UsingHadith Corpus in Learning Arabic as a Second Language

Wan Muhammad Wan Sulong[1], **Rabihah Mohd Ghazali**[2]

[1]Faculty of Modern Languages and Communication, Universiti Putra Malaysia
[2]Teacher Education Institute, Islamic Studies Campus, Selangor, Malaysia

_____

Abstract: The Hadith corpus is a rich linguistic and literary resource for Muslims, as well as a religious and legislative resource. The Hadith corpus contains the Prophet's discourse, which holds a special place in Muslims' hearts. It contains the Prophet's (p.b.u.h.) words, actions, sayings, resolve, moral and ethical characteristics. The Hadith corpus is a vast resource and a vast treasure for the Arabic language, which can be useful while learning Arabic as a foreign language (Arabic L2 henceforth). The purpose of this study is to discuss a decent way to use the Hadith corpus in Arabic L2 to meet the demands of non-native speakers. Three significant corpora are used to analyse Hadith texts namely sunnah.com, islamarchive.cc, and islamweb.net. Based on the findings, this research reveals that if the Hadith text offered meets the learners' needs, it could be a valuable tool for Arabic L2. For students to understand the text correctly, it must also be annotated. In this case, Hymes' 'SPEAKING' model, which considers many contextual aspects that are assumed to be involved in speaking, might be utilised as a foundation for annotating the text's social and situational components. These processes will aid Arabic L2 in comprehending the true meaning of words or text-linguistic forms that may differ from their lexical meanings. Examining the context also enables the investigation of the Arabic language's social and cultural values and conventions.

Keywords: Hadith corpus, Prophet's discourse, sociolinguistic variation, corpus linguistic, natural language processing
_____

Introduction

The introduction of corpus linguistics, a relatively young discipline of science, has re-energized the field of modern linguistics. As indicated by Adolphs (2006), this discipline analyses language in particular through a data collection that is natural, real, and relevant to its application, comprising both written and transcribed speech data. According to its definition, a corpus is a written collection of data, both ordinary and digital data, containing various linguistic information at the level of words, structures, meanings, and discourses that can be used for studies.

As defined by Sinclair (1991), a corpus is a collection of data from many genres used to investigate various aspects of linguistics in defining the real phenomena of language use. According to Cook (2003), the term "collection" refers to a collection of digital texts collected based on specified criteria, whether in written or oral form, then kept and processed on a computer for the aim of linguistic research.

Furthermore, according to Nesselhauf (2011), a corpus is a text derived from the practical and natural use of spoken and written language that has been collected systematically. The purpose of systematic collection is to prepare the structure and content of the text-based on aspects outside the language, such as time, place, and others, and to facilitatefurther research.

The corpus created as a result of the digitization process can then be used in a variety of linguistic fields, including morphology, syntax, semantics, lexicology, and other linguistic fields such as translation, literature, and so on. There is already a wide range of linguistic data collected in many of the world's languages. In practice, the use of corpus linguistics has been greatly expanded and strengthened in other languages. Arabic is no exception, though its development may not be as rapid as the compilation of other languages' corpora. The available language corpus is derived not only from modern languages, but also from classical languages by utilising written documents that still exist, can be found, and studied(Nur Hizbullah, Fazlurrahman, Fuzi Fauziah, 2016).

According to Cheng (2012), corpus linguistics is an empirical method of linguistic analysis and description used to examine the language used by natural speakers. This method is becoming more popular because it can be used to examine various linguistic problems in a broad sense and can even reveal interesting, fundamental, and even surprising things from existing linguistic data.Artificial Intelligence (AI) and Natural Language Processing (NLP) developments have led to attempts to automate a variety of operations that typically require domain specialists. The study of religious texts to improve understanding and uncover new embedded information, as well as to use them as a tool in learning Arabic as a second language, is one of the research fields that has attracted AI researchers' interest. The goal of this study is to suggest a method for creating a Hadith corpus enabling non-native Arabic speakers to acquire Arabic as a second language. It is proposed that the variation approach in sociolinguistics be used to present

the content of the Hadith corpus in a way that will help Arabic L2 learners grasp a natural style of conversing in Arabic.

