

# Textbook Analysis Practice on the TOEFL Reading Sections by the Application of LDA Topic Modeling

## Abstract

Conventional textbook analysis primarily involves techniques like text analysis and statistics. This approach for handling and analyzing small sample data is not capable of managing and analyzing vast and intricate educational big data. With the advancements in digitization and constant reform in language assessment, the focus has shifted towards extracting information from text data and utilising digital technologies to predict the prospects of education. The utilization of big data and artificial intelligence technologies in learning and teaching is still at an early stage due to technical limitations, policy constraints, and other factors. This article attempts to present a preliminary concept that combines text mining and topic modelling to make exploratory predictions about the reading parts of *the Official Guide to the TOEFL iBT Tests* (2020) in an attempt to examine the practical significance and potential applications of the research outcomes and offer limited practical insights and concepts for the advancement and growth of education under the paradigm shift in educational research of the digital transformation of education.

Key words: digitization, reading sections, TOEFL textbook analysis, text mining, topic modelling, digital transformation

## 1. Introduction

Textbooks contribute to the development of the assessment and evaluation system for learners, as well as influencing the process of classroom teaching and learning (Mahmood, 2011). TOEFL is one of the most widely recognized English texts, and its textbooks can serve as useful teaching materials in classroom instructional practice. This study aimed to analyse the instructions provided in *the Official Guide to the TOEFL iBT Tests* (2020) using text mining and topic modelling techniques. The motivation behind this analysis was the lack of previous studies on TOEFL textbook instruction, the increasing importance of educational research in the times of educational digital transformation, together with the availability about educational big data. The goal was to

uncover the underlying educational principles within the instructional material. Educational research necessarily enhances its scope and comprehensiveness in order to effectively address the problems and opportunities arising from the digital revolution in education.

Halliday and Hasan (1976) pointed out that the hypothesis highlights that individuals' pre-existing knowledge and cognitive processes are significantly influenced by their existing information and knowledge structure. Louwerse and van Peer (2005) define a theme as the 'central idea of a text or discourse' that is typically associated with human activities, directly involving human interests and concerns, and often carrying emotional connotations. However, students, even those with extensive vocabulary and strong foundational skills, typically find it unsatisfying to comprehend and accurately respond to the English question format. Students tend to prioritize enhancing their vocabulary and honing their grammar abilities while neglecting the contextual backdrop and cultural implications of the text, resulting in an inaccurate comprehension of the author's true aim. This paper primarily employs the document analysis method to propose suggestions for topic modelling mining teaching techniques pertaining to the reading sections of *the Official Guide to the TOEFL iBT Tests* (2020). This is achieved through the application of the Latent Dirichlet Allocation (LDA) topic model technique, which reveals concealed subjects within the datasets.

## **2. Method**

### **2.1 Research Design and Sampling Method**

This article primarily incorporates the document analysis method to propose suggestions for topic modeling mining teaching techniques related to the reading parts of the *Official Guide to the TOEFL iBT Tests* (2020) via the Latent Dirichlet Allocation (LDA) topic model technique, which uncovers hidden subjects in the datasets. The document analysis approach for this exploration will focus solely on analyzing a textbook, specifically *the Official Guide to the TOEFL iBT Tests* (2020) as one of the EFL curriculum publications. In this study, the term reading instruction refers to the way reading is imparted in class, whereas the concept reading pedagogy refers to the proposed style of reading sections in the instructional material. Topic refers to the main subject of a written work, which is the word or phrase that all the works are associated with. The introduction of AI technology has created new requirements for the change of educational research paradigms in

response to the digitization of education. The conventional approach to educational studies depends primarily on strategies like textual evaluation and statistical assessment. This approach for organizing and evaluating small samples is not capable of processing and evaluating huge and intricate academic information. The digitization initiative has led to a focus on text data mining and the potential of digital technologies in predicting the prospects of education. Investigators' instructional selections could offer scientific insight for the advancement of education from a future standpoint. This will serve as a novel trajectory for future investigations and expeditions. Hence, educational researchers must focus on the incorporation of digital technology innovation and educational practice, which presents new requirements for the subject matter of educational research.

