Mining Text in Online News Reports of COVID-19 Virus: Key Phrase Extractions and Graphic Modeling

新冠病毒在線上新聞報導中之文本探勘:關鍵詞萃取及圖形建模

Wen-Ta Tseng¹

Received: 14 September 2020 / Revised: 12 October 2020 / Accepted: 18 November 2020 /

Published online: 19 December 2020

© National Taiwan Normal University 2020

Abstract

The recent emergence and spread of COVID-19 have altered the way the world operates. As this pandemic continues to run its course, both language educators and learners around the world are facing a unique set of challenges. In this day and age, there are no more relevant, pressing, or internationally ubiquitous news stories than those related to COVID-19. For L2 learners to have a seat at the global table, it is necessary to learn languages using news stories. Hence, the current study applied text mining techniques to explore and identify patterns among news stories related to COVID-19. In the study, a corpus collecting online news reports about COVID-19 was analyzed. A number of *R* packages including *readtext*, *tidytext*, *ggplot2*, and *ggraph* were jointly employed to extract key phrases and construct a graphic model underlying the news corpus. A popular term-extraction method often used in text mining—*term frequency-inverse document frequency* (TF-IDF)—was utilized to extract the key phrases from the news reports on the COVID-19 virus. A wordnet structure was then established to uncover potentially salient thematic components. The pedagogical implications for language education and vocabulary assessment are further discussed.

摘要

在這個時代，沒有比與新冠病毒的新聞報導更緊迫或國際上無處不在的新聞報導。因此，本研究應用文本探勘技術來探索和識別與新冠病毒有關的新聞報導模式。這項研究對關於新冠病毒在線新聞報導的語料庫進行了分析。數個*R*套裝程式，包括*readtext*、*tidytext*、*ggplot2*和*ggraph*同時合用於提取關鍵語和構建一個圖形模型。本研究運用文本探勘中經常使用的常用術語提取方法「術語頻率- 向文檔頻率 (TF-IDF)」來提取有關新冠病毒新聞報導中的關鍵詞，並從而建立了一個字網結構以揭示潛在且顯著的主題組成元素。此研究進一步討論研究結果於語言教育與詞彙測驗之意涵。

Keywords Text mining · COVID-19 · TF-IDF · wordnet structure

✉ Wen-Ta Tseng
wenta.tseng@ntust.edu.tw

關鍵詞 文字挖掘, 新冠病毒, 術語頻率- 向文檔頻率, 字網結構。 .

Introduction

The recent emergence and spread of COVID-19 have altered the way the world operates. Individuals, communities, corporations, and governments have all been forced to quickly adapt to the new regulations and social distancing procedures put in place. While society as a whole embraces this “new normal”, people with chronic illness, including those who need elective treatment and psychiatric care, are losing the ability to receive treatment. Furthermore, economies around the world are facing hardships as leaders employ a variety of strategies to keep their citizens safe and slow the spread of COVID-19, while the world races toward a cure.

As this pandemic continues to run its course, educators and learners around the world are facing a unique set of challenges. With summer vacation coming to an end, students and teachers are tasked with acclimating to socially distanced learning environments and/or virtual classrooms. There is no script or playbook to follow in this situation, so creative tactics, trial-and-error, and adaptation must all be employed to come up with new, creative solutions to the challenges faced in education.

News about the pandemic is constantly being updated, tasking language educators with a unique set of challenges during this pandemic. As the global lingua franca, the English language is spoken as a first language by over 350 million people worldwide and as a second language by almost 500 million. Because of this distinction, English is a major language of many countries and the main language used for international business, aviation, diplomacy, technology, science, and more. As such, English language educators and learners should actively pursue studying English using content relevant to their educational and professional aspirations and goals.

Content and language integrated learning (CLIL) began in the European Union in the 1990s as a way for learners to study English as a second language (L2) using content that was interesting, relevant, and useful for future career goals [1, 2]. CLIL views L2 learning as a means for accessing important information which combines critical thinking, creative use of the L2, strengthening of problem solving and communication skills, and learning by doing [3]. Teachers in a CLIL environment act as facilitators and let learners take the lead on pursuing their learning aims, thus giving learners increased confidence [4] and motivation [5].

