Journal of Positive School Psychology
2022, Vol. 6, No. 3, 5034 – 5044
http://journalppw.com

Analysis of Covid-19 related Phrases Using Corpus-Based Tools: Dualisms Language & Technology

Wong Wei Lun[a], Mazura Mastura Muhammad*[a] , Muhamad Fadzllah Zaini[a], Ashrol Rahimy Damit[b], Carrine Teoh-Ong[c], Charanjit Kaur Swaran Singh[a], Norhayati Yusoff[d]

[a]*Sultan Idris Education University, Malaysia*
[b]*University Brunei Darussalam, Brunei*
[c]*Bond Holdings, Malaysia*
[d]*College Matriculation Perak*

**corresponding author: mazura@fbk.upsi.edu.my*

Abstract

Since 2020, COVID-19 has become a contemporary topic of study in a wide variety of disciplines. One of these is education, since the linguistic competence of Malaysian International high school learners is unknown as physical schooling has been replaced by online learning, which involves no formal language tests. Hence, a corpus-based analysis of salient phrases related to COVID-19 among Malaysian international high school learners through data visualisation makes a useful contribution because it is concerned with both COVID-19 and the linguistic competence of such learners. The study's main objective is to recognise these salient phrases in extended writing. A quantitative strategy was employed, focusing on quantitative computational linguistics analysis with a corpus-based methodology. On a methodological level, it analyses salient phrases, concordances, log-likelihood values, and their semantic domains using LancsBox, trigram from N-grams, a log-likelihood calculator, and the USAS semantic tagger. The findings include the ten most salient phrases discovered for the trigram and related concordances, as well as their associated semantic domains. Overall, the study reports the linguistic competence evident among Malaysian international high school learners post-COVID-19 and contributes to the corpus-based education research literature.

Keywords: Corpus-based, phrases, COVID-19, LancsBox, Malaysia.

1. INTRODUCTION

COVID-19 began posing a hazard to every country in 2020 (Venkateswarlu et al., 2022). According to Gupta et al. (2022), it is a contagious infection so physical personal contact should be avoided. The optimal response to this pandemic was to impose lockdowns and quarantine procedures. Each country employed this policy to contain the pandemic. As a result, virtually every profession was impacted, most notably education. All educational institutions were closed and online education became the primary mode of instruction (Aaraj et al., 2022; Wong et al., 2022).

COVID-19 is gradually becoming a current research trend in every field (Li & Jiang, 2021). Numerous educational studies are COVID-19-related. This heralds a new era of education in the context of the pandemic, in which further studies and conclusions must be available for policymakers, instructors, and educators to develop future teaching and learning plans.

According to Cambridge International Examination (2017), international high school learners in Malaysia should demonstrate knowledge and understanding of various vocabulary items, including phrases. For Jo (2022) and Lokker and Jezrawi (2022), extended writing is a more appropriate method of language assessment since it incorporates various language components such as vocabulary, phrases, and syntax. Bychkovska (2021) and Wang et al. (2020) highlighted the importance of phrases in extended writing.

Numerous issues were readily noticed once COVID-19 had begun to adversely affect education (Dereso et al., 2022; Silva et al., 2022; Wong et al., 2022). First, due to online learning and intermittent attendance, the linguistic competence of Malaysian international high school learners was found to be uncertain (Hargreaves, 2021; Hartshorn & McMurry, 2020). The students' English instructors made no precise assumptions or conclusions about their linguistic abilities, but did this mainly through writing. Moreover, Singh et al. (2017), and Don and Sriniwass (2017) noted that international high school learners were recognised as displaying an ineffective usage of phrases in lengthy writing. Following that, a dearth of a corpus for Malaysian high school students connected to COVID-19 was noted (Juan et al., 2020) when a corpus-based

approach was used (Zaini, 2020; Zhang & Hao, 2018; Muhammad et al., 2017). Thus, the current study was undertaken to address the four issues raised above by using corpus-based analysis to identify salient phrases related to COVID-19 from the extended writing generated by Malaysian international high school learners. The main research question was formulated as follows: What are the most salient phrases identified from the extended writing produced by Malaysian international high school learners?

Simultaneously, this study addresses the research gaps indicated by Zamin et al. (2019) and Nair and Hui (2018) in their corpus-based studies of phrasal verbs in Malaysian secondary school textbooks and the common descriptive writing errors evident among Malaysian Chinese private school learners. These gaps are bridged by this study by highlighting the use of COVID-19-related phrases in the extended writing provided by Malaysian international high school learners.

Finally, the study's findings are critical for English educators since they offer a methodological reference for teaching phrases during COVID-19. Additionally, educational institution managers may wish to incorporate corpus-based approaches into their regular teaching and learning processes. Lastly, Malaysian policymakers must possess a benchmark for corpus-based research in education during the pandemic.

2. LITERATURE REVIEW

One approach and two theories are involved in this current study. First, the corpus-based approach drives the whole study. The concept of the corpus-based approach (henceforth, CBA) is discussed rigorously in this section. This is followed by explanations of the learning theory of constructivism and N-gram theory, the main underpinning theories used in this analysis of the extended writing produced by Malaysian international high school learners. A theoretical framework is illustrated below, with a discussion.

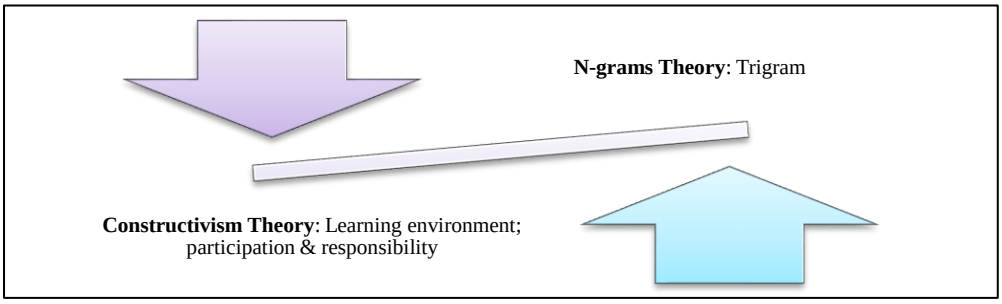


Fig. 1. Theoretical Framework.