To elucidate the mechanisms by which text comprehension occurs, one must possess a deep comprehension of the cognitive processes involved in representing, processing, and utilizing knowledge. Yang Xue (2019) asserts that when individuals comprehend novel information, they must establish connections between unfamiliar concepts and their existing knowledge and past encounters. A cognitive framework allows readers to organize and interpret new information based on their pre-existing knowledge, hence enhancing comprehension. Hence, this study will conduct qualitative research through document analysis to explore the TOEFL English textbook's topic model and suggest strategies with the aim of uncovering or forecasting educational patterns, eventually better instructing reading.

## **2.2 Strengths and Limitations**

According to Halliday and Hasan (1976), a significant aspect of English language acquisition includes developing the ability to accurately anticipate and make predictions. Prior predictions before the reading phase assist pupils in accessing relevant prior knowledge and activating essential mental frameworks. Iftita and Yusnita (2019) stated that the English textbook instructions cover all themes. Themes will establish the writer's main point. Instructional clauses must include it to support English textbook learning objectives. Flake (2017) found a high link between curriculum and student achievement. Thus, in the context of English reading, teachers should give priority to cultivating students' fundamental English language knowledge, enhancing their linguistic

framework, and establishing a strong groundwork for further comprehension and analysis. The studies on the washback effects of English language tests are notably inadequate, which stands in stark contrast to their significance. Therefore, additional investigations should be dedicated to this subject matter (Watanabe, 1996).

International standard language tests, especially TOEFL (the recent reform in 2019 and 2023), have undergone continuous reform and updates to keep up with the changing times. The associated texts are undergoing updates, necessitating a commensurate adjustment in teaching materials and objectives for educators to align with the current era. Digital technology allows for the gathering of extensive data on teaching and learning, offering diverse evidence to support educational research. The language instruction materials undergo regular updates to keep up with the rapid advancements in emerging digital technologies. Nevertheless, conventional empirical and theoretical models may struggle to keep pace with the progress of contemporary times. This exploration aims at examining the practical significance and potential applications of the research outcomes and offer limited practical insights and concepts for the advancement and growth of education under the paradigm shift in educational research of the digital transformation of education.

## **2.3 Sampling Method**

According to Patton (2014), document analysis is a highly beneficial research methodology that has been employed for numerous years. This approach involves the examination of diverse forms of written materials, such as books, newspaper stories, academic journal articles, and institutional reports. Any document that includes written content has the potential to be used for qualitative analysis. Meanwhile, Hani Morgan (2021, p. 66) has put forward that “during the selection process, researchers need to do more than just check for authenticity, credibility, representativeness, and meaning. They also need to decide on a sampling technique to construct the corpus that will allow them to achieve the goals of the research study.” Within the realm of educational research, there is a growing need to cultivate researchers’ ability to think critically about educational data, especially in the context of the ongoing digital transformation in education. Also, international standard language tests, especially TOEFL (the recent reform in 2019 and 2023), have undergone continuous reform and updates to keep up with the changing times. The associated texts are always undergoing updates,

necessitating a commensurate adjustment in teaching materials and objectives for educators to align with the current era. Digital technology allows for the gathering of extensive data on teaching and learning, offering diverse evidence to support educational research. The language instruction materials undergo regular updates to keep up with the rapid advancements in emerging digital technologies. Nevertheless, conventional empirical and theoretical models may struggle to keep pace with the progress of contemporary times.

Conventional educational research primarily aims to comprehend and elucidate educational phenomena based on theoretical knowledge and quasi-experiments. It often investigates small samples or specific research scenarios. The digitization of education has resulted in the emergence of educational big data, which encompasses valuable educational information. The choice of reading passages utilized in reading teaching might impact learners' understanding and proficiency with specific text types (Williams, 2005). As Hani Morgan (2021) claimed, Braun and Clarke (2006) noted that thematic analysis is a highly adaptable method that enables researchers to apply it to various sorts of studies and choose a study design that aligns with their specific interests and experience. Tok (2010) asserted that textbooks provide instructors with a structure and assist them in attaining the subject matter's goals. Therefore, for achieving digital education transformation, the full digitised application of TOEFL's reading sections by a topic-modelling qualitative analysis will be a driving force as an efficient teaching model for teachers to be more closely in line with international standards and effectively improve students' language ability.

