Mahrajan Arabi
Proceeding International Conference in Arabic Festival (INCAFA) 2025

UTILIZATION OF CORPUS LINGUISTICS IN THE QATAR DEBATE BOOK: A STUDY OF HIGH FREQUENCY WORDS AND THE STRUCTURE OF ARABIC DEBATES

Muhammad Ahsan Thoriq, Hanik Mahliatussikah
Darul Faqih Indonesia, Malang, Indonesia
e-mail correspondence : ahsan.thoriq31@gmail.com

ABSTRACT

Vocabulary has an important role in learning a foreign language, but students have difficulty determining what vocabulary is used in the world of debate. This research is motivated by the needs of students learning Arabic debate. The purpose of this research is to produce and describe a vocabulary list with the frequency of its use, High Frequency Words-HFW, in the Qatar debate book. The method used in this study is a quantitative method to obtain research data and a qualitative method to analyze and conclude research results. The results of the study stated that there were 8,724 types of words (different words), and 110,624 tokens (number of occurrences of words). The results of this research are needed for Arabic students, teachers, writers, and debaters, especially in Indonesia. HFW is very useful in learning foreign languages, it is proven that it can produce special lists of words and terms that can be categorized according to the structure of delivery of Arabic debates so that students can understand vocabulary that is often used in debates.

Keyword: Linguistic Corpus; High Frequency Words; Qatar Debate Book

مستخلص البحث

تملك المفردات دورة مهمة لدى تعليم اللغة الأجنبية. لكن المشكلة هو صعوبة متعلمي اللغة في تعيين ما من مفردات يستخدمونها في عالم المناظرة. تستند هذه الدراسة إلى حاجة الطلاب المتعلمين للمناظرة باللغة العربية. وتهدف هذه الدراسة إلى إنتاج ووصف قائمة المفردات ذات التكرار العالي (High Frequency Words) في كتاب المناظرات القطري. وقد استُخدم في هذه الدراسة المنهج الكمي للحصول على بيانات البحث، والمنهج النوعي لتحليل النتائج واستنتاجها. وتشير نتائج الدراسة إلى وجود ٨٧٢٤ نوعًا من الكلمات (كلمات مختلفة)، و١١٠٦٢٤ تكرارًا للكلمات. وتُعد نتائج هذه الدراسة ضرورية للمتعلمين، والمعلمين، والكتاب، والمناظرين باللغة العربية، لا سيما في إندونيسيا. وتُعد الكلمات عالية التكرار مفيدة جدًا في تعلم اللغات الأجنبية، إذ ثبت أنها قادرة على إنتاج قائمة كلمات ومصطلحات يمكن تصنيفها وفقًا لبنية الإلقاء في المناظرة باللغة العربية، مما يُمكن المتعلم من فهم المفردات المستخدمة بشكل متكرر في المناظرات.

كلمات أساسية: مدونة لغوية؛ تكرار الكلمات العالية؛ كتب مناظرة قطر

Learning Arabic with the debate method is one of the keys to making Arabic experience an escalation of speakers around the world, this is due to several factors, one of which is the

direction of Arabic learning orientation which is turning into a global communication tool (Kementerian et al., 2019). In its implementation, debate requires four skills at once to hear, speak, read, and write (Thoriq et al., 2022). But, the basis of all these skills is vocabulary.

A learning will not be separated from the guidebook used (Siagian, 2020). The guidebook contains selected topics in which there are thousands of words used repeatedly that are even just one time to find. The words in the book elicits students' curiosity, especially about the meaning of words, so they can use them as needed.

Words become one of the important parts in the guidebook used in learning. Johns & Wilke (2018)) state that words are basic things that must be mastered. Lightbown & Spada (2017) reveal that communication is the art of using words placed in the right order, perfectly pronounced, or characterized by precise grammatical morphemes. Communication becomes tedious and even incomprehensible if we do not use the right words (Nafinuddin, 2020). Based on this opinion, it can be concluded into two things about the word, namely important and true.

Understanding of words with a high level of occurrence (frequency of use) in guidebooks can help learners understand guidelines independently or with guidance (Wahyuningtyas & Kesuma, 2021). Based on several studies of the role of a vocabulary in learning foreign-language shows that the use of high-frequency vocabulary (HFW) plays an important role in determining the success of foreign language learning (Siagian, 2020). Teachers sometimes focus on text according to the topics provided in the teaching material, so high-frequency words (which are in the manual) tend to be missed.