2.1. Corpus-Based Approach (CBA)

First, the CBA uses the corpus as a source of pre-corpus linguistic descriptions. According to Muhammad et al. (2016), it is a method for delineating or exemplifying contemporary notions that had not primarily been established through primary reference to a corpus. During the experimental phase, research methodologies vary according to the specified or demonstrated theory. However, this process frequently includes speech-tagged corpora; information on words and the frequency of grammar elements; and interpretative analysis that incorporates co-occurrence features. This is more adaptable and robust than the corpus-driven method in terms of its applicability by researchers from different fields. As a result, the CBA is viewed as a tool or a set of abilities rather than a study subject. Numerous studies on corpora have been undertaken by systemic functionalists, most notably Zaini et al. (2021) and De Fina (2020). Meanwhile, Laske (2020) explored the diachronic evolution of language. Simultaneously, several recently completed studies on language variation (Devereaux, 2019; Villena-Ponsoda, 2019) fall under the heading of CBA.

The studies and research cited above demonstrate that the CBA is beneficial as its assists academics and linguists to justify and enhance theoretical assertions. Therefore, this study used the CBA to evaluate the ten most frequently used phrases related to COVID-19 among Malaysian international high school students in their extended writing.

2.2 Constructivism

The fundamental premise of constructivist theory is that students construct learning when they gain experience due to what they know (O'Connor & McCurtin, 2021). That is, learners make their meaning from their experiences. Constructivist thought is based on several cognitive theories developed by Piaget and Vygotsky. According to the former, learners learn actively, build systems, assimilate and accommodate all forms of science, and so forth. Vygotsky is responsible for social constructivism, group work, and internships, among other developments. Thus, it could be argued that constructivism is accountable for both "top-down" and "bottom-up" learning methodologies. This means that the teacher presents the broad concept

first and the learners subsequently receive the specifics. The teacher does not teach the details in this way, so learners will have difficulty grasping the details (Aljohani, 2017).

Constructivism views knowledge generation as an active process, in which learners construct cognitive structures through their interactions with their environment (Al-Sabaawi et al., 2021). Cognitive interaction occurs to the extent that their cognitive design builds reality. The mental form must constantly be modified and adapted to meet the requirements of the environment and the changing organism. Adjustment occurs regularly throughout the reconstruction phase (Amineh & Davatgari, 2015).

The most critical aspect of constructivism philosophy is that the learner should prioritise the learning process (Wu et al., 2021). Learners must actively increase their own knowledge, rather than relying on others. Learners must take ownership of their educational outcomes. Their inventiveness and vitality would then enable them to gain prominence in their cognitive lives. The focus of learning is experimental, a humanitarian adaptation based on tangible laboratory experience and dialogues with classmates, who then contemplate and create new concepts. As a result, the emphasis on education and teaching is not on educators but learners (Li et al., 2021).

Finally, it can be argued that constructivist learning is concerned with the following: (a) emphasising genuine learning in a meaningful context, (b) prioritising the process, (c) instilling learning in the framework of social experience, and (d) learning to construct experience.

Given this explanation and elaboration, constructivism was chosen as the learning theory for this study of Malaysian international high school learners in COVID-19 because this approach addressed the theme within constructivism that learners acquire phrases meaningfully, given their existing difficulties. The learning process was prioritised since Malaysia's Ministry of Education implemented online schooling as an immediate measure for education during the pandemic. Learners may have subsequently gained knowledge of COVID-19 phrases from their social experience and media outlets such as Facebook. Eventually, all actual experiences were transformed into knowledge.

2.3 N-gram Theory

Meanwhile, an N-gram is defined as a sequence of n-token words (Sidorov, 2019). N-grams, a computational linguistics invention, are used in the probabilistic N-gram method to predict the next word following a sequence of (n-1) words. For example, one may estimate the probability that the phrase "I scream" is followed by the word "loudly" by dividing the number of instances of the (4-gram) "I shout out loudly" in a text by the number of the cases of the (n-1 gram) "I scream".

The term N-gram is frequently used in corpus linguistics. Putri et al. (2021) conducted a preliminary evaluation to determine the superior performance of machine learning algorithms in recognising abusive language and hate speech on Twitter. The researchers employed a variety of machine learning methods, including Naive Bayes, Support Vector Machine, and Random Forest Decision Tree as classifiers, along with weighted word and character N-grams from Term Frequency – Inverse Document Frequency as feature extraction. The experiments employed a five-fold cross-validation strategy and were evaluated using F-1-Scores. After the investigation, the best F-1-Score was obtained for each dataset using a Support Vector Machine classifier using word N-gram features. In contrast, Zampieri (2021) emphasised the distribution of N-grams across varied corpora rather than the N-grams themselves.

Finally, as a focal point of this study, the trigram function in N-grams was chosen as the computational linguistic tool for analysing written phrases referring to COVID-19 produced by Malaysian international high school learners. Only two-, three- and four-word phrases in the top ten were interpreted and discussed in this study.

3. METHODS

The primary approach of this study followed a quantitative research design (Creswell & Creswell, 2017), which was used in conjunction with a corpus-based approach to bolster the study. Firstly, 100 learners from a Malaysian international school were recruited using a cluster sampling technique. They were in Year 7 and aged between 14 and 15. To protect the data from gender bias, 50 male and 50 female learners were chosen. They were required to write an extended text about COVID-19. Two modes of submission were made available: Google Classroom and paper-based essays. For data analysis, all the writing was transferred to Microsoft Word.

First, 100 extended writing samples were analysed for the ten most salient phrases. The trigram function in N-grams under Lancsbox was used to determine the most salient phrases (Taylor, 2018). Subsequently, the corresponding concordances were captured by screenshots. Additionally, a log-likelihood calculator was used to compare the salient phrases to the reference corpus, the Lancaster-Oslo-Bergen Corpus. The USAS semantic tagger was used to identify the semantic domains of the salient phrases (Rüdiger, 2020).

4. RESULTS AND DISCUSSIONS

This section summarises the results and discussions of the ten most salient phrases of trigram, along with their associated concordances, log-likelihood values, and semantic domains.

LancsBox was used to analyse 40,924 tokens for the ten most salient phrases. Based on the study, these ten salient phrases were derived from trigram.