Hani Morgan (2021) has put forward the thematic analysis of documents as a qualitative method. According to Yang Xue (2019), the content schema comprises two facets. One factor is the contextual or sociocultural setting in which reading materials are situated, while the other factor is the prior information that readers possess. One of the key determinants of learners' English reading comprehension is their ability to effectively activate prior knowledge. The learners' comprehension capacities may be hugely influenced by the content of the schema. Learners can enhance their comprehension by enhancing their reading background information. So, aiming to improve students' prior reading knowledge, Latent Dirichlet Allocation (LDA) can be applied to *the Official Guide to the TOEFL iBT Tests* (2020) for finding hidden topics. The results can serve as a more objective and scientific resource for researchers to observe and analyse educational phenomena. LDA topic

modelling analysis approaches offer an analytical tool for educational research, which can visually represent intelligent analysis of educational data. Based on the data correlations, the tools can efficiently uncover the association between data through statistical analysis and make intelligent predictions.

### 3. Relative Ethics Consideration

Patton (2015) has pointed out that document analysis, having been employed for several years, can be viewed as an esteemed research methodology, whose approach involves the examination of diverse document genres, such as novels, newspaper stories, academic papers, and institutional reports, and any textual material is likely to be used as material for qualitative analysis. The data are sourced from openly published textbooks: *the Official Guide to the TOEFL iBT Tests* (2020), designed for individuals taking tests and educators involved in English language acquisition according to the Educational Testing Service. Meanwhile, the data collection process adheres to copyright regulations. Generally, the analysis of data derived from published articles, particularly in the realm of non-commercial academic research, can be regarded as “fair use”. Furthermore, it is crucial to continuously recognize and give credit to the ETS and *the Official Guide to the TOEFL iBT Tests* (2020) when using any information from the guide.

According to Hani Morgan (2022, p. 70), she said that “the process of conducting a document analysis starts with finding the documents for the study. In selecting the documents, researchers need to consider several factors. Flick (2018) mentioned four factors to use when deciding which documents to include: 1. Authenticity 2. Credibility 3. Representativeness 4. Meaning.” When applying qualitative document analysis (QDA), the mentioned four factors should be taken into consideration.

As an openly authoritative official publication, *the Official Guide to the TOEFL iBT Tests* (2020) is a primary source without any modifications or inauthenticity that is created and managed by the Educational Testing Service (ETS), the largest organization in the world dedicated to educational assessment and research. Their objective is to promote excellence and fairness in education and the delivery of equitable, accurate, and dependable evaluations. To be clear, as Flick (2018) strives to

do, it is of great importance to assure the precision of all the materials (referred to as the textbooks) included in the analytical procedure. Nevertheless, one of the advantages of *the Official Guide to the TOEFL iBT Tests* (2020) is that this textbook does really provide assurances or guarantees regarding the precision, thoroughness, or appropriateness of the content for any specific use. Based on this fact, the authenticity and credibility of this textbook and its origin can be greatly ensured.

Meanwhile, as Hani Morgan (2022) pointed out, “representativeness has to do with how typical a document is.” Merriam and Tisdell (2016) have asserted that examining written materials, such as books and scholarly articles, might be advantageous due to the enduring nature of the information. Investigators have the potential to exert an impact on those interviewed through surveys or inspections. Nevertheless, as they perform a documented investigation using already existing documents, the information remains unaltered. As an authoritative official publication, this textbook can offer information about the TOEFL examination setting policy in which study participants are involved. It serves as an example of material that provides context if one wants to use a clever expression. Documents serve as a means of observing previous occurrences, offering both background and professional insights. Such information and insights can aid researchers in comprehending the foundational origins of language teaching challenges and can reveal practical texts for test-takers and relative educators. The researcher can utilize data extracted from documents, such as written records, to provide a contextual framework for the data obtained during instructional activities.

Additionally, *the Official Guide to the TOEFL iBT Tests* (2020), to some degree, can facilitate learners and educators in gaining familiarity with the test’s delivery method and experiencing the process of answering questions within specified time limits. According to McGrath (2002), textbooks are crucial as they include the necessary information, establish the course of study, and offer guidance for delivering instruction. This official guide aims to provide assistance in comprehending the language abilities required for achieving success in both the test and the classroom. It is meaningful to utilizing the practice exams in this book to identify the learners’ most deficient talents.