Basically, from these words can be taken advantage of as well as possible as a basic element in language learning (Rajeg, 2022). For example, in the Arabic debate book, speakers can obtain a structure of delivering a raw Arabic debate by native speakers. This delivery structure serves as a guide for the learner to find out what vocabulary is used by native speakers in Arabic debates, so that irregular communication does not occur in official Arabic debate events in Indonesia (Agung, 2020).

Based on the explanation above, the background of the study is a high-frequency word from the Qatar debate guidebook to bring up a standard Arabic debate delivery structure so that it can be used as a learning support and needs to be known and notified to learners and the use of corpus linguistics in determining high-frequency words from the book. Based on the background of the study, the research problem is how to utilize corpus linguistics in determining high-frequency words in the Qatar debate guidebook so as to obtain a standard Arabic debate delivery structure.

A corpus is a collection of texts, written or spoken, stored on a computer (O'Keeffe et al., 2007). Arum & Winarti (2020) corpus is said to be "natural" because the text collected is a text that is produced and used naturally and is not created (as it is) such as the handbook, the textbook, and many more. Based on this explanation, the corpus is a collection of texts (written or spoken) that are natural as in textbooks and stored on a computer.

Corpus linguistics is concerned with understanding how people use language in various contexts (Crawford & Csomay, 2015). Adolphs (in, Hizbullah et al., 2019) revealed that corpus

linguistics specifically examines language through a set of data that is natural, real as it is used, both written data and transcribed oral data. In this study, corpus linguistics was used to analyze a set of data recovered from the textbook "al-Mursyid fii Fanni al-Munadzarah" digitally or using software that runs on a computer..

In relation to the corpus there are several popular applications, one of which is AntConc which was developed by Laurence Anthony, a software program that is very useful for corpus linguistic analysis and is currently available for PC and Mac (Anthony, 2004). This application is for analyzing lexical and grammatical. Apart from these applications, there is also software developed by the Lexical Computing company founded in 2003 by lexicographer and research scientist Adam Kilgarriff, namely Sketch Engine. Its aim is to enable people who study language behavior (lexicographers, researchers in linguistic corpus, translators or language learners) to search for large collections of texts that are appropriate to complex and linguistically motivated questions (Kostka, 2022). Sketch Engine got its name after one of its main features, word sketching: a one-page, automatic, corpus derivative of grammatical behavior and word collocation (Kilgarriff et al., 2014). Currently, it supports and provides corpora in more than 90 languages.

Sketch Engine is able to facilitate several types of analysis such as KWIC (keyword in context', n-grams, collocates, and word lists. Word lists analysis can create simple word lists with the sum of frequencies and ratings from your own data set (Suchomel, 2021). In fact, San Martín & Trekker (2021) mention that Sketch Engine hosts a comprehensive set of tools, including a powerful concordancer, word and keyword frequency generators, tools for cluster and lexical set analysis, and word distribution plots.

Frequency analysis allows researchers to identify the words that occur most frequently in a given corpus, and then compare and contrast them with other words. Johns & Wilke (2018) states that a list of 100-200 high-frequency words would form more than 50% of the words in daily communication. Students who know the core of 200 or more high-frequency words with vision will have a solid basis for reading. It can be concluded that high-frequency words should be conveyed to language learners, especially foreign languages and second languages.

The high-frequency word in this study comes from the book 'Qatar Arabic debate guide' which is a guide for Arabic debate learners in the world. This book is a basic book for knowing debate strategies and the language used in debates. From this book, a structure of delivering Arabic debates is needed as a reference for learners to debate in a structured manner.

The research has been carried out by Ahsanuddin et al., (2020) On the development of the Corpus of Language, Literature, and Art. This research shows that the digital corpus that contains writing about language, literature, and art can be used well by final year students in compiling a thesis. The reason is that this development is very helpful for students to more easily find references in their research.

Research has been carried out by Thoriq et al., (2022) entitled Design and Build an Arabic Debate Corpus. The results of this study are in the form of the website of the Arabic Debate Corpus (KorDa). The data presented in the corpus is the result of a transcript from the Qatar

Debate championship video. The benefit of the study is that students can analyze the language of high-level word frequencies of native Arabic speakers.

The two studies are created a corpus website to be used as learning. The difference with this study lies in the use of the existing linguistic corpus to determine high-frequency words for language learning as a means of communication.