English medium of instruction (EMI) is an educational approach that parallels CLIL, and has been gaining popularity in L2 classrooms worldwide, especially in the Asia-Pacific region [6]. Over time, CLIL and EMI have proven effective at endowing L2 learners with skill-specific language proficiency as well as life-long learning abilities. In order for learners to further themselves by becoming fluent in more technical areas of the English language, they should employ the tactics of CLIL and EMI. For instance, learners in a CLIL course are required to be familiar with the knowledge and content around a specific topic. To achieve this familiarity, the fundamental step to be taken is to acquire the key terms/phrases of a domain-specific subject or topic. Likewise, for L2 learners to have a seat at the global table, it is also necessary to learn by using news reports. In this day and age, there are no more relevant, pressing, or internationally

ubiquitous news stories than those related to COVID-19. As the Internet usage surges to 70% during the pandemic [7], the need to utilize digital sources of information and news as material for learning English is more important than ever, giving CLIL/EMI a chance to strengthen as a tool for language education.

As information technology continues to advance at an incredible rate, the Internet is being integrated into the personal and professional lives of people all over the world. With this rise in the Internet use comes a massive amount of unstructured text data produced by users on social networks, web pages, forums, blogs, comment sections, review sites, and more. The information that can be collected from these countless data points offers valuable and authentic insights into the way people around the world think and behave, but such a massive amount of data is difficult to interpret on a large scale and across so many different platforms. This also holds true of the news reports of COVID-19 after its outbreak in January 2020.

In light of the global impact brought by COVID-19, the content of online news reports from this unprecedented pandemic may have scholastic value based on its potentials to inform the theory and practice of CLIL/EMI in a timely manner. However, there have been millions of online news reports about COVID-19 from all over the world since the outbreak began which would be impossible for researchers to sort and analyze manually. Although mainstream computer software typically featured for analyzing massive textual data can effectively create a list of keywords or key phrases from a large corpus, the outcome of this conventional approach may lack sensitivity in grasping subtle yet critical nuances between distinct documents in the same corpus. Since the development of COVID-19 news stories was quite dramatic before and after the outbreak, identifying keywords and key phrases without the ability to account for the distinct episodes of the event cannot precisely and genuinely establish the norm to which these key notions and core concepts of the entire corpus are referred. To this end, the current study aims at addressing this research gap by employing an important data analysis technique—*text mining*—to effectively extract and validly classify key information from a corpus including millions of online news reports about COVID-19. More importantly, the study also aims at constructing a graphic model to depict the semantic structure underlying the unstructured data of the online news corpus to critically inform instruction and learning in the context of CLIL and EMI.

Text Mining

Text mining “seeks to extract useful information from unstructured textual data through the identification and exploration of interesting patterns” ([8], p. 227). Text mining is not only considered more valuable than data mining but also far more complex as it employs software which brings together elements of database systems, artificial intelligence, machine learning, and mathematical statistics to filter vast quantities of unstructured data. Once this data is filtered, useful and valuable patterns emerge and can be explored and leveraged. These patterns could be difficult or impossible for human readers to identify using skills like skimming, parsing, critical thinking, and linguistic analysis. Text mining programs typically utilize a “bag-of-words” approach which moves through text swiftly and efficiently.

In the bag-of-words approach, each word is considered a unique piece of the document being mined. Bag-of-words is favored in text mining because it can be performed quickly and does not require technical expertise or a large budget [9]. The results of bag-of-words text mining are organized attributes and patterns called *term document matrices* (TDM) or *document term matrices* (DTM) which can then be integrated into a machine learning framework [10]. In a TDM, each document represents a column, while individual words or word groups make up the rows. Conversely, DTM columns are comprised of individual words or word groups, while rows are made up of documents.

To clarify how bag-of-words text mining operates, consider the following example based on the dictionary (W) containing all words that appear one or more times in a corpus of documents (D) [9, 10]. A single document (d_n) would be represented as a vector of weights ($w_{1n}, \dots, w_{|W|n}$). These weights ($w_{in} \in \{0, 1\}$) signify the absence or presence of a certain word in a document. The regularity of the i_{th} word in the n_{th} document is also measured to calculate how frequently these words appear in a given document. Normalization can be further utilized in order to quantify the frequency of words as represented by numbers between 0 and 1, regardless of the overall length of a document. Sequences of words or characters which form phrases, called n -grams (e.g., bigrams for two word phrases, and trigrams for three word phrases), can also be identified within a text. When a corpus (D) is mined using bag-of-words, the corpus can be operationalized as a matrix made up of document vectors (rows) and terms (columns) which allow for decomposition techniques such as dimensionality reduction and k-means clustering analysis.