#### 4. Research Tool

According to Bowen (2009), qualitative document analysis offers a systematic and methodical approach for extracting significance from written testimony. This investigation attempts to explore the reading sections of the TOEFL iBT Tests to uncover the knowledge since its latest reform (2019 and 2023). Text mining and topic modelling can effectively identify and detect research themes and underlying trends via a thematic qualitative analysis, which can serve as a reliable indicator for effective teaching.

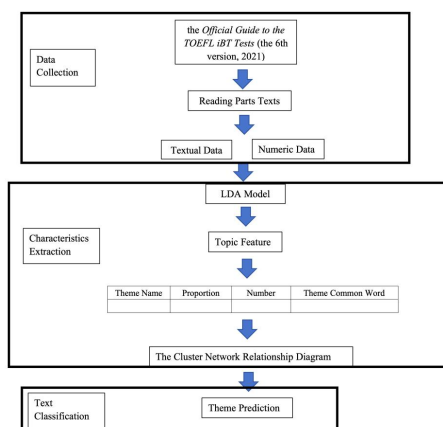
Latent Dirichlet Allocation (LDA) is a widely-used machine-learning means for uncovering latent topics within a collection of documents (Blei *et al.*, 2003). With the advancements of digitization technology, the Latent Dirichlet Allocation (LDA) topic modeling has become one of the most popular modelling methods and is being used by more and more researchers to model and analyze large-scale document data. The LDA topic model method can automatically mine the correlation between topics and words in document data, helping us better understand the internal structure and logic behind the information. According to Blei *et al.* (2003), the basic idea is: (1) Each document is a random mixture of several topics in a certain proportion. For example, in a two-topic model, document 1 can be considered to be a mixture of 10% topic A and 90% topic B, while document 2 is a mixture of 70% topic A and 30% topic B, and so on. The higher the probability, the more relevant the topic is to the document. (2) Each of them consists of a random mixture of various words in a specific proportion, and the distribution of words can represent the characteristics of the topic.

According to Yokie and Thomas (2018, p. 25), advocates of particular syllabus tasks argue that certain courses priorities the development of every skill necessary within the educational framework and spend more time looking over past content rather than introducing novel concepts (Usiskin, 2003). In contrast, specialized courses direct the attention of learners towards particular categories of knowledge while maintaining the broader framework and organized aspects of the educational programme (Cuoco, Goldenberg, & Mark, 2010). Also, Sangari and Mashatan (2022) noted that topic modelling is a robust text analysis technique used to analyse a wide range of journals and explore advances in research in different fields such as data science, decision sciences, computer

science, and SCM. It assists in obtaining valuable knowledge and identifying scholarly changes related to newly developed technologies. Through the examination of vocabulary usage across papers, topic modelling can discern latent subjects and cluster associated keywords. According to Jelodar *et al.* (2019), the objective of Latent Dirichlet Allocation (LDA) lies in uncovering the underlying subjects present in the corpus by analyzing the words. The most probable terms inside the various subjects supply an obvious hint of the topical content that the theme of the discussion is related to. However, so far, there have been few thorough studies that have thoroughly examined the theme parts of *the Official Guide to the TOEFL iBT Tests* (2020), especially by the Latent Dirichlet Allocation (LDA) topic model, and reliably predicted the results according to theme qualities and statistical parameters. This investigation successfully solves the deficiencies in the current field of research.

This study improves upon the conventional Latent Dirichlet Allocation (LDA) model by applying it specifically to the reading component of TOEFL iBT tests. The newly developed model, known as the Domain-Specific LDA (DS LDA), has enhanced skills in subject modelling. In order to predict the theme analysis in the reading parts, a random forest classifier uses the thematic characteristics produced by DS LDA. The content mining system designed for reading section analysis is depicted in Figure 1. It consists of three tiers: data collection, characteristic extraction, and text classification. For each document in the corpus, LDA defines the following generative process: 1) For each document, a topic is extracted from the topic distribution; 2) Extract a word from the word distribution corresponding to the extracted topic; 3) Repeat the above process until every word in the document has been traversed (Wu, Son and Wang, 2020b).