The study was conducted on the utilization of corpora by Syarofi & Nugraha (2022). The results of this research indicate that out of the 30 religious terms with the highest frequency, terms related to Islam had the highest frequency, followed by Hinduism, Christianity, Confucianism, and the lowest frequency was associated with Buddhism. These frequencies do not comprehensively reflect Indonesia's demographic constellation because they only represent Muslims as the majority group in Indonesia, while the frequencies of terms from other religions do not reflect the percentage of followers for each respective religion.

METHODS

The corpus data used for this study is a Qatar Arabic debate guidebook that has become a reference for Indonesian debaters. The data is considered valid because it represents a debate guidebook that guides debates around the world. The application used in analyzing data is Sketch Engine, this application was chosen because it detects Arabic more accurately and less messy.

This research was conducted in several stages: first, collecting Arabic debate guidebooks and selecting the Sketch Engine application. Second, re-type the data in word and change it in TXT after it classifies the data (erasing the punctuation) and presenting it as needed and continuing with the sorting of unnecessary vocabulary such as people's names, foreign vocabulary, and references. The next step is processing data using the Sketch Engine application to obtain a list of words with the highest to lowest frequency of occurrence. After that the final step is to analyze the data. The flow of this research is as follows:

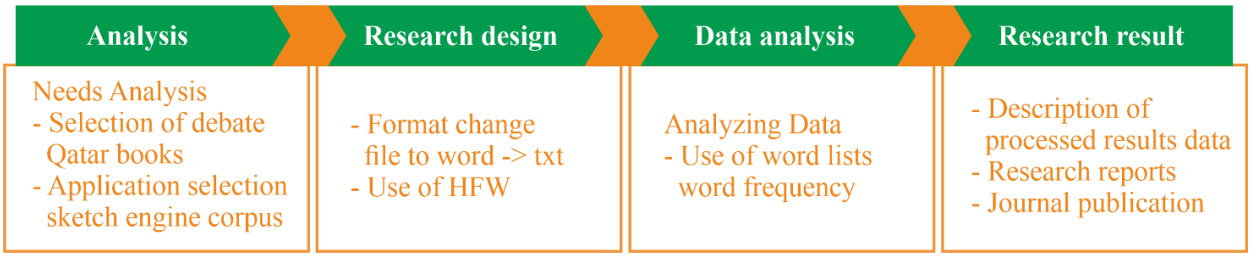


Figure 1. Research Flow

For data analysis using the linguistic corpus of the Sketch Engine application, researchers carried out several steps. The steps are as follows:

1. Choosing Arabic as the language of linguistic analysis
2. Uploaded files (PDF, word, rar, TXT) on the add menu 'recently used corpora'

3. Choose the Wordlist menu for high-frequency word analysis.
4. On the Bacis menu select 'word' and 'All' to analyze all the words that we have presented.
5. Find High frequency words

Data analysis was conducted to produce HFW that can be used by Arabic debate learners as material for producing correct sentences and according to the rules of native Arabic speakers and also finding the use of words in the standard Arabic debate delivery structure. Data analysis was conducted to produce HFW that can be used by Arabic debate learners as material for producing correct sentences and according to the rules of native Arabic speakers and also finding the use of words in the standard Arabic debate delivery structure.