Development of Hadith Corpus

There are two primary sources of reference in Islam: the Quran and hadiths. The Quran is a collection of Allah's words revealed to the Prophet Muhammad PBUH. Hadith refers to the chronicle of Prophet Muhammad PBUH's acts, statements, and quiet approbation. The 'isnad' and the'matn' are the two fundamental components of hadith. The term "isnad" refers to the chain of narrators who passed down the hadith, whereas "matn" refers to the context or body of the hadith. Hadith is classified by Muslim clerics into three categories:*sahih* (genuine), *hasan* (excellent), and *da'if* (weak) (Siti Syakirah Sazali, 2020).

The Quran has received more attention as a primary source. Al-Rabia has taken one step toward compiling the general corpus of Arabic, which includes the Quran as one of the sources (et.al., 2014). The King Saud University-Corpus of Classical Arabic is the name of the corpus (KSUCCA). As the name implies, they are only interested in the corpus of Arabic from the classical period, and the Quran is assumed to be from that period. They reasoned that because the classical Arabic pattern is the foundation of Arabic linguistic theory, it should be regarded as a credible source. This corpus contains 50 million words from classical Arabic literature and is used to study the lexical-semantic distribution of the Qur'an and classical Arabic using a computational linguistic approach, among other things. However, in addition to describing the structure of the corpus, their study also describes some aspects of language change from classical to modern times.

Although the importance of Hadith is second only to that of the Quran, most laws and legislations are derived from it due to its broader scope and incorporated details. However, Islamic computational studies have primarily focused on the Quran, leaving Hadith largely unexplored. According to Altammami, 2020, one possible reason is Hadith's vast and varying literature with inconsistent structure, which makes collecting them in a well-structured corpus a difficult task.Hadith computation research is still in its early stages (Bounhas, 2019). Nonetheless, as mentioned by Azmi et al., 2019, there is an annual increase in the number of published papers, indicating that it is gaining wider attention from multidisciplinary researchers.

Sunnah.com is one of the enormous human efforts devoted to compiling Hadith corpus and aligning Arabic Hadiths with their English translations. Several websites host Hadith books and compile a massive collection of hadith. Islamarchive.cc, for example, contains over 430,000 hadiths from 238 approved hadith books. Another example of a useful website is islamweb.net, which contains a wealth of Islamic resources, including Hadith.These three sources are used as material in this study to evaluate how the hadith text is presented, as well as to suggest an appropriate approach for how the hadith text is presented to Arabic L2learners.

Position of Hadithamong non-Arabic speaking Muslims

Hadith, or the Prophet's discourse, is one of the most appropriate texts to teach non-native Muslim speakers in Arabic L2 for the following reasons:

1- The position of religious discourse: The Prophet's discourse occupies a high place in the hearts of Muslims as a second legislative source after the Holy Qur'an. They have familiarized themselves with it and pledged to listen and study it and love to understand it and act upon it because they have long ago paid their attention to it in terms of education and learning in schools, mosques, institutes, and universities. The importance of studying the language of the Prophet's discourse is that it helps them understand Islamic teachings and the meanings of Arabic words used in acts of worship. Understanding Islamic knowledge is one of the primary goals that drive Muslims to study the Arabic language. The study conducted by Lubna Abdul Rahman et al., 2007 involving 49 participants in the short-term Arabic language course at the International Islamic University Malaysia and the International Islamic University Institute of Selangor (KUIS) indicated the objectives of their study of the Arabic language.93.9% of them answered that their goal in studying the Arabic language is to fulfill the requirements of worship, which mostly use the Arabic language, and to deepen Islamic teachings. 44.9% of them answered that it meets the needs of the times, while 28.6% of them meet the professional requirements and 22.4% to acquire the skill in multiple languages.

2- Loving the Prophet(p.b.u.h.): In the hearts of Muslims, Prophet Muhammad (p.b.u.h.)is held in high regard. They have a strong desire to learn about his life and follow his Sunnah because it leads them down the right path. Without Sunnah, it is impossible to understand Islam. Because of their intense love for him and a desire to learn his Sunnah,

his speech become one of the subjects that should be used in learning the Arabic language, because their love for the owner of the speech compels them to learn his speech and language.

