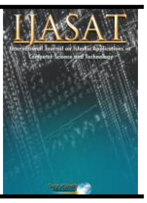


The Arabic–English Parallel Corpus of Authentic Hadith

Shatha Altammami[1,2,a], Eric Atwell[2,b] Ammar Alsalka[2,c]

[1]King Saud University (KSU), Saudi Arabia
[2]University of Leeds, UK

[a]Scshal@leeds.ac.uk, [b]E.S.Atwell@leeds.ac.uk, [c]M.A.Alsalka@leeds.ac.uk

Abstract

We present a bilingual parallel corpus of Islamic Hadith, which is the set of narratives reporting different aspects of the Prophet Muhammad's life. The Hadith collection is extracted from the six canonical Hadith books which possess unique linguistic features and patterns that are automatically extracted and annotated using a domain-specific tool for Hadith segmentation. In this article, we present the methodology of creating the corpus of 39,038 annotated Hadiths which will be freely available for the research community.

Keywords**:** Hadith, Parallel Corpus, NLP, Language Resource.

1. Introduction

One of the emerging trends of artificial intelligence (AI) and natural language processing (NLP) is their use in multidisciplinary research. This phenomenon is evident in the constant increase of scientific papers that describe computerising domain-specific tasks that range from cancer detection to opinion mining. One of the research areas that has sparked interest in AI methods is the study of religious texts to enhance understanding and discover new embedded knowledge. However, the main obstacle of such studies is the lack of annotated corpora suitable for religious-oriented text-mining tasks.

In this work, we aim to enrich an under-resourced religious texts, Islamic Hadith, which is the set of narratives reporting the words, actions and habits of the Prophet Muhammad. Although Hadith’s importance is second to the Quran’s (the Muslim holy book), most laws and legislation are obtained from Hadith due to its larger scope and incorporated details. Yet, Islamic computational studies have focused on the Quran, leaving Hadith relatively unexplored. One possible reason is Hadith’s vast and varying literature with inconsistent structure that makes collecting them in a well-structured corpus a challenging task.

Although there is research in the area of Hadith computation, it is still in its infancy, with limit contributions as concluded in recent surveys (Bounhas, 2019) (Azmi, Al-Qabbany, & Hussain, 2019). Researchers gather their own dataset from different sources and sometimes manually process them (Luthfi, Suryana, & Basari, 2018). This indicates that the field is lacking adequate language resources and reusability is limited, since the collected datasets are not published for use in other research projects. Hence, it is impossible to establish benchmarks, compare results or set evaluation measures (Guellil, Saâdane, Azouaou, Gueni, & Nouvel, 2019), which makes establishing a Hadith Common Dataset Initiative an imperative.

Through this work, we initiate the Hadith Common Dataset by introducing a freely available parallel corpus of Hadith in its original classical Arabic text and its corresponding English translations obtained from well-known Hadith books. To the best of our knowledge, no parallel corpus of Hadith is freely available to the research community. The accessible data is scattered around the web in an unstructured format. In fact, resources regarding Classical Arabic text constitute only 11% of the available Arabic resources (Guellil et al., 2019).

We named this language resource the *Leeds and King Saud University (LK) Hadith corpus* to represent the collaboration between King Saud University and the University of Leeds. The corpus will be released as part of this submission via a University of Leeds repository[1]. In the next few lines, we give a brief overview of Hadith and its literature. Then we discuss related work of existing corpora and elaborate on differences and extra features incorporated within our LK corpus. After that, we describe the methodology with which we collected data and built a corpus. Finally, we evaluate the corpus and discuss future directions for enlarging it.

2. Background

Muslims believe the Quran is God's divine words, which enjoined them to follow the guidance of Prophet Muhammad in their laws, legislation and moral guidance. This clear instruction to emulate the Prophet and follow his judgements is necessary because not all Islamic laws and regulations are mentioned in the Quran.

This act of reporting the different aspects of the Prophet's life became known as Hadith, which is an Arabic word for *speech*, *report* or *narrative*. Hadith types vary, perhaps being a short sentence or long paragraph describing what the Prophet said in a specific incident, the Prophet's conversation with someone or a story told by the Prophet's companions that explains the Prophet's actions in a specific matter like prayers.