4.1 Salient Phrases

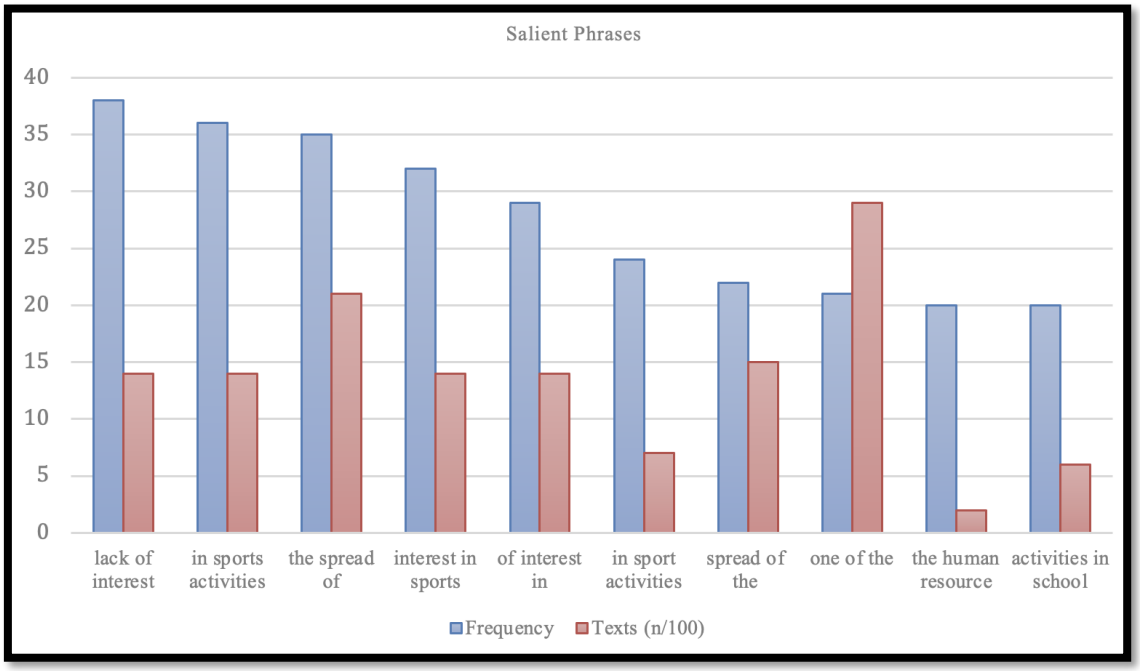


Fig. 2. Salient Phrases.

To answer Research Question One, the ten most salient phrases analysed from 100 extended writing were: *lack of interest, in sports activities, the spread of, interest in sports, of interest in, in sport activities, spread of the, one of the, the human resource* and *activities in school*. Referring to Figure 2, the phrase *the spread of* appeared in 21 texts out of 100, despite its frequency of occurrence being 35.0, compared to the top-ranked phrase *lack of interest*, which has a frequency of occurrence of 38.0. The implication is that in their extended writing, 21 international high school learners probably employed *the spread of* twice. Only 14 texts contained the top phrase *lack of interest*, which had the most significant frequency of 38.0. As a result, 14 international high school students were found to frequently use this phrase in their extended writing.

Another expression used was *in sport activities*. With a frequency of occurrence of 24.0, this phrase was discovered in only seven extended writing texts. This implies that seven international high school learners used repetitious behaviour when utilising this phrase. Furthermore, the phrase *the human resource* occurred 20.0 times; however, only two texts contained this phrase. These two learners were likely to have repeated the statement ten times apiece.

One of the is another phrase that occurred infrequently (*f*=21), yet this was located in 20 extended writing texts. Even with the highest frequency of 38, *lack of interest* was only identified in 14 extended writing texts. This demonstrates that the phrase *one of the* is well-known among the 100 Malaysian international high school learners since twenty of them used it a total of 21 times.

Furthermore, an intriguing outcome was the usage of *sport* and *sports*. The findings indicated that the utilisation of *sports* was more prevalent than the term *sport* itself. This occurrence, however, demonstrated that school learners were uncertain about the use of the single and plural forms of 'sport' in writing, particularly in phrasal words. Thus, English teachers could use this component to enhance learning.

4.2 Concordances

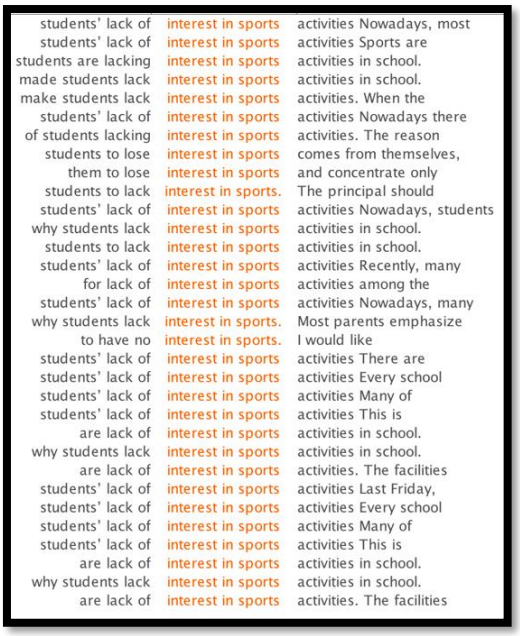
The concordances for the ten salient phrases from trigram were identified and are graphically shown. The context of concordances was adjusted to three so the results can be better understood.

Fig 3. Concordances of *lack of interest*.

the beginning of	the spread of	the disease, there
new guidelines to slow	the spread of	the coronavirus, including
significant health concerns,	the spread of	the coronavirus will
balance between controlling	the spread of	the disease and
and help prevent	the spread of	this virus. We
established to control	the spread of	coronavirus entail the
aimed at curbing	the spread of	COVID-19. Some such
globally to halt	the spread of	COVID-19. Today, millions
extent, has controlled	the spread of	the virus. However,
chances to control	the spread of	the disease. In
contamination will prevent	the spread of	infectious agents. Practices
needed to stem	the spread of	the virus. Epidemics
place to prevent	the spread of	the virus also
rush to reduce	the spread of	the disease by
killed due to	the spread of	this disease. The
quarantine to control	the spread of	the havoc of
vital to control	the spread of	the virus because
Wuhan, China. However,	the spread of	the disease is
are now online.	The spread of	the virus has
killed due to	the spread of	this disease. The
quarantine to control	the spread of	the havoc of
vital to control	the spread of	the virus because
required to contain	the spread of	the coronavirus, it
way of mitigating	the spread of	avian flu. It
way of mitigating	the spread of	avian flu. It
around the world.	The spread of	infectious diseases among
racing to slow	the spread of	the disease by
of June 2020.	The spread of	COVID-19 has sent
governments grappled with	the spread of	the disease by
growth associated with	the spread of	the virus may
Novel CoronaVirus. However,	the spread of	Covid 19 has
lockdown so that	the spread of	this virus can
that due to	the spread of	Covid 19 from
precautions so that	the spread of	this virus can
order to control	the spread of	Covid 19.