**Figure 1. LDA Theme Text Analytic**



## 4.1 Data Collection

The largest organization in the world that focuses on educational assessment and research, the Educational Testing Service (ETS), developed and oversaw *the Official Guide to the TOEFL iBT Tests* (2020), which is an openly released resource. Given this fact, the genuineness and reliability of this textbook and its source may be significantly guaranteed.

The collected data includes textual elements, such as project descriptions that contain extracted topical features, as well as numerical components, such as the theme name, its proportion, number, common term, and distribution over the entire passage. The suggested predictive approach utilizes a collection of five qualities to generate a random forest model, and provides the basis for creating a cluster network relationship diagram.

## 4.2 Characteristics Extraction

As Vayansky and Kumar (2020) have narrated, during the clustering process, words are initially allocated to themes in a random manner, and the system can calculate the conditional likelihood of topics inside each file in an iterative manner. Feature extraction through classification analysis is essential as it categorizes texts into subjects, making it a critical process. After extracting data from the target textbooks, sentences are subjected to lexical separation, which involves breaking them down into individual words and identifying the part of speech (POS) of every single word. Only nouns are considered significant for extraction in this analysis, as they are the most representative elements in the written content. In order to enhance the process of extracting nouns, stop words are omitted. The thematic features and motivations derived from the analysis are combined with the theme name, its proportion, number, common term, and distribution throughout the passage to predict the contextual theme of the target reading passage. Via text feature extraction, unstructured data can first be converted into structured data, which is fully prepared for the next step of mining and processing in order to construct a thesaurus of high-frequency keyword distribution.

## 4.3 Text Classification

Through the establishment about a thesaurus of high-frequency keyword distribution, the next step

is to evaluate the model in an attempt to check whether the topics make sense, whether the words belong to the correct topics, etc. Subsequently, priority should be given to the theme analysis by interpreting the text data based on the obtained theme distribution so that readers can understand the hidden information and constructs in the text data. In the LDA model, weight is an important concept, which refers to the importance of a topic in a document or a word in a topic. Usually, a numerical value serves as the weight in the LDA model. The larger the value in the figures stands for the higher the contribution to relative topic or document. When describing and analyzing model results, you can refer to the weight of each topic.

#### 4.4 Data Analysis

In the LDA topic model, its visual mining output results have three major features: 1. Directly output the keywords and their weights contained in each topic; 2. Directly output the probability table that each text belongs to each topic; 3. Visualization of LDA results. The LDA model believes that the frequency of each word appearing in a document can be understood as the superposition of two processes: one is to see what topics the document contains and determine a topic from it; the other is to see what words the topic contains and determine a word from it. In this way, we can analyze the multiple themes behind this document, as well as the most important theme. Similarly, we can see the most popular words under each topic and deduce what the topic is about.

Here is a text called *The Mystery of Yawning* (p. 519) from *the Official Guide to the TOEFL iBT Tests* (2020) to present how to apply the LDA model to theme classification. For instance, when setting one theme word about this text, determining the number of topics that yield the most meaningful results: yawning. After converting the text into a text suitable for classification tasks, this article counts the word frequency of each keyword. However, feature words with smaller word frequencies have relatively little impact on many tasks. Therefore, this article selects the 30 feature words with the highest word frequency and uses these word frequencies to model, thereby improving accuracy. The top 30 most relevant terms for the theme word yawning will also be generated as follows: Table 1.