Unlike the Quran, Hadith was not documented immediately after the Prophet's death. Instead, it was passed down the generations verbally by scholars, each mentioning the person from whom they heard the Hadith. However, some dishonest people deliberately have fabricated material and ascribed it to the Prophet. This led to the development of Hadith science, in which scholars study the chain of narrators and their biographies to accept or reject the Hadith teaching, the process of which formed the unique structure of Hadith.

Hadith consists of two parts, as shown in Figure 1. The *Isnad* is shown in bold, representing the reverse chronological chain of narrators followed by the *Matn*, which is the actual teaching. The *Isnad* can be translated to mean *support*, since it is used to identify the authenticity of Hadith following the narrator's genealogy. It is a meta-data that is useful for authenticity but does not add useful information to the context of the actual narration (*Matn*). Therefore, in designing our corpus, it is crucial to separate the *Isnad* from the *Matn* to allow researchers to focus on their text of interest.

It is worth noting that the *Isnad* segment is not definite in all Hadiths since some Hadiths consists of irregular structures. Nonetheless, the *Isnad* is meant to facilitate identifying the Hadith authenticity by studying the genealogy of the narrators who reported the Prophet’s words or actions. Hence, in our corpus we incorporated the Prophet in the *Matn* instead of *Isnad*. This is to ensure when the *Matn* components are extracted from the corpus, it is clear whether the words are of the Prophet or his companion as shown in Figure 8.

[1] https://doi.org/10.5518/480.

حَدَّثَنَا يَحْيَى بْنُ بُكَيْرٍ، حَدَّثَنَا اللَّيْثُ، عَنْ عُقَيْلٍ، عَنِ ابْنِ شِهَابٍ، قَالَ أَخْبَرَنِي أَنَسُ بْنُ مَالِكٍ، أَنَّ رَسُولَ اللَّهِ صلى الله عليه وسلم قَالَ " مَنْ أَحَبَّ أَنْ يُبْسَطَ لَهُ فِي رِزْقِهِ، وَيُنْسَأَ لَهُ فِي أَثَرِهِ، فَلْيَصِلْ رَحِمَهُ ".

Yahya bin Bakir told us that Alith told him from Aqeel from Ibn Shihab who said Anas bin Malik told me that the prophet peace be upon him (PBUH) said, "Whoever loves that he be granted more wealth and that his lease of life be prolonged then he should keep good relations with his kith and kin.

Figure 1: Hadith example, Isnad in bold

3. Related Work

There are many existing Arabic corpora (Atwell, 2018). However, we are only interested in those that include Hadith or classical Arabic text in general. Although there were attempts to collect a Hadith corpus, researchers are still forming their own data, which suggests the non-existence of a well-structured common resource dedicated to Hadith (Bounhas, 2019). For example, there are large corpora which incorporate Hadith (Al-Thubaity, 2015). The KSU 50-million-word corpus of classical Arabic is designed to help researchers understand the use of words during the period of Quran revelation (Alrabiah, 2013).

There are other corpora which incorporate Hadith books, namely the Historical Arabic Corpus, or HAC (Hammo, Yagi, Ismail, & AbuShariah, 2016), which contains 45 million words from different time periods. Moreover, Tashkeela (Zerrouki & Balla, 2017) is a 76-million-word vocalised corpus of text that represents classical and modern Arabic books.

Another interesting project called the Open Islamicate Texts Initiative (OpenITI) is an international collaboration that incorporates other projects under its umbrella, including KITAB. They aim to incorporate Persian and other languages, forming a very large Islamic corpus (Belinkov, Magidow, Barrón-Cedeño, Shmidman, & Romanov, 2018).

A Hadith corpus was recently published with the same aim as ours. It includes Hadith from four books scraped from different websites that cover several languages like Arabic, English and Urdu (Mahmood, Ullah, K., Ramzan, & Ilyas, 2018). Our corpus is different since it is an Arabic–English parallel corpus of the six canonical Hadith books. However, we might investigate extending our corpus by merging their corpus with ours and applying AI to align the Urdu translations.

Alosami presented a Sunnah Arabic Corpus in their PhD as an annotated linguistic resource which comprises 144,000 words extracted from the Riyadu Assalihin Hadith book (Abdulrahman Alosaimy & Atwell, 2017). Although it is a valuable work, the corpus is relatively small since Hadiths are extracted from only one book.