Fig. 5. Concordances of *the spread of*.

Fig. 4. Concordances of *in sports activities.*

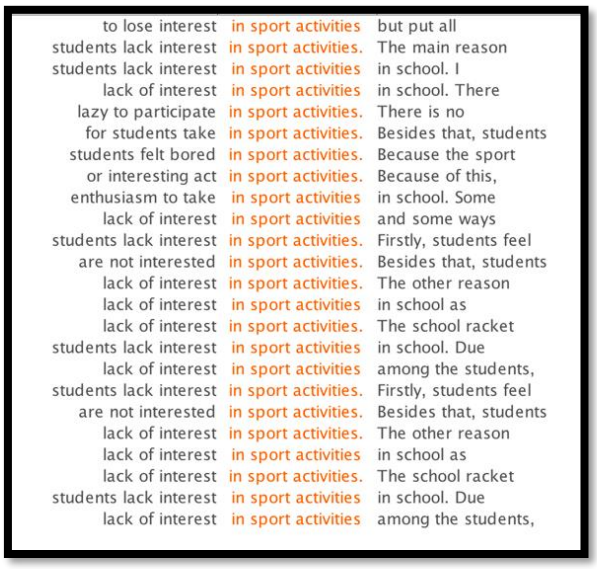


students' lack of	interest in sports	activities Nowadays, most
students' lack of	interest in sports	activities Sports are
students are lacking	interest in sports	activities in school.
made students lack	interest in sports	activities in school.
make students lack	interest in sports	activities. When the
students' lack of	interest in sports	activities Nowadays there
of students lacking	interest in sports	activities. The reason
students to lose	interest in sports	comes from themselves,
them to lose	interest in sports	and concentrate only
students to lack	interest in sports.	The principal should
students' lack of	interest in sports	activities Nowadays, students
why students lack	interest in sports	activities in school.
students to lack	interest in sports	activities in school.
students' lack of	interest in sports	activities Recently, many
for lack of	interest in sports	activities among the
students' lack of	interest in sports	activities Nowadays, many
why students lack	interest in sports.	Most parents emphasize
to have no	interest in sports.	I would like
students' lack of	interest in sports	activities There are
students' lack of	interest in sports	activities Every school
students' lack of	interest in sports	activities Many of
students' lack of	interest in sports	activities This is
are lack of	interest in sports	activities in school.
why students lack	interest in sports	activities in school.
are lack of	interest in sports	activities. The facilities
students' lack of	interest in sports	activities Last Friday,
students' lack of	interest in sports	activities Every school
students' lack of	interest in sports	activities Many of
students' lack of	interest in sports	activities This is
are lack of	interest in sports	activities in school.
why students lack	interest in sports	activities in school.
are lack of	interest in sports	activities. The facilities

Fig. 6. Concordances of *interest in sports*.

reasons students' lack	of interest in	sports activities Nowadays,
reasons students' lack	of interest in	sports activities Sports
reasons students' lack	of interest in	sports activities Nowadays
reasons students' lack	of interest in	sports activities Nowadays,
reasons students' lack	of interest in	sports activities Recently,
reason for lack	of interest in	sports activities among
reasons students' lack	of interest in	sports activities Nowadays,
reasons students' lack	of interest in	sports activities There
nowadays are lack	of interest in	sport activities in
for student's lack	of interest in	sport activities and
reasons students' lack	of interest in	sports activities Every
to students' lack	of interest in	sport activities. The
reasons students' lack	of interest in	sports activities Many
reasons students' lack	of interest in	sports activities This
students are lack	of interest in	sports activities in
students are lack	of interest in	sport activities in
students are lack	of interest in	sports activities. The
students are lack	of interest in	sport activities. The
the overwhelming lack	of interest in	sport activities among
reasons students' lack	of interest in	sports activities Last
reasons students' lack	of interest in	sports activities Every
to students' lack	of interest in	sport activities. The
reasons students' lack	of interest in	sports activities Many
reasons students' lack	of interest in	sports activities This
students are lack	of interest in	sports activities in
students are lack	of interest in	sport activities in
students are lack	of interest in	sports activities. The
students are lack	of interest in	sport activities. The
the overwhelming lack	of interest in	sport activities among

Fig. 7. Concordances of *of interest in*.



to lose interest	in sport activities	but put all
students lack interest	in sport activities.	The main reason
students lack interest	in sport activities	in school. I
lack of interest	in sport activities	in school. There
lazy to participate	in sport activities.	There is no
for students take	in sport activities.	Besides that, students
students felt bored	in sport activities.	Because the sport
or interesting act	in sport activities.	Because of this,
enthusiasm to take	in sport activities	in school. Some
lack of interest	in sport activities	and some ways
students lack interest	in sport activities.	Firstly, students feel
are not interested	in sport activities.	Besides that, students
lack of interest	in sport activities.	The other reason
lack of interest	in sport activities	in school as
lack of interest	in sport activities.	The school racket
students lack interest	in sport activities	in school. Due
lack of interest	in sport activities	among the students,
students lack interest	in sport activities.	Firstly, students feel
are not interested	in sport activities.	Besides that, students
lack of interest	in sport activities.	The other reason
lack of interest	in sport activities	in school as
lack of interest	in sport activities.	The school racket
students lack interest	in sport activities	in school. Due
lack of interest	in sport activities	among the students,

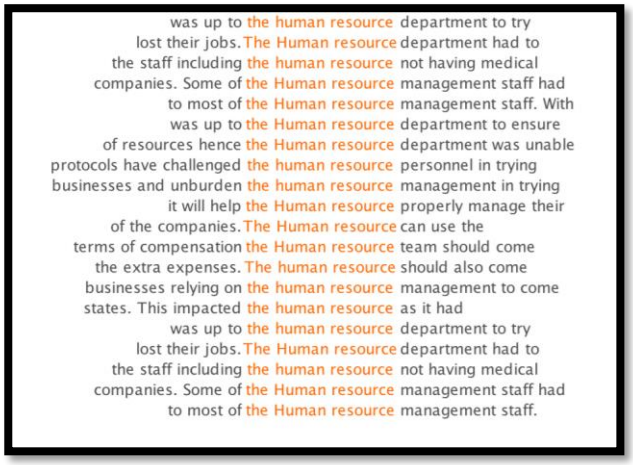
Fig. 8. Concordances of *in sport activities*.