RESULTS AND DISCUSSIONS

High Frequency Words

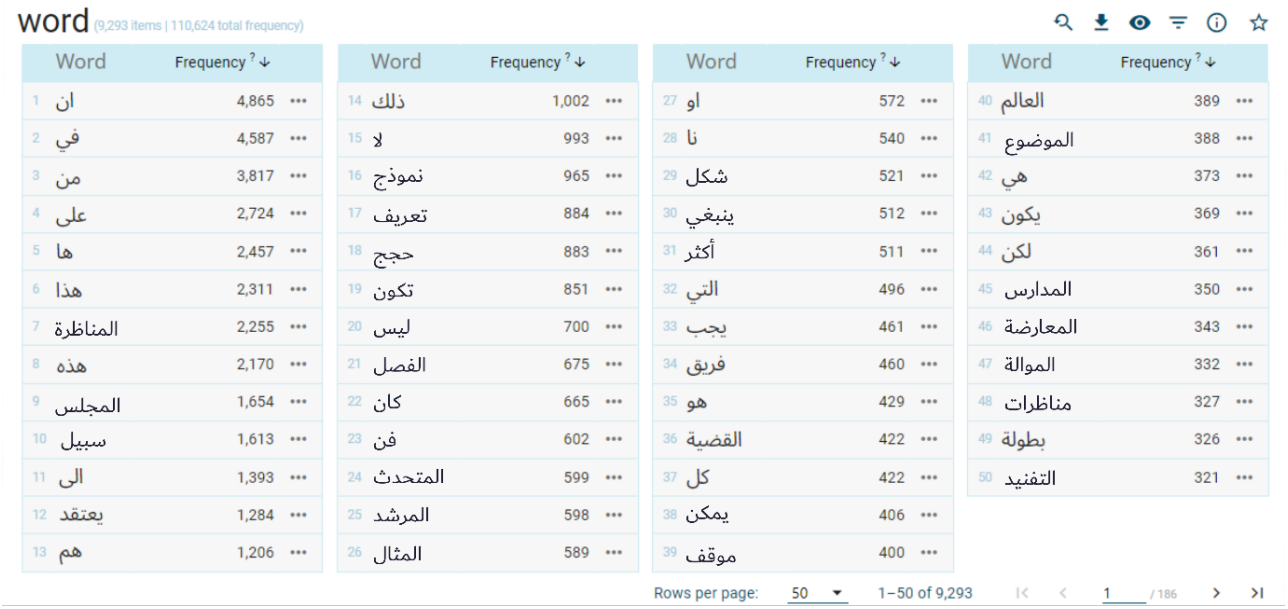
The Sketch Engine is a leading corpus tool widely used in lexicography and linguistic analysis. It is an online corpus query system that allows users to analyze the linguistic properties of pre-loaded corpora or explore their own using a set of in-built tools (Evans, 2022). The software has been used for discovering patterns of normal usage in various aspects of English, ranging from morphology to discourse and pragmatics, and has been found to be valuable in fulfilling different needs, including language study, teaching, writing, and translating (Crosthwaite & Baisa, 2023).

Sketch Engine provides a variety of analysis tools for linguistic exploration and research. It is an online corpus query system that allows users to analyze the linguistic properties of a range of pre-loaded corpora or to explore their own using a set of in-built tools (Wehrmeyer, 2023). The software is widely used in lexicography and is known for its powerful analytical functionalities, making it a first choice solution for publishers, universities, translation agencies, and national language institutes throughout the world. Some of the analysis tools and statistics offered by Sketch Engine include word or lemma frequency, part-of-speech frequency, n-gram frequencies, phrase and multiword frequency, high-quality keyword and term extraction, and co-occurrence analysis using linguistic criteria via the word sketch (Pang & Wang, 2023). These features make Sketch Engine a valuable resource for linguistic analysis and research.

High-frequency words play an essential role in corpus linguistics. They are words that appear frequently in a language, such as "the," "and," or "is," and their knowledge is essential for comprehension of academic spoken English. High-frequency words facilitate comprehension and signify the difficulty of a text, facilitating the customization of reading materials for learners (Dang et al., 2022). In corpus linguistics, word frequency aids in the identification of keywords that reflect a text's primary themes, which is essential for summarizing, indexing, and information retrieval tasks. Additionally, corpus analysis can be used to calculate word frequency in various ways, depending on the research questions and data, and can help to identify high-frequency and low-frequency words (Dang, 2021). High-

frequency word lists generated from corpora can also be used in the selection of vocabulary for second language teaching.

Through data processing according to the steps taken, there are 9,293 words types obtained using the Word List feature in the Sketch Engine application. After reducing the data, 8,724 different types of words were found and the number of occurrences of these words reached 110,624. The results of this data analysis present a list of words sorted by their level of use in the Qatar debate guidebook. There are 186 columns in the results of the analysis because the researcher made one column into 50 words so that each word can be seen clearly. The following is the data for the 50 words that have the highest frequency in the book which are explained in table 1.



word (9,293 items | 110,624 total frequency)

	Word	Frequency		Word	Frequency		Word	Frequency		Word	Frequency
1	ان	4,865	14	ذلك	1,002	27	او	572	40	العالم	389
2	في	4,587	15	لا	993	28	نا	540	41	الموضوع	388
3	من	3,817	16	نموذج	965	29	شكل	521	42	هي	373
4	على	2,724	17	تعريف	884	30	ينبغي	512	43	يكون	369
5	ها	2,457	18	حجج	883	31	أكثر	511	44	لكن	361
6	هذا	2,311	19	تكون	851	32	التي	496	45	المدارس	350
7	المناظرة	2,255	20	ليس	700	33	يجب	461	46	المعارضة	343
8	هذه	2,170	21	الفصل	675	34	فريق	460	47	الموالة	332
9	المجلس	1,654	22	كان	665	35	هو	429	48	مناظرات	327
10	سبيل	1,613	23	فن	602	36	القضية	422	49	بطولة	326
11	الى	1,393	24	المتحدث	599	37	كل	422	50	التفنيد	321
12	يعتقد	1,284	25	المرشد	598	38	يمكن	406			
13	هم	1,206	26	المثال	589	39	موقف	400			