In addition, a recent study surveyed and enumerated the freely available Arabic corpora. It mentions the existence of one Hadith corpus; however, it was not accessible and was not mentioned or used in the literature (Zaghouani, 2017). This indicates a common problem where a dataset is lost. It occurs when researchers share data on personal websites that become absolute after time. Therefore, we attempt to mitigate that by sharing our corpus on the University of Leeds repository. The below table highlights the difference between our corpus and existing ones.

Table 1: Corpora Comparison

Corpus	Hadith Only	All Canonical Books	Isnad Segmented from Matn	Parallel	Available Online
(Al-Thubaity, 2015)		x			x
(Alrabiah, 2013)		x			x
(Hammo et al., 2016)		x			x
(Zerrouki & Balla, 2017)		x			x
(Belinkov et al., 2018)		x			x
(A Alosaimy & Atwell, 2017)	X			x	
(Mahmood et al., 2018)	X		X		x
(LK) Hadith corpus	*X*	*x*	*X*	*x*	*x*

4. Corpus Data Source

In the Islamic literature, there are six canonical Hadith books known as Al-Sihah al-Sittah, which translate as 'The Authentic Six' and are considered the most reliable. These collections include Hadiths that cover almost all aspects of life in providing proper guidance in Islam, ranging from methods of performing prayers to elaborating on monetary transaction guidelines. These canonical books are Sahih Bukhari, Sahih Muslim, Sunan Abu Dawood, Sunan Altarmithi, Sunan Ibn Maja and Sunan Al-Nasai, and they form the basis for Islamic Hadith books. Although they are called the ‘Authentic Six’, not all incorporated Hadiths possess the same degree of authenticity. Consequently, they were called the ‘Authentic Six’ based on the majority of Hadiths in these books (Khan, 1987).

The canonical books organise Hadiths into a topology of topics into three levels, which we called Book – Chapter – Section. Each chapter is dedicated to one theme. Within the chapter, there are several sections that the author used to indicate a ruling on specific matters, given the incorporated Hadiths as evidence. The structure of these books is illustrated in Figure 2. Each Hadith consists of two parts, *Isnad* and *Matn*, and some books add a comment by the author, usually regarding the authenticity of the Hadith.

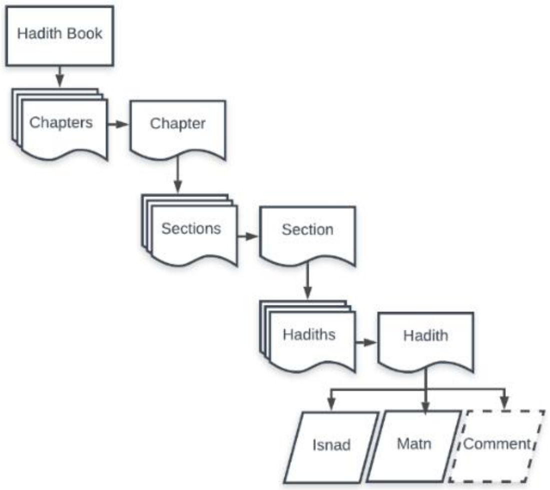
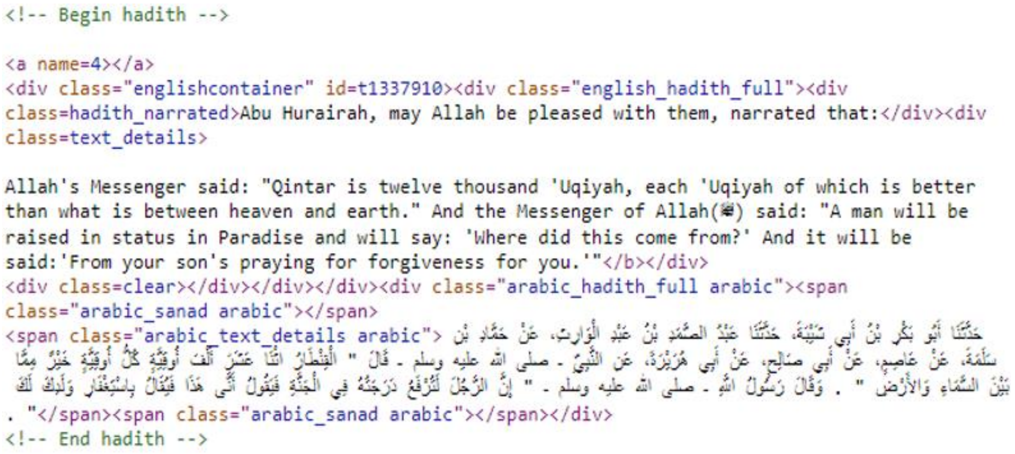