3- The Prophet's eloquence: God Almighty bestowed eloquence of speech upon His Messenger, and he is regarded as the most eloquent of Arabs in speech, the clearest in speech, and the highest in eloquence. This feature compels Muslims to study his words and emulate his discourse and speaking styles.

Hadith Corpus for Arabic L2

If the manner the Hadith corpus is presented satisfies the needs of the learners, it could be a useful tool for Arabic L2. Furthermore, the text must be annotated for learners to understand it correctly. The sociolinguistic approach of variation can be employed in this circumstance. Chambers (2002) defined sociolinguistic variation as the process of studying language use in natural social settings and identifying linguistic variants according to their social distribution. The importance of language variation research rests in its capacity to show how speakers adapt their speech to situations that include diverse aspects such as audience, setting, and themes. As a result, we can identify specific language patterns employed by specific individuals of society at a given moment, while also reflecting dominant sociocultural traits for the speech community at that time.

In this context, Dell Hymes proposed the "ethnography of speaking" (1972) descriptive model of speech communication and established its components based on the interaction between language and other aspects of social life, which results in a variety of language environments. Hymes (1986) proposes a method for defining such social aspects of discourse and their relationship to the various language styles used in that discourse. It also seeks to highlight the importance of such a relationship between language style and the context of a situation in achieving communication goals.He established a framework that considers a variety of factors that are thought to play a role in speaking. S for Setting and Scene, P for Participants, E for Ends, A for Act of Sequence, K for Key, I for Instrumentalities, N for Norms of Interaction and Interpretation, and G for Genres are the acronyms Hymes employs for the many aspects he considers to be relevant (Hymes, 1986). What Hymes gives in his 'SPEAKING' formula, as Wardhaugh, 2002 points out, is an essential reminder that communication is a complex activity and that each given amount of talk is a piece of skilled work. It is skillful in the sense that the speaker must demonstrate sensitivity to and understanding of each of the eight factors to be successfulin communication. (Sulong, Abdussalam, 2013)

To adopt Hymes' approach to Hadith corpus in Arabic L2, the presentation of the Hadith corpus offered to learners should consider the following processes:
First:Rearranging the dialogue texts in the prophetic discourse to make them easier to read for learners and to make the interlocutors more visible. Here are some instances of dialogue texts from the Prophet's discourse before and after they were arranged:

حَدَّثَنَا هَدَّابُ بْنُ خَالِدٍ الْأَزْدِيُّ حَدَّثَنَا هَمَّامٌ حَدَّثَنَا قَتَادَةُ حَدَّثَنَا أَنَسُ بْنُ مَالِكٍ عَنْ مُعَاذِ بْنِ جَبَلٍ قَالَ: كُنْتُ رِدْفَ النَّبِيِّ صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ لَيْسَ بَيْنِي وَبَيْنَهُ إِلَّا مُؤْخِرَةُ الرَّحْلِ فَقَالَ: يَا مُعَاذَ بْنَ جَبَلٍ! قُلْتُ: لَبَّيْكَ رَسُولَ اللهِ وَسَعْدَيْكَ. ثُمَّ سَارَ سَاعَةً ثُمَّ قَالَ يَا مُعَاذَ بْنَ جَبَلٍ قُلْتُ لَبَّيْكَ رَسُولَ اللهِ وَسَعْدَيْكَ ثُمَّ سَارَ سَاعَةً ثُمَّ قَالَ يَا مُعَاذَ بْنَ جَبَلٍ قُلْتُ لَبَّيْكَ رَسُولَ اللهِ وَسَعْدَيْكَ قَالَ هَلْ تَدْرِي مَا حَقُّ اللهِ عَلَى الْعِبَادِ قَالَ قُلْتُ اللهُ وَرَسُولُهُ أَعْلَمُ قَالَ فَإِنَّ حَقَّ اللهِ عَلَى الْعِبَادِ أَنْ يَعْبُدُوهُ وَلَا يُشْرِكُوا بِهِ شَيْئًا ثُمَّ سَارَ سَاعَةً قَالَ يَا مُعَاذَ بْنَ جَبَلٍ قُلْتُ لَبَّيْكَ رَسُولَ اللهِ وَسَعْدَيْكَ قَالَ هَلْ تَدْرِي مَا حَقُّ الْعِبَادِ عَلَى اللهِ إِذَا فَعَلُوا ذَلِكَ قَالَ قُلْتُ اللهُ وَرَسُولُهُ