Term Weighting

In text mining, researchers weight each term in a document according to its level of importance. In this way, the efficiency of mining and usability of data are both increased as more important words are weighted heavier than less important words, and stop words might be removed altogether. The most common term weighting method refers to *term frequency–inverse document frequency* (TF-IDF) [11]. In TF-IDF, word weights are calculated proportionally to the words within a document and reversely proportional to words in other texts within a corpus. This is the most popular method, but also the most complex and labor intensive, since the entire corpus must be entered in the system for a thorough investigation. The orthogonal form of the mathematical equation of TF-IDF is depicted as follows:

$$W_{ij} = \log \frac{N}{DF_i} (1 + \log TF_i)$$

In this equation, w_{ij} refers to the weight of a word (t_i) observed in a document (d_j). N is the total number of documents in the corpus, TF_i points to the occurrence of a word in the document, and DF_i is the number of documents which include the word within the corpus (Qaiser & Ali). In sum, the aim of TF-IDF is to identify the words that are particularly important in the context of each document by inhibiting the impact of commonly used words across the documents while uplifting the weight of words that are unique, yet less used, in the corpus. Due to the popularity of TF-IDF in the field of text mining, the current investigation employs TF-IDF to identify the key phrases (i.e.,

bi-grams) within a US news corpus in order to inform the practice of English language instruction and learning.

Descriptions of the News Corpus of COVID-19

In the current study, a corpus collecting online news reports about COVID-19 was analyzed (http://data.gdeltproject.org/blog/2020-coronavirus-narrative/live_online/news/20191101-20200326-covid19.csv.gz). The time period of the online news reports about COVID-19 fell between January and March of 2020. The rationale for studying the news corpus was that in January of 2020, the whole world did not pay enough attention to the disastrous impact that might be incurred by COVID-19. By March 2020, after the spread of the virus due to a careless attitude, the whole world was experiencing an unprecedented health crisis. Hence, it is informative in practice to investigate whether differences exist in the lexical expressions that mirror the trajectory of the impact brought about by COVID-19.

As shown in Table 1, the online news corpus to be analyzed in total includes 5,269,595 words. The number of words collected in January, February, and March was 247,674, 1,253,810, and 3,768,111, respectively. Under the theoretical framework of text mining, the mini-corpus of January, February, and March 2020 was operationalized as three separate documents.

As reflected by the number of words collected in the three different time points, the impact of COVID-19 became extremely widespread and serious shortly after February because the number of words in relation to the news reports of COVID-19 in March was 3 times more than that of February and 15 times more than that of January. To systematically mine the text within the news corpus, a number of *R* packages were jointly used in the current investigation: *readtext*, *tidytext*, *ggplot2*, and *ggraph*. In light of the uniqueness of the corpus, the current study aims to address the following research questions:

RQ1: In what ways may the key phrases to be identified vary across the three time-related mini-corpora?

RQ2: How are the key phrases of the entire news corpus interrelated?

Results

RQ1: In what ways may the key phrases to be identified vary across the three time-related mini-corpora?

As reported in Table 2, in the document of online news released in January of 2020, the top 10 key phrases were *patent site*, *prevent respiratory*, *vaccination accounts*, *surveillance systems*, *leading clinicians*, *espionage charges*, *serial conspiracy*, *false suggestions*,

Table 1 The number of words in the three documents and the entire corpus

Document	January 2020	February 2020	March 2020	Total
Word count	247,674	1,253,810	3,768,111	5,269,595

Table 2 The key phrase extractions and their TF-IDF indices

Document	Key phrases	N	TF	IDF	TF-IDF
January	Patent site	80	.0013	1.10	.0014
	Prevent respiratory	80	.0013	1.10	.0014
	Vaccination accounts	80	.0013	1.10	.0014
	Surveillance systems	46	.0008	1.10	.0009
	Leading clinicians	40	.0007	1.10	.0008
	Espionage charges	32	.0005	1.10	.0006
	Serial conspiracy	32	.0005	1.10	.0006
	False suggestions	29	.0005	1.10	.0006
	Excessive anxiety	69	.0011	0.41	.0004
	Fake miracle	26	.0004	1.10	.0004
February	Rapid developments	342	.0012	1.10	.0013
	World health	291	.0010	1.10	.0011
	Rumors creating	217	.0008	1.10	.0009
	Carnival celebrations	204	.0007	1.10	.0008
	Closed theater	204	.0007	1.10	.0008
	Performance canceled	204	.0007	1.10	.0008
	Canceled school	181	.0007	1.10	.0008
	Barring flights	125	.0005	1.10	.0006
	Authorities' incompetence	124	.0005	1.10	.0006
	Pneumonia epidemic	112	.0004	1.10	.0004
March	Coronavirus pandemic	2523	.0029	.41	.0012
	National lockdown	607	.0007	1.10	.0008
	Corona-19 outbreak	1260	.0015	.41	.0006
	NHS guidance	455	.0005	1.10	.0006
	Jointly combating	445	.0005	1.10	.0006
	Disinformation campaign	1160	.0013	.41	.0005
	National quarantine	407	.0005	1.10	.0005
	Moderate symptoms	385	.0004	1.10	.0004
	Crushing US	364	.0004	1.10	.0004
	Combating fraud	349	.0004	1.10	.0004