Rows per page: 50 1–50 of 9,293 1 / 186

Table 1. 50 High Frequency Words in the Qatar Debate Handbook

The first column (far left) shows word ranking based on the number of words in the data presented sequentially. The more often the word appears, the ranking will be higher. The HFW list is suggested to be a mandatory vocabulary that must be mastered by novice debaters. It is believed that this will motivate the debater and give them a sense of optimism in producing their own sentences. Research on how to acquire a multiple vocabulary in order to communicate well and perfectly has been carried out. One of the factors that make new vocabulary easy to memorize is the frequency with which it is seen, heard, and understood. This is because the meaning of words can be done by seeing, hearing, and daily habits (Ramadan & Mulyati, 2020).

Exposure to high-frequency words increases the likelihood that learners will understand and remember the meanings of new words (Qin, 2023). Studies have shown that vocabulary acquisition is a function of frequency, and that learners need to be exposed to a word multiple times before achieving a significant effect on vocabulary knowledge acquisition (van Heuven et al., 2023). Additionally, repetition and multiple exposures to vocabulary items are important for vocabulary acquisition, and instruction of high-frequency words known and used by mature language users can add productively to an individual's language ability (Yeldham, 2022). Therefore, incorporating high-frequency words into vocabulary instruction and providing learners with multiple exposures to these words in various contexts can help improve vocabulary acquisition and retention.

The high-frequency words contained in the Qatari debate book are different from the results of high-frequency word study in the native Arabic debate competition held by Qatar. The study of HFW from the Qatar debate competition for native speakers was carried out by researchers by transcribing the discussions at the final of the Qatar debate competition in 2020 as representatives of other debaters. The following are ten HFWs based on the results of research conducted by researchers from Qatar debate books and Qatar debate competitions.

I	II
Qatar Debate Championship Vocabulary Frequency	High Frequency Words of Qatar Debate Book
1. القضية	1. أن
2. محور	2. في
3. أن	3. من
4. في	4. على
5. هدف	5. ها
6. ما	6. هذا
7. المناظرة	7. المناظرة
8. هذه	8. هذه
9. المجلس	9. المجلس
10. سبيل	10. سبيل

Table 2. Ten differences in WFH in debate books and debate competitions

From the table above, there are same four words, while the top six words are different. This shows that the words written in the book are very different from the language that emanates from the spoken Arabic directly. The language written in books is more formal and structured, while the spoken language seems informal. This is what researchers hope that novice debaters are more accustomed to using formal language as written in books for communication in Arabic debate competitions. So that the language is more structured and easily to understand by the judges.

The novice debaters may find it more comfortable to use formal language, as it can make them sound more professional and knowledgeable. Formal language can also help establish credibility and authority in a debate round (Zakaria et al., 2023). However, it is essential to strike a balance between formality and clarity, as using overly complex or obscure language can confuse the audience and hinder effective communication. Incorporating persuasive language in debates can be beneficial for novice debaters, as it can help move judges past what their argument would have done naturally, especially in rounds discussing individuals (Calafato, 2019). This can be achieved by expanding their vocabulary and using powerful words that resonate with the audience. It is crucial for novice debaters to understand that debate is just a structured way to have a conversation about choices, and they should approach debate rounds with the same mindset as they would when trying to convince someone on the street to do something. By focusing on clear, everyday language and logical reasoning, novice debaters can improve their performance and become more comfortable with the debate format (Liliarti & Kuswanto, 2018).