beginning of the	spread of the	disease, there was
guidelines to slow the	spread of the	coronavirus, including closing
Health Organization (WHO).	Spread of the	coronavirus is causing
health concerns, the	spread of the	coronavirus will impact
between controlling the	spread of the	disease and keeping
has controlled the	spread of the	virus. However, most
to control the	spread of the	disease. In order
to stem the	spread of the	virus. Epidemics have
to prevent the	spread of the	virus also require
to reduce the	spread of the	disease by screening
of the rapid	spread of the	disease, face-to-face learning
to control the	spread of the	havoc of this
to control the	spread of the	virus because it
China. However, the	spread of the	disease is not
now online. The	spread of the	virus has encouraged
to control the	spread of the	havoc of this
to control the	spread of the	virus because it
to contain the	spread of the	coronavirus, it may
allowing for uncontrolled	spread of the	new infection, one
to slow the	spread of the	disease by testing
grappled with the	spread of the	disease by closing
associated with the	spread of the	virus may affect

Fig. 8. Concordances of *spread of the.*

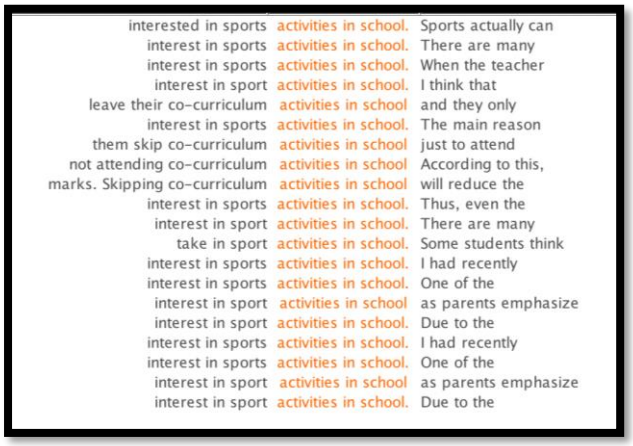
on top of	one of the	world's roofs—especially for
to climb on	one of the	peaks. Though it
Detroit, which is	one of the	cities around the
voice come over	one of the	walkie-talkies, "We've got
voice come over	one of the	walkie-talkies, "We've got
the services rendered .	One of the	solutions was there
conference setting, remain	one of the	most beneficial aspects
digital tools. Similarly,	one of the	most beneficial aspects
with regard to	one of the	most important questions
to climb on	one of the	peaks. Though it
every day. It's	one of the	reasons decorated actor
every day. It's	one of the	reasons decorated actor
of the language.	One of the	oldest pandemic-related words is quarantine. This originally referred to the
parental factor is	one of the	reasons too. The
activities in school.	One of the	reason is the
activities in school.	One of the	reason is the
By stressing every	one of the	countries it touches,
and epistemic values.	One of the	sources of science
Covid 19 is	one of the	most fatal virus
social distancing is	one of the	major steps that
the services rendered.	One of the	solutions was there

Fig. 9. Concordances of *one of the.*



was up to	the human resource	department to try
lost their jobs.	The Human resource	department had to
the staff including	the human resource	not having medical
companies. Some of	the Human resource	management staff had
to most of	the Human resource	management staff. With
was up to	the Human resource	department to ensure
of resources hence	the Human resource	department was unable
protocols have challenged	the human resource	personnel in trying
businesses and unburden	the human resource	management in trying
it will help	the Human resource	properly manage their
of the companies.	The Human resource	can use the
terms of compensation	the Human resource	team should come
the extra expenses.	The human resource	should also come
businesses relying on	the human resource	management to come
states. This impacted	the human resource	as it had
was up to	the human resource	department to try
lost their jobs.	The Human resource	department had to
the staff including	the human resource	not having medical
companies. Some of	the Human resource	management staff had
to most of	the Human resource	management staff.

Fig. 10. Concordances of *the human resource*.



interested in sports	activities in school.	Sports actually can
interest in sports	activities in school.	There are many
interest in sports	activities in school.	When the teacher
interest in sport	activities in school.	I think that
leave their co-curriculum	activities in school	and they only
interest in sports	activities in school.	The main reason
them skip co-curriculum	activities in school	just to attend
not attending co-curriculum	activities in school	According to this,
marks. Skipping co-curriculum	activities in school	will reduce the
interest in sports	activities in school.	Thus, even the
interest in sport	activities in school.	There are many
take in sport	activities in school.	Some students think
interest in sports	activities in school.	I had recently
interest in sports	activities in school.	One of the
interest in sport	activities in school	as parents emphasize
interest in sport	activities in school.	Due to the
interest in sports	activities in school.	I had recently
interest in sports	activities in school.	One of the
interest in sport	activities in school	as parents emphasize
interest in sport	activities in school.	Due to the

Fig. 11. Concordances of *activities in school*.

Concordances of salient phrases at the time suggested a relationship. The phrases *lack of interest, in sports activities, interest in sport, of interest in, in sport activities*, and *activities in school* were all connected because, according to the results, several school learners constructed the longer phrase: …*lack of interest in sport(s) activities in school*. Similarly, another point of agreement on school activities was the impact of skipping co-curricular activities; for example, several students wrote *Skipping co-curriculum activities in school will reduce*... Thus, it could be claimed that the key phrases employed focused on the concept of the learners' involvement in school co-curricular sporting activities. This topic appears to be a major issue for Malaysian international high school learners during COVID-19.

On the contrary, the concordances of *the spread of* and *spread of the* concerned COVID-19, as evidenced by the following examples:...*slow/prevent/curbing the spread of the virus/coronavirus/COVID-19*... Both of these prominent expressions have concordances with COVID-19. Compared to other phrases involving interests, sports, or school, COVID-19-related phrases appear to have been used less frequently by the international high school learners.

Subsequently, the phrase *the human resource* developed concordances that were unrelated to COVID-19, interest, sports, or school, and was utilised in a business management context. This seemed strange because this presented a completely different context than the others previously discussed. It might be deduced that some high school learners had been exposed to such contexts during COVID-19 through their parents or social media.

The phrase *one of the* has concordances with pandemic remedies, the benefits of digital gadgets, actors, quarantine, school, science, and COVID-19. According to the findings, *one of the* is a practical phrase because it may be employed in various settings. It was discovered that twenty international high school learners were proficient in utilising *one of the* in their extended writing.

4.3 Log-likelihood Values and Key N-grams

The log-likelihood values of the ten salient phrases identified from trigram are presented in Figure 12. Furthermore, the key N-grams that emerged are presented.