excessive anxiety, and *fake miracle*. Clearly, these key phrases in general pointed to the uncertainty and doubt of the health impact that might be brought about by COVID-19. Immediate and effective actions had yet to be taken by most western countries. As indicated by key phrases such as *patent site*, *prevent respiratory*, *vaccination accounts*, and *leading clinicians*, one of the main concerns at this stage appeared to be around whether or not a new vaccine should be invented against the new virus, and foreign research institutes were competing to acquire patents for their so-called new drugs. However, the other main concern of the news reports at this stage seemed to be more politically implicated than reports from January of 2020. Foreign governments cast doubt on the source of COVID-19 and critically questioned the motive of the discovery of COVID-19, as suggested by key phrases like *espionage charges*, *serial conspiracy*, *false suggestions*, *excessive anxiety*, and *fake miracle*.

Furthermore, in the document of online news released in February of 2020, the top 10 key phrases were *rapid developments*, *world health*, *rumors creating*, *carnival celebrations*, *closed theater*, *performance canceled*, *canceled school*, *barring flights*, *authorities' incompetence*, and *pneumonia epidemic*. Essentially, these 10 key phrases highlighted the swift global spread of the COVID-19 virus and the fact that nearly all foreign governments became fully aware of the immeasurable impact of the COVID-19 virus. Following this realization, governments began taking actions such as canceling large-scale social and educational activities nationwide to prevent further outbreak of the virus. However, because the COVID-19 virus had already been globally and widely spread in February 2020, it was too late for foreign governments to implement any systematic post hoc safeguards against the new virus. The key phrase *authorities' incompetence* rightly reflected this fact.

Finally, in the document of online news released in March 2020, the top 10 key phrases referred to *coronavirus pandemic*, *national lockdown*, *corona-19 outbreak*, *NHS guidance*, *jointly combating*, *disinformation campaign*, *national quarantine*, *moderate symptoms*, *crushing US*, and *combating fraud*. It was not until March that formal terms were coined and used by the world which explicitly acknowledged the extreme seriousness and contagiousness of the COVID-19 virus (i.e., *coronavirus pandemic* and *corona-19 outbreak*). At this stage, it had been confirmed by scientists that the virus could still be highly contagious even with *moderate symptoms*. The key phrase *crushing US* precisely portrayed the image of how COVID-19 devastated the USA in March. Upon the recognition of its destructive impact on human health, more extreme nation-level measures were taken to fight against the spread of the virus. The clustering of key phrases including *national lockdown*, *jointly combating*, *combating fraud*, and *national quarantine* critically suggested the insufficiency or failure of prior safeguards in preventing further outbreak of the virus. In the face of the global health crisis, unfortunately, numerous waves of diplomatic conflicts between countries were instigated by the *disinformation campaign* regarding the source of the COVID-19 virus. This *disinformation campaign* came at an unfortunate time when the entire human race should have stood and fought together to save lives during the outbreak of one of the deadliest viruses in the twenty-first century.

- *RQ2: How are key phrases of the entire news corpus interrelated?*

Figure 1 illustrates the overall wordnet system comprising the key phrases identified in the whole news corpus. It could be seen that in the entire wordnet structure, a number of salient small wordnet systems could be observed. Understandably, the first wordnet sub-system centered around the core word *coronavirus*, which co-occurred with *virus*, *outbreak*, *epidemic*, *crisis*, *pandemic*, and *misinformation*. Furthermore, the two keywords *pandemic* and *misinformation* functioned as the bridging nodes to which the other two wordnet sub-systems, which centered around the core words *health* and *spreading*, were jointly connected. In the wordnet sub-system centered around the core word *health*, a number of keywords co-occurred with it (i.e., *world*, *public*, *global*, *organization*, *experts*, *authorities*). In the wordnet sub-system which revolved around the core word *spreading*, the keywords that co-occurred mainly carried derogatory sentiment (i.e., *fake*, *rumors*, *misinformation*, and *disinformation*). Through the word *fake* as a bridging node, still another wordnet sub-system which revolved around the core word *news* could be further uncovered, and the words that typically co-occurred with *news* were predominantly associated with journalism (e.g., *conference*, *press*, *media*). Notably, “fake news” is a term that has been coined by Trump; this term