Figure 2: Canonical Hadith books structure

We wanted to maintain this structure in our corpus. Hence, we sought electronic sources of Hadith that followed this structure. Several websites host Hadith books; however, they did not meet our requirements. For example, *aHadith.co.uk* contains the English version of the Hadith with the section and chapter titles removed. Another example of valuable websites is *islamweb.net*, which hosts a huge number of Islamic resources, including Hadith. However, it does not fulfil our purpose of a parallel corpus of English- and Arabic-aligned Hadiths. Only *sunnah.com* met our needs. It maintained the structure of the books, and the English translation is aligned in parallel with the Arabic Hadith at the narrative level.

5. Corpus Collection and Creation

We developed software to scrape *sunnah.com* pages and extracted information from every Hadith. However, HTML tags were not used consistently. This could be due to its being built by a group of web developers. For example, the Arabic *Isnad* is not separated from the *Matn* in most Hadiths, despite the existence of an HTML tag <*Arabic_sanad arabic*> dedicated to *Isnad*, as shown in Figure 3.



```
<!-- Begin hadith -->

<a name=4></a>
<div class="englishcontainer" id=t1337910><div class="english_hadith_full"><div
class=hadith_narrated>Abu Hurairah, may Allah be pleased with them, narrated that:</div><div
class=text_details>

Allah's Messenger said: "Qintar is twelve thousand 'Uqiyah, each 'Uqiyah of which is better
than what is between heaven and earth." And the Messenger of Allah(ﷺ) said: "A man will be
raised in status in Paradise and will say: 'Where did this come from?' And it will be
said:'From your son's praying for forgiveness for you.'"</b></div>
<div class=clear></div></div></div><div class="arabic_hadith_full arabic"><span
class="arabic_sanad arabic"></span>
<span class="arabic_text_details arabic"> حَدَّثَنَا أَبُو بَكْرِ بْنُ أَبِي شَيْبَةَ، حَدَّثَنَا عَبْدُ الصَّمَدِ بْنُ عَبْدِ الْوَارِثِ، عَنْ حَمَّادِ بْنِ
سَلَمَةَ، عَنْ عَاصِمٍ، عَنْ أَبِي صَالِحٍ، عَنْ أَبِي هُرَيْرَةَ، عَنِ النَّبِيِّ - صلى الله عليه وسلم - قَالَ " الْقِنْطَارُ اثْنَا عَشَرَ أَلْفَ أُوقِيَّةٍ كُلُّ أُوقِيَّةٍ خَيْرٌ مِمَّا
بَيْنَ السَّمَاءِ وَالأَرْضِ " . وَقَالَ رَسُولُ اللَّهِ - صلى الله عليه وسلم - " إِنَّ الرَّجُلَ لَتُرْفَعُ دَرَجَتُهُ فِي الْجَنَّةِ فَيَقُولُ أَنَّى هَذَا فَيُقَالُ بِاسْتِغْفَارِ وَلَدِكَ لَكَ
. "</span><span class="arabic_sanad arabic"></span></div>
<!-- End hadith -->
```

Figure 3: HTML example of one Hadith in Sunnah.com

To overcome this, a Hadith segmentation tool was developed (Altammami, Atwell, & Alsalka, 2019) to automatically segment *Isnad* from *Matn*. Using this tool, we could maintain a consistent segmentation of *Isnad* and *Matn*.

5.1 Hadith Segmentation Tool

In the following lines, we briefly describe the functionality of the second version of our Hadith segmenter. We plan to elaborate on the details of this second version in another article. The Hadith segmenter pipeline is shown in Figure 4, where it applies the following:

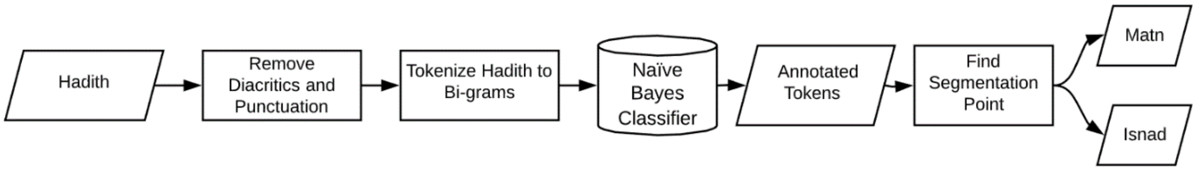


Figure 4: Processes of Hadith Segmenter

1. The first step is to pre-process the Hadith to remove diacritics, punctuations and extra white spaces.
2. Then it tokenises the pre-processed Hadith into bigrams of words.
3. After that, it labels every token as *Isnad* or *Matn* by using a Naive Bayes Classifier trained on 4,000 segmented Hadiths.
4. Finally, a rule-based algorithm is applied to segment the Hadith based on the labelled bigrams.