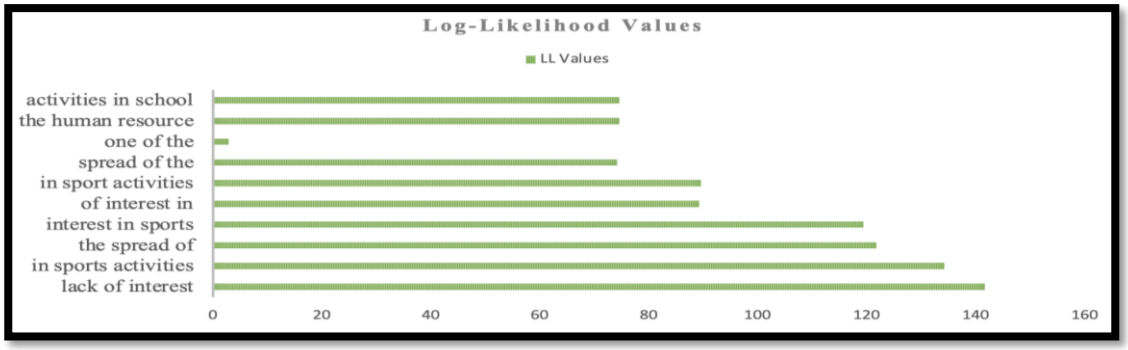


Fig. 12. Log-Likelihood Values.

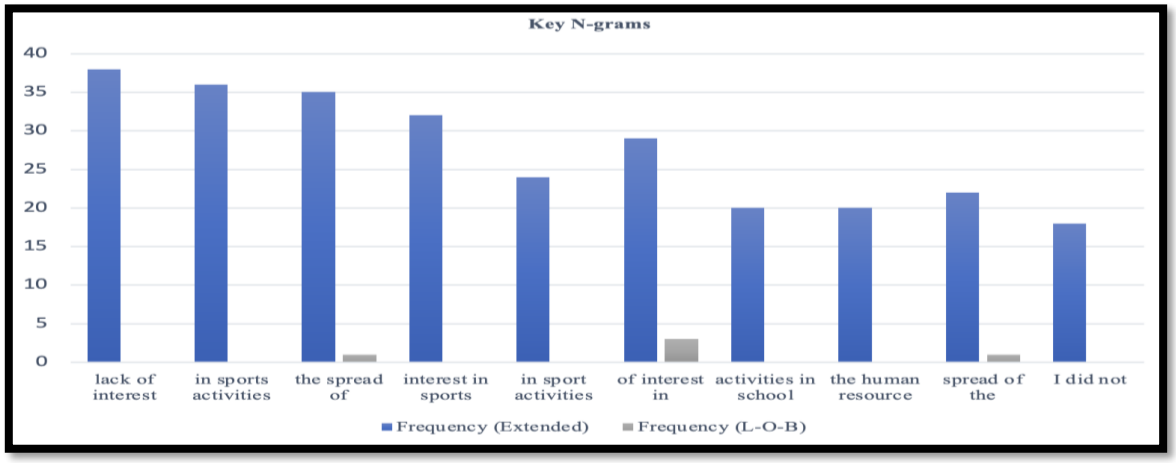


Fig. 13. Key N-grams.

Except for phrase *one of the*, every salient phrase in Figure 12 was found to be significant, based on the log-likelihood values and their significance level of <0.0001. This is consistent with the findings presented in Table 3, since all the nine salient phrases with high log-likelihood values and a significance level of <0.0001 were discovered to be positive key N-grams, except for the phrase *one of the*, which was replaced with the phrase *I did not*.

Following that, Figure 12 displays the log-likelihood values for each of the ten most salient phrases. Except for *one of the*, each phrase has a positive and high log-likelihood value. This demonstrates that *one of the* is a less significant phrase. Additionally, as Figure 13 shows, *one of the* is missing from the list of key N-grams; it was replaced with *I did not*. This demonstrates that, compared to L-O-B, *one of the* was not a critical phrase in the corpus of international high school learners in this study. Likewise, *the spread of*, *of interest in* and *spread of the* had incidence frequencies of 1,3, and 1 in L-O-B, respectively. This indicates that, except for *one of the*, the remaining nine salient phrases proved to be noteworthy in the corpus built by the 100 extended writing pieces generated by the Malaysian international high school learners.

4.4 Semantic Domain

Table 4 denotes and illustrates the semantic taggers and their meanings of the ten most salient phrases extracted from the 100 extended essays created by the Malaysian international high school learners. The primary category of each semantic tagger is identified, along with the frequency distribution in Figure 13.

Table 4. Meaning and Semantic Taggers of Ten Salient Phrases.

Phrases	Semantic Taggers	Meaning of Semantic Taggers
lack of interest	lack_A9- of_Z5 interest_X5.2+	A9-: giving Z5: grammatical bin X5.2+: interested/excited/energetic
in sports activities	in_Z5 sports_K5.1 activities_A1.1.1	Z5: grammatical bin K5.1: sports A1.1.1: general actions/making
the spread of	the_Z5 spread_F1 of_Z5	Z5: grammatical bin F1:food Z5: grammatical bin
interest in sports	interest_X5.2+ in_Z5 sports_K5.1	X5.2+: interested/excited/energetic Z5: grammatical bin K5.1: sports
of interest in	of_Z5 interest_X5.2+ in_M6	Z5: grammatical bin X5.2+: interested/excited/energetic M6: location and direction
in sport activities	in_Z5 sport_K5.1 activities_A1.1.1	Z5: grammatical bin K5.1: sports A1.1.1: general actions/making
spread of the	spread_F1 of_Z5 the_Z5	F1:food Z5: grammatical bin Z5: grammatical bin
one of the	one_Z8 of_Z5 the_Z5	Z8: pronouns Z5: grammatical bin Z5: grammatical bin
the human resource	the_Z5 human_S2/I3.1mf[i1.2.1 resource_S2/I3.1mf[i1.2.2	Z5: grammatical bin S2/I3.1mf[i1.2.1: people/ work and employment (generally) S2/I3.1mf[i1.2.2 : people/ work and employment (generally)
activities in school	activities_A1.1.1 in_Z5 school_P1/H1c	A1.1.1: general actions/making Z5: grammatical bin P1/H1c:education in general/ architecture, houses and buildings

According to Table 4, the ten major categories of semantic domains were identified as follows:

- A: general and abstract terms
- F: food and farming
- H: architecture, buildings, houses, and the home
- I: money and commerce
- K: entertainment, sports, and games
- M: movement, location, travel, and transport
- P: education
- S: social actions, states, and processes
- X: psychological actions, states, and processes
- Z: names and grammatical words

Figure 13 depicts the frequency of each semantic domain reported in Table 4 to highlight the most salient semantic domain.