The following is the text as it appears on sunnah.com, islamarchive.cc, and islamweb.net:
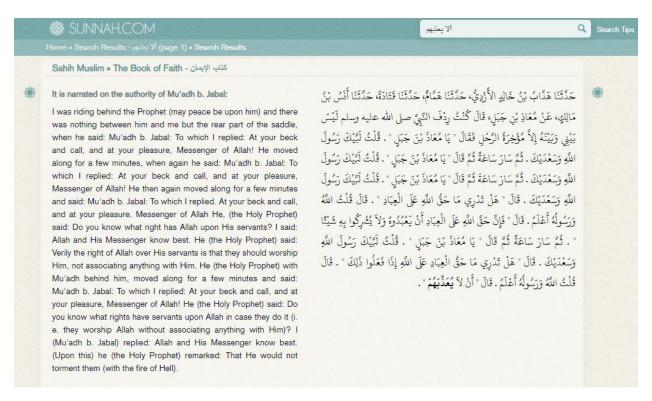
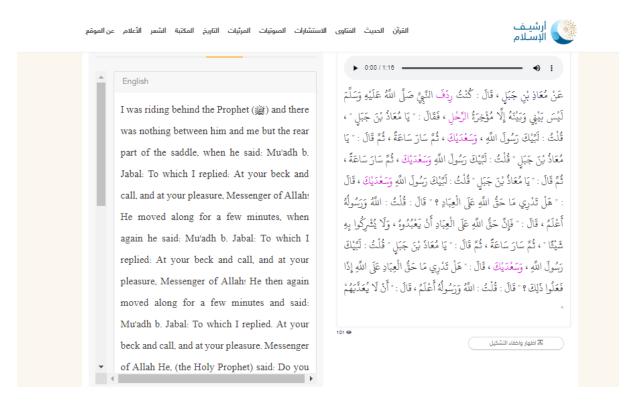Figure 1: The text in sunnah.com

Figure 2: The text inislamarchive.cc

Figure 3: The text inislamweb.net

Figures 1, 2, and 3 show the arrangement of the text in all three sources. The proposed arrangement is as follows:

عَنْ مُعَاذِ بْنِ جَبَلٍ قَالَ: "كُنْتُ رِدْفَ النَّبِيِّ صَلَّى اللّٰهُ عَلَيْهِ وَسَلَّمَ لَيْسَ بَيْنِي وَبَيْنَهُ إِلَّا مُؤْخِرَةُ الرَّحْلِ،

1. فَقَالَ (رسولُ الله) : "يَا مُعَاذَ بْنَ جَبَلٍ!"،

2. قُلْتُ (أي معاذ) : "لَبَّيْكَ رَسُولَ اللّٰهِ وَسَعْدَيْكَ"، [ثُمَّ سَارَ سَاعَةً]،

3. ثُمَّ قَالَ: "يَا مُعَاذَ بْنَ جَبَلٍ!"،

4. قُلْتُ: "لَبَّيْكَ رَسُولَ اللّٰهِ وَسَعْدَيْكَ" [ثُمَّ سَارَ سَاعَةً]،

5. ثُمَّ قَالَ: "يَا مُعَاذَ بْنَ جَبَلٍ!"،

6. قُلْتُ: "لَبَّيْكَ رَسُولَ اللّٰهِ وَسَعْدَيْكَ"،

7. قَالَ: "هَلْ تَدْرِي مَا حَقُّ اللّٰهِ عَلَى الْعِبَادِ؟"،

8. قَالَ: قُلْتُ: "اللّٰهُ وَرَسُولُهُ أَعْلَمُ"،

9. قَالَ:"فَإِنَّ حَقَّ اللّٰهِ عَلَى الْعِبَادِ أَنْ يَعْبُدُوهُ وَلَا يُشْرِكُوا بِهِ شَيْئًا" [ثُمَّ سَارَ سَاعَةً]،

10. قَالَ: "يَا مُعَاذَ بْنَ جَبَلٍ!"،

11. قُلْتُ: "لَبَّيْكَ رَسُولَ اللّٰهِ وَسَعْدَيْكَ"،

12. قَالَ: "هَلْ تَدْرِي مَا حَقُّ الْعِبَادِ عَلَى اللّٰهِ إِذَا فَعَلُوا ذَلِكَ؟"،

13. قَالَ: "قُلْتُ: "اللّٰهُ وَرَسُولُهُ أَعْلَمُ"،

14. قَالَ: "أَنْ لَا يُعَذِّبَهُمْ".