In this version, we addressed the limitation of the first version by identifying cases of Hadith with irregular structures. For example, some Hadiths consist of parallel *Isnad*, where there are two chain of narrators followed by Prophetic words (*Matn*). The first version segmented Hadith at the first chain and incorporated the second chain with the *Matn*. We enhanced the algorithm to deal with such cases and segmented it correctly, as illustrated in Figure 5, below.

Isnad	Matn
حدثنا مسدد قال حدثنا يحيى عن شعبة عن قتادة عن أنس رضي الله عنه عن النبي صلى الله عليه وسلم وعن حسين المعلم قال حدثنا قتادة عن أنس	عن النبي صلى الله عليه وسلم قال لا يؤمن أحدكم حتى يحب لأخيه ما يحب لنفسه
Mosadad said Yahya told us Shoba heard Qatada from Anas may Allah be pleased with him, that he heard the Prophet (PBUH), and from Husayn al-Muallim said Qatada told us that Anas said that	the Prophet (PBUH) said: "No one of you becomes a true believer until he likes for his brother what he likes for himself".

Figure 5: Hadith with parallel *Isnad*

5.2 Corpus Annotation

After scrapping the contents of *sunnah.com*, we started building our corpus by annotating Hadith components and its meta data, including book, chapter, section and Hadith number. To annotate the Arabic Hadith, we fed it into the segmenter tool to extract *Isnad* and *Matn*. Finally, every Hadith and its meta data are saved in a record where they are separated by commas. Hence, the CSV (comma separated values) file format is used with UTF-8 encoding. Such annotation could be easily converted to XML format that can be used across different systems. Every CSV file contains the following information for every Hadith record. An example of how one Hadith record is captured in a CSV file is broken down for readability in Figures 6, 7 and 8.

- Chapter Number: The chapter number where the Hadith is listed.
- Chapter English: Title of the chapter in English.
- Chapter Arabic: Title of chapter in Arabic.
- Section Number: The section number where the Hadith is listed.
- Section English: Title of the section in English.
- Section Arabic: Title of the section in Arabic.
- Hadith number: The sequential number of the Hadith.
- English Hadith: The whole English Hadith consists of *Isnad* and *Matn*.
- English *Isnad*: The name of the first narrator in English.
- English *Matn*: The actual Hadith teaching in English.
- Arabic Hadith: The whole Arabic Hadith consists of *Isnad* and *Matn*.
- Arabic *Isnad*: The chain of narrators in Arabic.
- Arabic *Matn*: The actual Hadith teaching in Arabic.
- Arabic Comment: An optional value that contains the scholar’s comment on the authenticity of the Hadith.
- English Grade: The degree of authenticity in the transliteration.
- Arabic Grade: The degree of authenticity in Arabic.

Chapter Number	Chapter English	Chapter Arabic	Section Number	Section English	Section Arabic	Hadith Number
10	The Book on Jana"iz (Funerals)	كتاب الجنائز عن رسول الله صلى الله عليه وسلم	1	What Has Been Related About Reward For The Sick	باب مَا جَاءَ فِي ثَوَابِ الْمَرِيضِ	966

Figure 6: Example of Hadith record extracted from Sunan Tarmizi Chapter 10 – Part 1

English Hadith	English Isnad	English Matn	English Grade
Aishah narrated that: The Messenger of Allah said: "The believer is not afflicted by the prick of a thorn or what is worse (or greater) than that, except that by it Allah raises him in rank and removes sin from him."	Aishah narrated that:	The Messenger of Allah said: "The believer is not afflicted by the prick of a thorn or what is worse (or greater) than that, except that by it Allah raises him in rank and removes sin from him."	Sahih