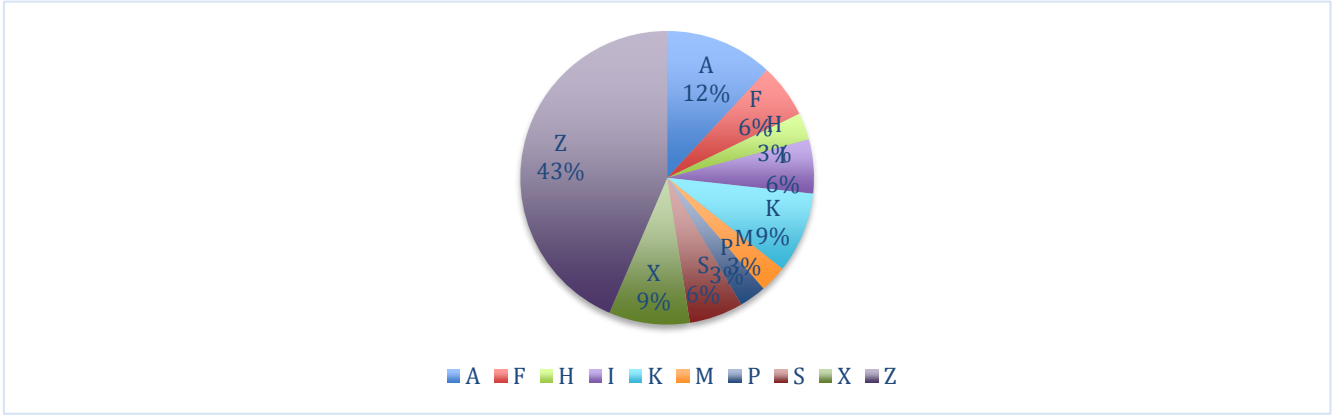


Fig. 14. Frequency of Each Semantic Domain.

These findings demonstrate that Z (names and grammatical words) was the most prominent category in the semantic domain, as determined by the ten most salient phrases retrieved from the 100 extended writing pieces written by the Malaysian international high school learners. The semantic domain analyses were classified into the following categories: (a) A: general and abstract terms, (b) F: food and farming, (c) H: architecture, buildings, houses, and the home, (d) I: money and commerce, (e) K: entertainment, sports, and games, (f) M: movement, location, travel, and transport, (g) P: education, (h) S: social actions, states, and processes, (i) X: psychological actions, states, and processes and (j) Z: names and grammatical words. According to Table 4 and Figure 13, semantic domain category Z had the most significant proportion, 43%, showing that it was the most prevalent semantic domain among the ten semantic domains analysed using semantic taggers. According to the analysis, all ten salient phrases had a word classified into the Z category, which was a grammatical bin of words such as *of, in*, or *the*. This implies that the international high school learners had successfully grasped the use of *in, of*, and *the*.

According to Figure 14, the semantic domain categories A and K had proportions of 12% and 9%, respectively. Four phrases were associated with semantic domain A: *lack of interest, in sports activities, in sport activities*, and *activities in school*. The terms associated with semantic domain A were *lack* and *activities*. On the other hand, three phrases relating to semantic domain X were found, including *lack of interest, interest in sports*, and *of interest in*, with the term *interest* serving as the primary keyword for this semantic domain.

The semantic domains H, M, and P tended to have the lowest percentage, 3%. These semantic categories were associated with only two phrases: *activities in school* and *of interest in*. The term *school* was semantically related to domains H and P. Similarly, the term *in* fell within the semantic scope of M. One could therefore argue that the international high school students used fewer words in the H, M, and P semantic domains.

5. CONCLUSION

The data indicate that the ten most salient phrases in 100 extended writing pieces created by Malaysian international high school students were connected to sports and COVID-19. This demonstrates that the international high school students were aware of the virus and athletic activities had become a focal point of discussion during the pandemic.

This data provide insights into the English writing skills of international high school learners after the outbreak of COVID-19. Moreover, English instructors were likely to be familiar with the key phrases used during COVID-19 and in their preferred learning resources. The findings serve as a guide for English educators as they plan their future instruction.

Numerous constraints on this study were recognised. Firstly, the token (word) count was insignificant compared to 1 million tokens. While the number would be sufficient for a small learner corpus, having additional tokens could result in more rigorous findings. Furthermore, the total number of participants in the research was limited to 100. Meanwhile, only trigram is used in this study, rather than other forms of N-grams.

Future researchers could bridge the gaps identified through these constraints. To begin with, additional tokens are recommended. The extended writing genre was chosen for this study; however, other writing genres would be feasible. Additionally, the research participants in future studies could be obtained from a diverse range of

educational levels rather than international high schools. Since trigram was employed in this study, future researchers could analyse their data using bigram or qualgram.

References

Aaraj, S., Farooqui, F., Saeed, N., & Khan, S. (2022). Impact of Covid pandemic and hybrid teaching on final year MBBS students' end of clerkship exam performance. *Pakistan Journal of Medical Sciences, 38*(1), 113-117.

Aljohani, M. (2017). Principles of "constructivism" in foreign language teaching. *Journal of Literature and Art Studies, 7*(1), 97-107.

Al-Sabaawi, M. Y. M., Dahlan, H. M., Shehzad, H. M. F., & Alshaher, A. A. (2021). A model of influencing factors of online social networks for informal learning in research institutes. *Social Network Analysis and Mining, 11*, 68.

Amineh, R. J., & Asl, H. D. (2015). Review of constructivism and social constructivism. *Journal of Social Sciences, Literature and Languages, 1*(1), 9-16.

Bychkovska, T. (2021). Effects of explicit instruction on noun phrase production in L2 undergraduate writing. *Journal of English for Academic Purposes, 54*, 101040.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589-597.

Cambridge International Examinations. (2017). *Cambridge IGCSE: A guide for universities*. Cambridge Assessment.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Dereso, C. W., Meher, K. C., & Shobe, A. A. (2022). Covid-19 pandemic and strategizing the higher education policies of public universities of Ethiopia. *International Journal of Sociotechnology and Knowledge Development (IJSKD), 14*(2), 1-16.

Don, Z. M., & Sriniwass, S. (2017). Conjunctive adjuncts in undergraduate ESL essays in Malaysia: Frequency and manner of use. *Moderna språk, 111*(1), 99-117.