Debate Delivery Structure

Debate Delivery Structure is a crucial aspect of debating, as it involves organizing and presenting arguments in a clear and logical manner. A formal debate typically follows a specific structure, which includes the following components (Firoozi et al., 2019). The introduction is the first speech of the debate, in which the team presents the topic and the resolution (the statement to be debated). The first speaker of the affirmative team presents the arguments for the resolution. They should clearly state their position and provide logical reasoning and evidence to support their claims. The second speaker of the affirmative team responds to any counterarguments or rebuttals presented by the opposing team. They should address specific points raised by the opposing team and provide additional evidence or reasoning to strengthen their position. The first speaker of the opposing team presents the arguments against the resolution. They should clearly state their position and provide logical reasoning and evidence to support their claims, addressing the points made by the affirmative team. The second speaker of the opposing team responds to any counterarguments or rebuttals presented by the affirmative team. They should address specific points raised by the affirmative team and provide additional evidence or reasoning to strengthen their position. The conclusion is the final speech of the debate, in which the team summarizes their main points and restates their position on the topic. They should reinforce their arguments and provide a final reason for the audience to accept their viewpoint (Dbabis et al., 2018). In addition to these components, there are different models and methods for structuring a debate argument, such as the Toulmin model, the Rogerian approach, and the Monroe's motivated sequence. These models provide a

framework for organizing and presenting arguments in a clear and logical way, and they can be useful for structuring debates and persuasive writing.

The structure of the Arabic debate is an important element that allows debaters to speak in a structured manner, this structure becomes a scoring point for the judges to determine the better group. Today, structure has become a strategy for debaters to construct arguments. The structure that the researchers designed was obtained from Arabic debate guidebook.

The Arabic debate guidelines are centered on the Arabic debate guidelines published by Qatar. In Arabic debates there are certain levels according to ability and experience. In the Qatar Arabic debate manual, there are three levels of a debater, namely: beginner, intermediate, and expert (Quinn, 2009). The three levels are as follows:

a. Beginner: is a debater who has no experience in the field of Arabic debate, or has limited experience and no more than two years of learning Arabic debate.

b. Intermediate: is a debater who understands the basics of Arabic debate well and has studied Arabic debate for more than two years.

c. Expert: is a debater who understands the basics of Arabic debate perfectly and can always practice and hone his skills in Arabic debate well.

Meanwhile, the level chosen by the researcher is the beginner level because it is considered the most suitable for learning in Indonesia. The results of this high-frequency word study on the Qatar debate book found the right and standard words in conveying debate structures for the beginner level. This structure will later become a reference for producing words and constructing arguments in Arabic debates. The following is a table of word findings based on the debate delivery structure:

Standard Arabic debate delivery structure		
No	Section	Word
1	مجاملة ابتدائية	ما (15) إلا (13) شيء (10) عن (9) انقلب (5) إلى (3) ضدّه (3)
2	سلام واحترام	سيدي (9) العدل (7) رئيس (7) الحكام (5) المجلس (3)
3	مجاملة	كلنا (56) أن (43) إذا (39) نعلم (34) في (30) الملح (9) زاد (4)
4	تأييد الفريق	أننا (34) لهذا (29) صف (27) من (19) الموالة (18) نرى (9) ما (5)
5	تعريف	هو (43) لهذا (40) عن (33) أن (31) و (28) يجب (24) هي (20) في (17) إذا (9)
6	تقسيم أدوار المتحدثين في القضية	أن (31) قبل (28) نعرض (23) موقفنا (20) لهذا (17) اليوم (9)
7	موقف أو هدف	إذا (29) إلى (27) لأن (23) كما (18) تقليل (15) الملح (14) الفساد (11)
8	حجة :	لهذا (11) حجتنا (9) الأولى (7) اليوم (4)

	منطوق	إلى (27) الفساد (22) يؤدّي (11) الإنتاج (8) ضعف (3)
	تعليل	معي (45) إلى (38) من (26) تخيلوا (26) دولة (22) تسعى (22) التنمية (17)
	تدليل	أو (54) من (51) فيها (43) إذ (33) أن (30) مثل (27) له (21) أهله (19)
	توكيد	و (34) هذا (33) كله (26) إلى (23) التي (19) حجتنا (15) ضعف (14) الإنتاج
9	تأييد الفريق	لذلك (38) يا (30) سيدي (28) نحن (25) هنا (23) في (21) عند (20) ما (20)
10	سلام	و (6) السلام (6)

Table 3. Structure of Arabic Language Debate Presentation

Based on table 3, it can be found that the vocabulary that is often used in Arabic debates is a particle. This word is often used because it makes communication simple and easy to understand. For example, the particle ه which refers to a noun and the particle ل which is connected with the word يحقّ become an inseparable idiom.