Translation:

> It is narrated on the authority of Mu'adh bin Jabal: I was riding behind the Prophet (p.b.u.h.) and there was nothing between him and me but the rear part of the saddle, when he said:
>
> 1. P: Mu'adh bin Jabal! [To which I replied]:
> 2. M: At your beck and call, and at your pleasure, Messenger of Allah! [He moved along for a few minutes, when again he said]:
> 3. P: Mu'adh b. Jabal! [To which I replied]:
> 4. M: At your beck and call, and at your pleasure, Messenger of Allah! [He moved along for a few minutes, when again he said]:
> 5. P: Mu'adh b. Jabal! [To which I replied]:
> 6. M: At your beck and call, and at your pleasure, Messenger of Allah! [He (the Holy Prophet) said]:
> 7. P: Do you know what right has Allah upon His servants? [I said]:
> 8. M: Allah and His Messenger know best. [He (the Holy Prophet) said]:
> 9. P: Verily the right of Allah over His servants is that they should worship Him, not associating anything with Him. [He (the Holy Prophet), with Mu'adh behind him moved along for a few minutes and said]:
> 10. P: Mu'adh bin Jabal! [To which I replied]:
> 11. M: At your beck and call, and at your pleasure, Messenger of Allah! [He (the Holy Prophet said]:
> 12. P: Do you know what rights have servants upon Allah in case they do it? [I replied]:
> 13. M: Allah and His Messenger know best. [(Upon this) he remarked]:
> 14. P: That He would not torment them (with the fire of hell).

The process of rearranging the text is as follows:
1- The chain of narrators and the majority of the narrators' names have been removed because they are unrelated to the learning L2 objectives, and only the narrator who participated in the speech with the Messenger of God(p.b.u.h.)has remained.
2- Sentences that are not from the dialogue are placed in square brackets to distinguish them from the dialogue.
3- Each sentence containing a dialogue is numbered.

Second:Select texts that contain words that they are familiar with and are free of strangeness. Words related to belief, worship, and morals, for example, are among those that Muslims are accustomed to hearing.

Third: Annotate the text with pieces of information linked to social andsituational context. In this regard, the Hymes SPEAKING model can be adopted. The context components that are thought to be useful in understanding the text can be presented as illustrated in the table below:

Table 1: Proposed contextual components of speech event to be annotated in Hadith text

Hadith	Components of speech events	Description
Hadith X	S-Scene and Setting	The time and place of speech, which are the specific physical circumstances in which it occurs.
	P-Participants	Various speaker-listener, addressor-addressee, or sender-receiver combinations
	E-End	The purpose or reason of the speaking
	Act of Sequence	It refers to the actual form and content of what is said.
	Key	The way the speaking is performed (tone, manner, delivery)
	Instrumentalities	The mode of communication used
	Norms of interaction and interpretation	The social rules of what is proper in conversation
	Genre	It refers to distinct types of speeches, such as poems,

		proverbs, riddles, sermons, prayers, and lectures.

The significance of the speech context stems from the fact that prophetic discourse is a social discourse whose methods can only be fully comprehended by looking at the setting in which it took place. It was necessary for the Messenger (p.b.u.h.) to speak eloquently. As a result, considering the context of speech occurrences can assist in elucidating the true meaning of words or linguistic forms that may differ from their lexical meanings. Furthermore, examining the context allows for the exploration of the social and cultural values and customs thatwere prevalent in the Prophet's era.

Conclusion

In summary, the findings of the study are as follows:

1- Hadith corpus, which comprises the Prophet's discourse, is beneficial to Arabic L2 learners because it is one of the most eloquent and brilliant literary texts after the Holy Qur'an. Furthermore, the Prophet's discourse is one of the pillars upon which the principles of the Arabic language, rhetoric, and other Arabic disciplines are built.