Fig. 1 The wordnet structure of online news corpus of the COVID-19 virus

together (i.e., “fake news” vs. “fake” and “news”) could be highlighted to shed further light on the misinformation spread in America and the distrust of the media that has become prevalent since Trump’s election in 2016.

There were still two more wordnet sub-systems that carried derogatory sentiment. One centered around the core word *conspiracy*, and the other one revolved around the core word *false*. In particular, the *false* wordnet sub-system was close to another wordnet sub-system which was politically implicated by key phrases such as *federal government*, *Chinese government*, *Chinese city*, and [New] *York city*. It was interesting to note that the larger wordnet sub-system established by *health* was surrounded by the three smaller wordnet sub-systems: *false*, *conspiracy*, and *government*.

In sum, most of the salient wordnet sub-systems clustered to the left of the entire wordnet structure. Based on the semantic structures underlying the wordnet sub-systems, four salient thematic components appeared to dominate the themes of news reports during the 3-month development of the COVID-19 virus investigated in the current work: *threat of virus*, *health concern*, *spread of falsehoods*, and *social-political friction*.

Discussion and Conclusion

This study employed text mining techniques to extract key phrases and built up a wordnet structure underlying a sizeable online news corpus in relation to the COVID-19 virus. Ten key phrases were separately extracted from the three different documents coded in accordance with the corresponding months in which the online news was broadcasted. It was found that the key phrases identified for each month meaningfully hallmarked the central or primary conceptions that were either directly or indirectly coined or implicated by the outbreak of the COVID-19 virus.

Hence, a comprehensive understanding of the background and meanings of the 30 key phrases identified will be useful for L2 learners and instructors to acquire fundamental knowledge about the social-historic background surrounding the outbreak of the COVID-

19 virus. In practice, it is suggested that these 30 key phrases be designed into L2 textbooks for CLIL or EMI courses. Since these 30 key phrases are time-dependent in nature, course material designers should take this feature into account so that L2 learners can be aware of why the human race reacted passively and suspiciously to the deadly COVID-19 virus even when its outbreak was repeatedly confirmed by numerous countries. Critically, acquiring these 30 key phrases helps learners to cultivate insights on the outbreak trajectory of the COVID-19 virus in the history of epidemiology.

In a similar vein, as per the graphic modeling outcome, there were four salient thematic components (i.e., *threat of virus*, *health concern*, *spread of falsehoods*, and *social-political friction*) existing within the news corpus. Unlike the timeline feature inherent in the 30 key phrases, the four thematic components are salient in equal weight across the three time points (i.e., 3 months) as defined in the news corpus. Hence, the course material designers may keep in mind the principle that the four thematic components should coexist in texts to depict a complete storyline regarding the impact brought about by the COVID-19 virus. It should be noted that the wordnet structure figure can be incorporated into textbooks of EMI/CLIL so that learners can have a high-level overview of this structure and develop a critical awareness of the semantic links underlying wordnet sub-systems.

More importantly, there has been a growing body of research focusing on assessing learners' vocabulary knowledge via computerized adaptive testing (CAT) over the past two decades (e.g., [12, 13–16]). These pioneering psychometric studies in general suggest the superiority of CAT in terms of its adaptiveness and measurement precision over traditional paper-and-pencil-based tests in assessing learners' vocabulary knowledge. Vocabulary knowledge is multidimensional in nature and is typically conceptualized as consisting of size and depth dimensions [17]. Given the psychometric strengths of CAT, applying text mining techniques such as term frequency–inverse document frequency (TF-IDF) analysis efficiently helps test designers uncover and identify important keywords/key phrases in a specific corpus. Once these have been identified, different facets of word knowledge associated with keywords/key phrases can be systematically operationalized into a test bank so that a CAT can be validly implemented. Keywords identified by the multiplication of both term frequency (TF) and inverse document frequency (IDF) are more representative and valid than choosing keywords based solely on term frequency (TF), a traditional frequency index typically used in the field of corpus linguistics. In a nutshell, the outcome of the current study suggests that the adoption of TF-IDF may scaffold CAT to establish a large, valid item bank which can then be used to assess numerous facets of vocabulary knowledge that are deemed important and fundamental in a corpus.