**Table 1: Top 30 Most Relevant Terms for Theme Word: Yawning**

|    | Words                  | Part of Speech      | Frequency | Number | tf-idf       | Theme Proportion |
|----|------------------------|---------------------|-----------|--------|--------------|------------------|
| 2  | yawning                | noun, proper noun   | 19        | 5      | -0.002149205 | yawn:100.00%     |
| 3  | yawn                   | verb, noun          | 18        | 4      | 0            | yawn:100.00%     |
| 4  | have                   | auxiliary verb,verb | 12        | 4      | 0            | yawn:100.00%     |
| 5  | theory                 | noun                | 9         | 4      | 0            | yawn:100.00%     |
| 6  | activity               | noun                | 8         | 2      | 0.002535414  | yawn:100.00%     |
| 7  | adult                  | noun                | 5         | 2      | 0.001584634  | yawn:100.00%     |
| 8  | before                 | adverb, preposition | 5         | 2      | 0.001584634  | yawn:100.00%     |
| 9  | lung                   | noun                | 5         | 1      | 0.002842429  | yawn:100.00%     |
| 10 | oxygen                 | noun                | 5         | 2      | 0.001584634  | yawn:100.00%     |
| 11 | often                  | adverb, preposition | 4         | 2      | 0.001267707  | yawn:100.00%     |
| 12 | about                  | adverb, preposition | 4         | 3      | 0.000553772  | yawn:100.00%     |
| 13 | breathing              | noun                | 4         | 2      | 0.001267707  | yawn:100.00%     |
| 14 | level                  | noun                | 4         | 3      | 0.000553772  | yawn:100.00%     |
| 15 | change                 | noun                | 4         | 2      | 0.001267707  | yawn:100.00%     |
| 16 | function               | noun                | 4         | 3      | 0.000553772  | yawn:100.00%     |
| 17 | accord                 | verb, noun          | 3         | 3      | 0.000415329  | yawn:100.00%     |
| 18 | people                 | noun                | 3         | 3      | 0.000415329  | yawn:100.00%     |
| 19 | breathe                | noun                | 3         | 1      | 0.001705457  | yawn:100.00%     |
| 20 | volunteer              | noun                | 3         | 2      | 0.00095078   | yawn:100.00%     |
| 21 | increase               | verb, noun          | 3         | 3      | 0.000415329  | yawn:100.00%     |
| 22 | sleep                  | noun                | 3         | 2      | 0.00095078   | yawn:100.00%     |
| 23 | life                   | noun                | 3         | 2      | 0.00095078   | yawn:100.00%     |
| 24 | fetus                  | noun, proper noun   | 3         | 1      | 0.001705457  | yawn:100.00%     |
| 25 | blood                  | noun                | 3         | 2      | 0.00095078   | yawn:100.00%     |
| 26 | the tiredness theory   | noun phrase         | 2         | 2      | 0.000633854  | yawn:100.00%     |
| 27 | the theory             | noun phrase         | 2         | 1      | 0.001136972  | yawn:100.00%     |
| 28 | skin conductance       | noun phrase         | 2         | 1      | 0.001136972  | yawn:100.00%     |
| 29 | physiological activity | noun phrase         | 2         | 1      | 0.001136972  | yawn:100.00%     |
| 30 | pure oxygen            | noun phrase         | 2         | 1      | 0.001136972  | yawn:100.00%     |
| 31 | the frequency          | noun phrase         | 2         | 1      | 0.001136972  | yawn:100.00%     |

In addition, the LDA topic model can also count the clustering of words and their co-occurrence frequencies and judge the topics in the document based on the obtained information. Keywords can be visually exhibited by a word cloud, especially in text mining application. Words that appear more frequently are displayed larger or more vividly, thereby rendering the keywords a cloud-like appearance. Colour pictures to perceive the main themes and core ideas of the text data in an attempt to visually highlight the keywords that use more frequently in the relative article so that web browsers can understand the gist of the text by just scanning the text at a glance. It mainly uses text mining and visualization technology. A personalized word cloud is not only a way of expressing research and analysis content but also a kind of “artwork” in advertising media. In the LDA topic model, a quick and convenient word cloud picture can be formed. Word clouds can visually display keywords, making them more intuitive and artistic. The following table 2 is the word cloud in *The Mystery of Yawning* (p. 519). From this visualization table, the theme and its relative words about this text can be vividly revealed. Also, the LDA topic model can automatically generate a visual co-occurrence knowledge map (Table 3) of research hot topic words, which can deeply display the word relationship. The LDA topic model, along with its visual representation, is an effective tool for analyzing text. The LDA topic model may extract topic information from the text by modelling the corpus, enabling us to comprehend the concealed structure within the data.