Figure 7: Continued Example of Hadith Record – Part 2

Arabic Hadith	Arabic Isnad	Arabic Matn	Arabic Comment	Arabic Grade
حَدَّثَنَا هَنَّادٌ، حَدَّثَنَا أَبُو مُعَاوِيَةَ، عَنِ الأَعْمَشِ، عَنْ إِبْرَاهِيمَ، عَنِ الأَسْوَدِ، عَنْ عَائِشَةَ، قَالَتْ قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم لاَ يُصِيبُ الْمُؤْمِنَ شَوْكَةٌ فَمَا فَوْقَهَا إِلاَّ رَفَعَهُ اللَّهُ بِهَا دَرَجَةً وَحَطَّ عَنْهُ بِهَا خَطِيئَةً . قَالَ وَفِي الْبَابِ عَنْ سَعْدِ بْنِ أَبِي وَقَّاصٍ وَأَبِي عُبَيْدَةَ بْنِ الْجَرَّاحِ وَأَبِي هُرَيْرَةَ وَأَبِي أُمَامَةَ وَأَبِي سَعِيدٍ وَأَنَسٍ وَعَبْدِ اللَّهِ بْنِ عَمْرٍو وَأَسَدِ بْنِ كُرْزٍ وَجَابِرِ بْنِ عَبْدِ اللَّهِ وَعَبْدِ الرَّحْمَنِ بْنِ أَزْهَرَ وَأَبِي مُوسَى . قَالَ أَبُو عِيسَى حَدِيثُ عَائِشَةَ حَدِيثٌ حَسَنٌ صَحِيحٌ .	حَدَّثَنَا هَنَّادٌ، حَدَّثَنَا أَبُو مُعَاوِيَةَ، عَنِ الأَعْمَشِ، عَنْ إِبْرَاهِيمَ، عَنِ الأَسْوَدِ، عَنْ عَائِشَةَ، قَالَتْ	قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم لاَ يُصِيبُ الْمُؤْمِنَ شَوْكَةٌ فَمَا فَوْقَهَا إِلاَّ رَفَعَهُ اللَّهُ بِهَا دَرَجَةً وَحَطَّ عَنْهُ بِهَا خَطِيئَةً	. قَالَ وَفِي الْبَابِ عَنْ سَعْدِ بْنِ أَبِي وَقَّاصٍ وَأَبِي عُبَيْدَةَ بْنِ الْجَرَّاحِ وَأَبِي هُرَيْرَةَ وَأَبِي أُمَامَةَ وَأَبِي سَعِيدٍ وَأَنَسٍ وَعَبْدِ اللَّهِ بْنِ عَمْرٍو وَأَسَدِ بْنِ كُرْزٍ وَجَابِرِ بْنِ عَبْدِ اللَّهِ وَعَبْدِ الرَّحْمَنِ بْنِ أَزْهَرَ وَأَبِي مُوسَى . قَالَ أَبُو عِيسَى حَدِيثُ عَائِشَةَ حَدِيثٌ حَسَنٌ صَحِيحٌ .	صحيح

Figure 8: Continued Example of Hadith Record – Part 3

The LK corpus structure is simple and resembles the original structure of the books. An illustration of the corpus is shown in Figure 9, where the ‘LK Hadith Corpus’ folder contains six folders representing the six canonical Hadith books. Within these folders, the CSV files represent the chapters in the book. For example, we created 97 CSV files under the Sahih Bukhari folder, which represent the number of chapters in the Sahih Bukhari book. The first CSV file is named *Chapter1.csv*, and it contains seven Hadith records.