The use of high-frequency words can enhance debate delivery by making arguments more persuasive and engaging. According to a resource from saskdebate.ca, persuasive language is crucial in debate, and the use of powerful words can improve a debater's style, winning them points with judges and making it easier for them to win. Expanding vocabulary with persuasive words can assist in moving judges past what the argument would have done naturally, particularly in rounds discussing individuals. Additionally, the use of persuasive language is most effective in rounds that are discussing individuals, and it can be critical to move judges in these instances. Incorporating high-frequency persuasive words into debate delivery can help debaters communicate their points more effectively and sway the audience in their favor. By using powerful language and persuasive words, debaters can enhance their style and make a more compelling case for their arguments. This can ultimately lead to a more successful debate performance.

CONCLUSIONS

The results of this study state that particles are the most contributors of high-frequency words (أن، في، من، على). So, it can be concluded that the list of words that have a high frequency is not only nouns and verbs. This is because particles make communication simpler and denser by becoming pronouns of nouns or idioms of verbs, thus making letter words more frequently used. Research that determines high-frequency words using corpus linguistics can be applied to written manuscripts (books, newspapers, magazines, etc.) or to oral manuscripts (audio and video) depending on needs.

This study can produce a list of specific words and terms that can be categorized according to the structure of the delivery of Arabic debates so that students can understand vocabulary that is often used in debates. In addition, this study describes words that are supported by facts in the form of numbers (frequency of occurrence).

REFERENCES

Agung, N. (2020). *Peningkatan Kemampuan Debat Bahasa Arab Mahasiswa Melalui Metode Suggestopedia*. *2*(1), 19–29.

Ahsanuddin, M., Ma'sum, A., & Ridwan, N. A. (2020). INVESTIGATING ARABIC CORPUS (KorSA) OF INDONESIAN UNDERGRADUATE THESIS ABSTRACTS. *Humanities & Social Sciences Reviews*, *8*(3), 920–927. https://doi.org/10.18510/hssr.2020.8396

Anthony, L. (2004). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, *7*(2), 7–13.

Arum, E. R., & Winarti, W. (2020). Penggunaan Linguistik Korpus dalam Mempersiapkan Bahan Ajar English For Specific Purpose Di Bidang Radiologi. *Jurnal Teras Kesehatan*, *2*(2), 58–69. https://doi.org/10.38215/jutek.v2i2.39

Calafato, R. (2019). The non-native speaker teacher as proficient multilingual: A critical review of research from 2009–2018. *Lingua*, *227*, 102700. https://doi.org/10.1016/j.lingua.2019.06.001

Crawford, W., & Csomay, E. (2015). *Doing Corpus Linguistics*. Routledge. https://doi.org/10.4324/9781315775647

Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, *3*(3), 100066. https://doi.org/10.1016/j.acorp.2023.100066

Dang, T. N. Y. (2021). High-frequency words in academic spoken English: Corpora and learners. *ELT Journal*, *74*(2), 146–155. https://doi.org/10.1093/ELT/CCZ057

Dang, T. N. Y., Webb, S., & Coxhead, A. (2022). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, *26*(4), 617–641. https://doi.org/10.1177/1362168820911189

Dbabis, S. Ben, Ghorbel, H., & Belguith, L. H. (2018). Sequential dialogue act recognition for Arabic argumentative debates. *Procesamiento Del Lenguaje Natural*, *60*(March 2019), 53–60. https://doi.org/10.26342/2018-60-6

Evans, M. (2022). Early Modern Digital Review Sketch Engine. *Renaissance and Reformation*, *44*(4), 217–223. https://doi.org/10.33137/rr.v44i4.38650

Firoozi, T., Razavipour, K., & Ahmadi, A. (2019). The language assessment literacy needs of Iranian EFL teachers with a focus on reformed assessment policies. *Language Testing in Asia*, *9*(1). https://doi.org/10.1186/s40468-019-0078-7

Hizbullah, N., Suryaningsih, I., Mardiah, Z., Al, U., & Indonesia, A. (2019). MANUSKRIP ARAB DI NUSANTARA Arabi : Journal of Arabic Studies. *Arabi: Journal of Arabic Studies*, *4*(1), 65–74.

Johns, J. L., & Wilke, K. H. (2018). High-Frequency Word: Some Ways to Teach and Help

Students Practice and learn Them. *TJLE: Texas Journal of Literacy Education*, *6*(1), 1–14.