2- The prophetic discourse was so accurately narrated that it resembled a recorded discussion using modern technological tools that record interviews and conversations today. The Prophet's speech became rich material and a huge treasure for Arabic subjects, and it was worthy of attention for Arabic language teachers to use it in teaching non-native learners of Arabic L2 styles of speech and communication.

3- Correct and sound educational approaches should be used while using Hadith corpus for non-native learners of Arabic L2, such as supplying appropriate text organization and selecting texts that contain terms and structures appropriate for their level.

4- Paying attention to the contextual components of speech, such as social and situational context, to help Arabic L2 learners understand the correct meaning of the discourse and its application in everyday discourse. By analyzing situational circumstances surrounding speech occurrences, the variation approach in examining the prophetic style of discourse is capable of developing the relationship between language and society, as well as bridging ways of speaking with specific social identities.

References

Adolphs, S. (2006) *Introducing electronic text analysis - A practical guide for language andliterary studies*. New York: Routledge

Alrabiah, M, Alhelewh, N, Al-Salman, A et al. (1 more author) (2014) *An empirical study on the Holy Quran based on a large classical Arabic corpus.* International Journal of Computational Linguistics (IJCL), 5 (1). 1 - 13. ISSN 2180-1266

Altammami, S., Atwell, E., Alsalka, A. (2020). Constructing a bilingual Hadith corpus using a segmentation tool. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3390–3398 Marseille, 11–16 May 2020

Azmi, A. M., Al-Qabbany, A. O., and Hussain, A.(2019). Computational and natural language processing based studies of hadith literature: a survey.*Artificial Intelligence Review*,52, 1369–1414.

Bounhas, I. (2019). On the usage of a classical Arabic corpus as a language resource: related researchand key challenges. *ACM Transactions on Asianand Low-Resource Language Information Processing(TALLIP)*, 18(3):23.

Chambers, J.K. (2002). Studying language variation: An informal epistemology. In Chambers, J.K., Trudgill,Peter, Estes,Natalie Schilling (Eds.), *The handbook of language variation and change* (pp.3-140). UK: Blackwell Publishers.

Cheng, W. (2012). *Exploring Corpus Linguistics: Language in Action*. Oxon: Routledge

Cook, G. (2003). Applied linguistics: Oxford introductions to language study (H. G. Widdowson, Series Ed.).Oxford: Oxford University Press.

Hymes, D. (1986). Models of the interaction of language and social life. In Gumperz, J. P., Hymes, D. (Eds.), *Directions in sociolinguistics, the ethnography of communication* (pp.50-66). UK: Basil Blackwell.

Lubna Abdul Rahman, Zalika Adam, Ku Fatahiyah Ku Azizan. (2007). Kesedaran Orang Awam terhadap Kepentingan Bahasa Arab: Kajian dalam Kursus Bahasa Arab (Arabic Short Course), in *Proceedings of*

Malaysia International Conference on Foreign Language 2007. Serdang: Department of Foreign Languages, Faculty of Modern Languages and Communications, UPM. Pp. 144-149.

Muslim, Ibn al-Hajjaj. (2006). *Sahih Muslim*. Riyadh: Dar Taybah.

Nesselhauf, N. (2011).*Corpus linguistics:A practical introduction.* Retrieved fromhttp://www.as.uniheidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf

Nur Hizbullah, Fazlurrahman, Fuzi Fauziah (2016).Linguistik korpus dalam kajian danpembelajaran bahasa Arab di Indonesia. *Prosiding Konferensi Nasional Bahasa Arab II*.Universitas Negeri Malang, Malang.Pp.385-393

S. S. Sazali, N. A. Rahman and Z. A. Bakar. (2020). Characteristics of Malay translated hadith corpus.*Journal of King Saud University– Computer and Information Sciences*, https://doi.org/10.1016/j.jksuci.2020.07.011

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sulong, W.M.W, Abdussalam, A.S. (2013). Variation Approach in Investigating Socio-cultural Aspects of the Discourse of Prophet Muhammad (P.B.U.H). In Eriko Yamato (Ed.),*Studies on Foreign Languages and Cultures in Malaysia* (pp. 144-163). Serdang: UPM Press.

Wardhaugh, Ronald. (2002). *Introduction to Sociolinguistics*. Oxford: Blackwell Publishers Ltd.