**Table 2: the Word Cloud about the Mystery of Yawning (p. 519)**

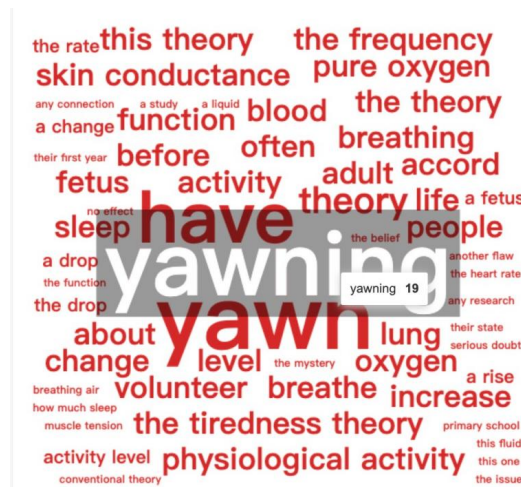
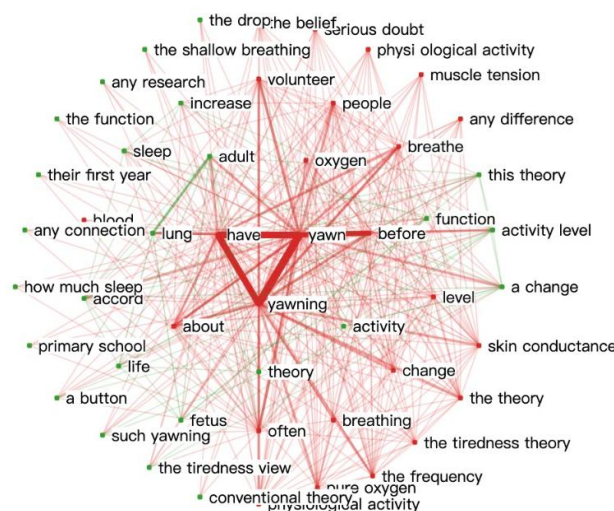


**Table 3: the Visual Co-occurrence Knowledge Map about *the Mystery of Yawning* (p. 519)**



## 5. Conclusion

As Halliday (2014) claimed, the concept of text topic refers to the fundamental method by which phrases connect together. This research proposes attempts to design specific hypotheses and measures in terms of the content, theme, and vocabulary of the TOEFL iBT Tests English theme corpus construction in order to provide a reference for the construction of the international language testing corpus. Topic models are extensively utilized in domains such as machine learning. Latent

Dirichlet Allocation (LDA), also known as invisible Dirichlet distribution, holds a significant role in the field of topic modelling. Through our research on English text analysis, the LDA topic model can be utilized to examine the clustering of terms and their co-occurrence frequencies. This allowed us to evaluate documents by analyzing the themes derived from the given information. It has a significant influence on forecasting future patterns in exams and greatly affects teaching methods. Effective instructions that align with learning objectives should clearly convey a narrative or a sequence of events (Levy, 1996). The construction of the TOEFL iBT Test English corpus will be an inherent necessity to promote the development of international English teaching, and it is an important material for cultivating English talents in line with international standards. It will play an indispensable role at present, promote the related theoretical research of international language testing, enhance the auxiliary work of the TOEFL iBT Tests English corpus classroom teaching, and boost the work of autonomous English learning.

Conventional educational research typically employs causal inference reasoning, which involves using experimental design and statistical analysis to ascertain the effects of a specific intervention on student learning. Even though the main idea or title of some articles can serve as a guarantee of their main theme, traditional text analysis fails to thoroughly explore the inherent relationships between words and uses the highest word frequency to model, thereby improving accuracy. However, text mining can now extract topic information from a corpus to model it, allowing us to see the hidden structure in the data. Nevertheless, conventional reasoning is undergoing a transformation due to the increasing prevalence of data-intensive scientific research. Research combines correlation analysis with causal inference, focusing on extracting the correlation between data and creating accurate predictions. The ongoing progress of digital transformation in education has led to a reevaluation of different types of educational data. The potential of educational big data has gradually become a dominant reform trend to support decision-making in education.