In conclusion, the current study mined the texts of a sizeable corpus in relation to the online news reports of the COVID-19 virus. The study applied TF-IDF, a classic, well-known method to extract representative phrases from distinct documents inherent within the corpus. A graphic model featuring the wordnet structure underlying the corpus was also constructed. The pedagogical implications for English language education were further provided. It is hoped that the outcome of the study may help elevate the inner awareness of the entire human race regarding how to closely cooperate together to defend against another type of deadly virus, if one should arise.

Appendix 1

#####Part (A): The following R codes are used to identify the key phrases#####

```
one <- data_frame(news, month)

text <- one %>%
  group_by(month) %>%
  mutate(linenum = row_number(),
  ignore_case = TRUE)%>%
  ungroup()

text_df <- mutate(text, text = text$news)

Covid_bigrams <- text_df %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2)

Covid_bigrams %>%
  count(bigram, sort = TRUE) %>%
  print(n=100)

bigrams_separated <- Covid_bigrams %>%
  separate(bigram, c("word1", "word2"), sep = " ")

bigrams_filtered <- bigrams_separated %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word)

bigram_counts <- bigrams_filtered %>%
  count(word1, word2, sort = TRUE)
bigram_counts

bigrams_united <- bigrams_filtered %>%
  unite(bigram, word1, word2, sep = " ")

list(bigrams_united$bigram)

bigram_tf_idf <- bigrams_united %>%
  count(month, bigram) %>%

bind_tf_idf(bigram, month, n) %>%
  arrange(desc(tf_idf))
```

#####Part (B): The following R codes are used to implement the graphic modeling###

```
set.seed(2016)
a <- grid::arrow(type = "closed", length = unit(.15, "cm"))
ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "red", size = 4) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()
```

References

1. Darwin, R., Lo, Y. Y., & Lin, A. M. Y. (2020). Examining CLIL through a critical lens. *English Teaching and Learning*, 44, 103–108.
2. Smit, U., & Dafouz, E. (2012). Integrating content and language in higher education: An introduction to English-medium policies, conceptual issues and research practices across Europe. *AILA Review*, 25, 1–12.
3. Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge: Cambridge University Press.
4. Marsh, D., & Langé, G. (2000). *Using languages to learn and learning to use languages*. Jyväskylä: University of Jyväskylä.
5. Lasagabaster, D. (2011). English achievement and student motivation in CLIL and EFL settings. *Innovation in Language Learning and Teaching*, 5(1), 3–18.
6. Brown, H., & Bradford, A. (2017). EMI, CLIL, & CBI: Differing approaches and goals. In P. Clements, A. Krause, & H. Brown (Eds.), *Transformation in English education* (pp. 328–334). JALT: Tokyo.
7. Beech, M. (2020). COVID-19 pushes up internet use 70% and streaming more than 12%. <https://www.forbes.com/sites/markbeech/2020/03/25/covid-19-pushes-up-internet-use-70-streaming-more-than-12-first-figures-reveal/#51b471263104>. Access date: August 01 2020.
8. Radovanović, M., & Ivanović, M. (2008). Text mining: Approaches and applications. *Novi Sad Journal of Mathematics*, 38, 227–234.
9. Huilgo, P. (2020). Quick introduction to bag-of-words (BoW) and TF-IDF for creating features from text. <https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/>. Access date: August 01 2020.
10. Kaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181, 25–29.
11. Kaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to. Documents. *International Journal of Computer Applications*, 181, 25–29.
12. Kremmel, B. (2018). *Development and initial validation of a diagnostic computer-adaptive profiler of vocabulary knowledge* unpublished doctoral dissertation, University of Nottingham, Nottingham.
13. Santos, V. D. O. (2017). *A computer-adaptive test of productive and contextualized academic vocabulary breadth in English (CAT-PAV): Development and validation*. Graduate Theses and Dissertations, 16292. Ames: Iowa State University.
14. Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education*, 97, 69–85.
15. Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement*, 35, 155–167.
16. Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.
17. Mizumoto, A., Sasao, Y., & Webb, S. A. (2019). Developing and evaluating a computerized adaptive testing version of the word part levels test. *Language Testing*, 36, 101–123.

Affiliations

Wen-Ta Tseng¹

¹ Department of Applied Foreign Languages, National Taiwan University of Science and Technology, Taiwan, Republic of China