Kementerian, A. R. I., Direktorat, M., & Direktorat, J. P. I. (2019). *Keputusan Menteri Agama Tentang Kurikulum PAI dan Bahasa Arab*. Kementerian Agama Republik Indonesia.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kostka, M. (2022). Pipeline Effectiveness in the Sketch Engine. *Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing*, 123–130.

Lightbown, P. M., & Spada, N. (2017). *How Languages are Learned* (Fourth). University Press.

Liliarti, N., & Kuswanto, H. (2018). Improving the competence of diagrammatic and argumentative representation in physics through android-based mobile learning application. *International Journal of Instruction*, *11*(3), 106–122. https://doi.org/10.12973/iji.2018.1138a

Nafinuddin, S. (2020). *Pengantar Semantik (Pengertian, Hakikat, dan Jenis)*. osf.io.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. University Press, Cambridge.

Pang, S., & Wang, K. (2023). Sketching the changing patterns in kaleidoscopes: New developments in corpus-based studies of translation features. *Research in Corpus Linguistics*, *11*(2), 79–102. https://doi.org/10.32714/ricl.11.02.05

Qin, M. X. (2023). A corpus-based approach to Chinese English lexis. *World Englishes*, *42*(2), 308–326. https://doi.org/10.1111/weng.12531

Rajeg, G. P. W. (2022). Kajian konkordansi korpus terhadap perilaku konstruksional makna literal dan metaforis pasangan verba sinonim pandang dan tatap. *Rinarxiv.Lipi.Go.Id*, *November 2021*.

Ramadan, S., & Mulyati, Y. (2020). Makna Kata dalam Bahasa Indonesia (Salah Kaprah dan Upaya Perbaikannya). *Ranah: Jurnal Kajian Bahasa*, *9*(1), 90. https://doi.org/10.26499/rnh.v9i1.1036

San Martín, A., & Trekker, C. (2021). Adapting word sketches for specialized knowledge extraction. *Proceedings of the 14th International Conference of the Asian Association for Lexicography (ASIALEX)*, 64–87.

Siagian, E. N. (2020). *High Frequency Words in Indonesian as Foreign Language at Beginers Level Esra Nelvi Siagian Badan Pengembangan dan Pembinaan Bahasa Kosakata memiliki peranan penting dalam pembelajaran bahasa kedua atau bahasa asing . Agar bahasa asing yang dipelajari dap*. *9*(2), 188–201. https://doi.org/https://doi.org/10.26499/rnh.v9i2.2320

Suchomel, V. (2021). Genre Annotation of Web Corpora: Scheme and Issues. *Proceedings of*

the Future Technologies Conference (FTC), *1*, 738–754. https://doi.org/10.1007/978-3-030-63128-4_55

Syarofi, A., & Nugraha, A. P. (2022). ANALISIS FREKUENSI PENGGUNAAN ISTILAH KEAGAMAAN DALAM BUKU AJAR BAHASA INDONESIA SMA KELAS X-XII: KAJIAN BERBASIS KORPUS. *Jurnal Pendidikan Bahasa Dan Sastra Indonesia Metalingua*, *7*(2), 123–130. https://doi.org/10.21107/metalingua.v7i2.16853

Thoriq, M. A., Ahsanuddin, M., & Suryadarma, Y. (2022). *Taṣmīm Mudawwanah Al-Munāẓarah Al-Arabiyyah 'ala Asās Al-Sam'iyyah Al-Baṣariyyah*. *8*(1), 109–127.

van Heuven, W. J., Payne, J. S., & Jones, M. W. (2023). SUBTLEX-CY: A new word frequency database for Welsh. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218231190315

Wahyuningtyas, D., & Kesuma, T. M. J. (2021). Pemanfaatan Linguistik Korpus dalam Menentukan Kata Berfrekuensi Tinggi pada Buku Sahabatku Indonesia BIPA 1. *Jurnal Bahasa …*, *3*(1), 60–69.

Wehrmeyer, E. (Ed.). (2023). *Advances in Sign Language Corpus Linguistics* (Vol. 8). John Benjamins Publishing Company. https://doi.org/10.1075/scl.108

Yeldham, M. (2022). Second language English listeners' relative processing of coherence-based and frequency-based formulas: A corpus-based study. *Applied Linguistics Review*, *13*(2), 287–317. https://doi.org/10.1515/applirev-2018-0093

Zakaria, A., Renandya, W. A., & Aryadoust, V. (2023). A Corpus Study of Language Simplification and Grammar in Graded Readers. *LEARN Journal: Language Education and Acquisition Research Network*, *16*(2), 130–153.