## Reference

- Blei, D., Edu, B., Ng, A., Jordan, M. and Edu, J. (2003). 'Latent dirichlet allocation', *Journal of Machine Learning Research*, 3, pp.993-1022. Available at:  
<https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?ref=https://githubhelp.com>
- Bowen, G.A. (2009) 'Document analysis as a qualitative research method', *Qualitative Research Journal*, 9(2), pp. 27-40. Available at: <https://doi.org/10.3316/QRJ0902027>
- Cuoco, A., Goldenberg, E.P. and Mark, J. (2010) 'Contemporary curriculum issues: organizing a curriculum around mathematical habits of mind', *The Mathematics Teacher*, 103(9), pp.682-688. doi:<https://doi.org/10.5951/mt.103.9.0682>.
- Flake, L.H. (2017) 'A look at the relationship of curriculum and instruction and the art and science of teaching', *Asian Journal of Education and Training*, 3(2), pp.82-85. Available at:  
<https://doi.org/10.20448/journal.522.2017.32.82.85>
- Halliday, M.A.K. and Hasan, Ruqaiya. (1976) *Cohesion in English*. London: Longman.
- Iftita Kharisma Rosi and Yusnita Sylvia Ningrum (2019) 'Textual meaning of English textbook instruction', *ELT Forum: Journal of English Language Teaching*, 8(2), pp. 184-189. Available at:  
<https://doi.org/10.15294/elt.v8i2.32956>
- Jelodar, H. et al. (2019) 'Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey', *Multimedia Tools and Applications*, 78(11), pp. 15169-15211. Available at:  
<https://doi.org/10.1007/s11042-018-6894-4>
- Levy, S. (1996) *Starting from Scratch*. Portsmouth: Heinemann Educational Books.
- Louwerse, M. and Peer, W.V. (2002) *Thematics: Interdisciplinary Studies*. Amsterdam: John Benjamins Publishing.
- Mahmood, K. (2011) 'Conformity to quality characteristics of textbooks: the illusion of textbook evaluation in Pakistan', *Journal of Research and Reflections*, 5(2), pp.170-190. Available at:  
<https://ue.edu.pk/jrre/articles/52006.pdf>
- McGrath, I. (2016) *Materials Evaluation and Design for Language Teaching*. Edinburgh: Edinburgh University Press.
- Merriam, S.B. and Tisdell, E.J. (2016) *Qualitative Research: a Guide to Design and Implementation*. San Francisco, CA: Jossey-Bass.
- Patton, M.Q. (2014) *Qualitative Research & Evaluation Methods*. Singapore: SAGE Publications.
- Roller, C.M. and Matambo, A.R. (1992) 'Bilingual readers' use of background knowledge in

learning from text', *TESOL Quarterly*, 26(1), pp. 129-141. Available at:

<https://doi.org/10.2307/3587372>

Tok, H. (2010) 'TEFL Textbook evaluation: from teachers' perspectives', *Educational Research and Reviews*, 5(9), pp. 508.

Usiskin Z (2003) 'The integration of the school mathematics curriculum in the United States: History and meaning', *Integrated mathematics: Choices and challenges*, pp.13-31.

Watanabe, Y. (1996) 'Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research', *Language Testing*, 13(3), pp.318-333. Available at:

<https://doi.org/10.1177/026553229601300306>

Williams, J.P. (2005) 'Instruction in reading comprehension for primary-grade students: a focus on text structure', *The Journal of Special Education*, 39(1), pp. 6-18. Available at:

<https://doi.org/10.1177/00224669050390010201>

Wu, J., Son, G. and Wang, S. (2020b) 'A competency mining method based on Latent Dirichlet Allocation (LDA) model', *Journal of Physics: Conference Series*, 1682(1), p.012059. Available at:

<https://doi.org/10.1088/1742-6596/1682/1/012059>

Xue, Y. (2019) 'The use of schema theory in the teaching of reading comprehension', *International Journal of Liberal Arts and Social Science*, 7(4), pp.58-63. Available at:

<https://ijlass.org/articles/7.4.6.58-63.pdf>

Yokie Prasetya Dharma and Joni, T. (2018) 'An analysis of English textbook relevance to the 2013 English curriculum', *Journal of English Educational Study*, 1(1), pp.24-33. Available at:

<https://doi.org/10.31932/jees.v1i1.277>