Gupta, R. K., Kunhare, N., Pateriya, R. K., & Pathik, N. (2022). A deep neural network for detecting coronavirus disease using chest x-ray images. *International Journal of Healthcare Information Systems and Informatics (IJHISI), 17*(2), 1-27.

Hargreaves, A. (2021). What the COVID-19 pandemic has taught us about teachers and teaching, 6(1).

Hartshorn, K. J., & McMurry, B. L. (2020). The effects of the Covid-19 pandemic on ESL learners and TESOL practitioners in the United States. *International Journal of TESOL Studies, 2*(2), 140-156.

Huang, J., Xie, Y., Meng, Y., Shen, J., Zhang, Y., & Han, J. (2020). Guiding Corpus-based Set Expansion by Auxiliary Sets Generation and Co-Expansion. *Proceedings of The Web Conference 2020* (pp. 2188-2198).

Jamal, J., Shafqat, A., & Afzal, E. (2021). Teachers' perceptions of incorporation of corpus-based approach in English language teaching classrooms in Karachi, Pakistan. *Liberal Arts and Social Sciences International Journal (LASSIJ), 5*(1), 611-629.

Jo, C. W. (2022). Mapping adolescent literacy across L1 backgrounds: Linguistic and discourse features as predictors of persuasive essay quality. *System, 104*,102698.

Juan S. S., Ismail, M. F. C., Ujir, H., & Hipiny I. (2020). Language Modelling for a Low-Resource Language in Sarawak, Malaysia. In Z. Zakaria & R. Ahmad (eds.). Advances in Electronics Engineering. Springer. *Lecture Notes in Electrical Engineering, 619*, 147-158.

Li, J., & Jiang, Y. (2021). The Research Trend of Big Data in Education and the Impact of Teacher Psychology on Educational Development During COVID-19: A Systematic Review and Future Perspective. *Frontiers in Psychology, 12*.

Li, M., Wu, M., & Chen, X. (2021). Research on cooperative foreign language teaching mode based on multimedia network technology. *ACM International Conference Proceeding Series*, 400-403.

Lokker, C. & Jezrawi, R. (2022). Evaluating reflective writing to guide curricular improvements in health informatics education. *Reflective Practice, 23*(1), 44-56.

Muhammad, M. M., Janoory, L., Janan, D., & Jack, C. S. (2016). The diachronic analysis of english songs from 1960-2010: A corpus-based study. *Conference MPAC, 22*, 226.

Muhammad, M. M., Hamzah, S. G., Bin Abdullah, S. K., & Jack, C. S. (2017). A study of closure in a nursing textbooks and journals: A corpus based study. *International Journal of Advanced and Applied Sciences, 4*(2), 96-105.

Nair, S. M., & Hui, L. L. (2018). An Analysis of Common Errors in ESL Descriptive Writing among Chinese Private School Students in Malaysia. International *Journal of Education and Practice, 6*(1), 28-42.

Noble, H., & Heale, R. (2019). Triangulation in research, with examples. *Evidence-Based Nursing, 22*, 67-68.

O'Connor, A., & McCurtin, A. A. (2021). Feedback journey: Employing a constructivist approach to the development of feedback literacy among health professional learners. *BMC Medical Education, 21*, 486.

Putri, S. D. A., Ibrohim, M. O., & Budi, I. (2021). Abusive language and hate speech detection for Indonesian-local language in social media text. *Lecture Notes in Networks and Systems, 251*, 88-98.

Rüdiger, S. (2020). *Corpus approaches to social media*. John Benjamins Publishing Company.

Sidorov, G. (2019). *Syntactic n-grams in computational linguistics*. Springer.

Singh, C. K. S., Singh, A. K. J., Razak, N. Q. A., & Ravinthar, T. (2017). Grammar errors made by ESL tertiary students in writing. *English Language Teaching, 10*(5), 16-27.

Silva O., Sousa Á., & Nunes J. (2022). Technology's impacts in the students of higher education in the Covid-19 pandemic period. In A. Mesquita, A. Abreu & J. V. Carvalho (Eds.). Perspectives and Trends in Education and Technology. Springer. *Smart Innovation, Systems and Technologies, 256*, 183-194.

Taylor, C. (2018). *Corpus approaches to discourse*. Routledge.

Venkateswarlu, B., Shenoi, V. V., & Tumuluru, P. (2022). CAViaR-WS-based HAN: Conditional autoregressive value at risk-water sailfish-based hierarchical attention network for emotion classification in COVID-19 text review data. *Social Network Analysis and Mining, 12*(1), 10.

Wang, H. C., Hsiao, W. C., & Chang, S. H. (2020). Automatic paper writing based on a RNN and the TextRank algorithm. *Applied Soft Computing, 97*, 106767.

Wong, W. L., Muhammad, M. M., Chuah, K. P., Saimi, N., Ma'arop, A. H., & Elias, R. (2022). Did you run the Telegram? Use of mobile spelling checker on academic writing. *Multilingual Academic Journal of Education and Social Sciences, 10*(1), 1–19.

Wong, W. L., Muhammad, M. M., Singh, C. K. S., Ping, C. K., Saimi, N., & May, Y. S. (2022). The usage frequency and user-friendliness of online platforms among pre-university students during Covid-19 pandemic. *International Journal of Academic Research in Progressive Education and Development, 11*(1),26-37.

Wu, W., Bakirova, G., & Trifonov, I. (2021). A shift towards visualization in elearning. *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT), 16*(6), 1-12.

Zaini, M. F., Sarudin, A., Muhammad, M. M., & Bakar, S. S. A. (2020). Representatif Leksikal Ukuran sebagai Metafora Linguistik berdasarkan Teks Klasik Melayu. *GEMA Online® Journal of Language Studies, 20*(2), 168-187.

Zaini, M. F., Sarudin, A., Muhammad, M. M., Osman, Z., Redzwan, H. F. M., & Al-Muhsin, M. A. (2021). House building tips (HBT) corpus dataset as a resource to discover Malay architectural ingenuity and identity. *Data in brief, 36*, 107013.

Zamin, A. A. M., Elfeky, M., Kamarudin, R., & Abd Majid, F. (2019). A Corpus-based Study on the use of Phrasal Verbs in Malaysian Secondary School Textbooks. *International Journal of Applied Linguistics and English Literature, 8*(6), 76-85.

Zampieri, M. (2021). *Similar languages, varieties, and dialects*. Cambridge University Press.

Zhang, J., & Tao, H. (2018). *Corpus-based research in Chinese as a second language*. In The Routledge handbook of Chinese second language acquisition (pp. 48-62). Routledge.