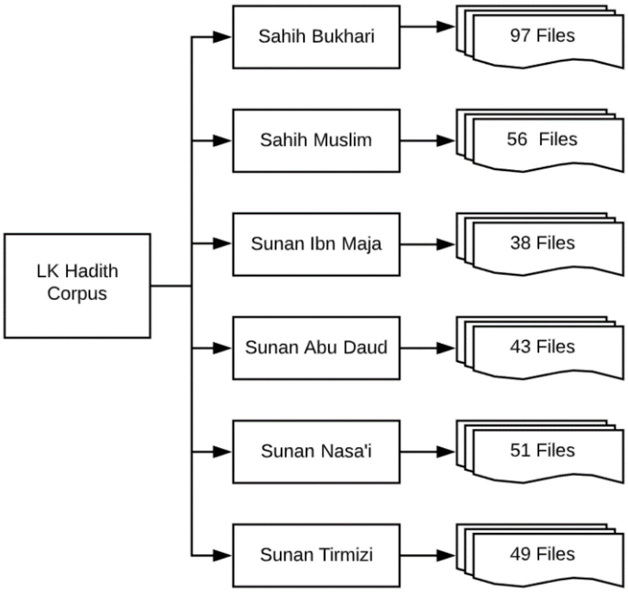


Figure 9: LK Hadith Corpus Structure

6. Corpus Content

The corpus includes 33,359 Hadith records of Arabic and an aligned English translation, making more than 10 million tokens. The number of tokens in the English part is larger than the Arabic. However, the Arabic part is richer in word type, as shown in Table 2. In Table 3, we demonstrate that the word frequencies of *Isnad* and *Matn* are representative in both languages.

Table 2: Corpus Statistics

	Chapter	Section	Hadith	Isnad	Matn	Comment	Grade	Total
Ar-Token	108,023	190,998	2,204,952	788,466	1,363,001	78,058	34,544	4,768,042
En-Token	159,474	300,182	2,501,072	298,665	2,203,070	-	35,799	5,498,262
Ar-Type	264	10,626	61,998	9,411	59,953	4,642	653	147,547
En-Type	467	9,866	43,547	7,375	9,411	-	48	70,714

Table 3: *Isnad* and *Matn* Word Frequencies

	Word	Frequency	Word	Frequency	Word	Frequency	Word	Frequency
Isnad	Narrated	24031	Abdllah	4651	بن	94616	عبد	22529
	That	13328	Umar	2747	عن	84976	الله	22500
	Said	12788	Abbas	2155	حدثنا	63156	أخبرنا	10961
	Bin	8790	Hraira	4117	قال	31375	سعيد	7447
	Reported	6248	Aisha	1518	أبي	26597	يحي	6949
Matn	He	64355	Man	5343	الله	87903	الناس	3206
	I	37090	Came	5745	عليه	45855	يوم	3085
	Allah	33525	Do	5751	صلى	43724	رجل	2781
	Messenger	25684	Asked	4405	وسلم	42417	الصلاة	2207
	Him	19744	Water	4328	قال	40724	الجنة	1578

Following the initial compilation of the dataset, manual intervention was necessary to clean up inconsistencies. We started with Sahih Bukhari, where we checked every Hadith against the PDF version of the book. We found very few mistakes in which a Hadith was placed under the wrong section or the English translation was for another Hadith, which is normal since human efforts are susceptible to mistakes.

Therefore, our Hadith corpus relies on the source. In other words, missing values or inconsistencies with the original book are dependent on the source. So far, we have checked Sahih Bukhari against the PDF version of the book, and we can safely say it is the gold standard of our corpus.

7. Potential Uses

The main objective of this project is providing the research community with a Hadith corpus that is well-annotated for diverse research purposes. Hence, every component is annotated to be easily extracted. Once researchers have access to a common dataset, it is possible to set benchmarks and compare results. Following are potential uses of the corpus:

1- To build ontologies that support Hadith authenticity by focusing on *Isnad*. Such systems could be tested using *Isnad* extracted from the LK Hadith corpus.
2- To use *Isnad* in a system that automatically draws the tree of narrators and where they lived to show how Hadith travelled through time and space.
3- To apply AI methods to *Matn* to automatically link Hadith to the Quran without Isnad affecting the results.
4- To use the English translation as training data for Hadith machine translation.

8. Conclusion

We have presented the creation of Leeds and King Saud University (LK) Hadith corpus using a domain-specific tool to segment and annotate Hadith components. This corpus is publicly available for researchers in Hadith computational studies to compare their findings and set benchmarks which are currently lacking. In the future, we plan to extend this corpus to include Hadith commentaries aligned with the Hadith at the narrative level and possibly include the translations of Hadiths in other languages.

9. Acknowledgement

The first author would like to thank King Saud University for sponsoring her PhD studies at the University of Leeds through the Saudi Cultural Bureau.

10. References

Al-Thubaity, A. O. (2015). A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-014-9284-1

Alosaimy, A, & Atwell, E. (2017). Sunnah Arabic Corpus: Design and Methodology. *Proceedings of the 5th International Conference on Islamic Applications in Computer Science and Technologies (IMAN 2017),* (December), 26–28. Retrieved from http://eprints.whiterose.ac.uk/125569/

Alosaimy, Abdulrahman, & Atwell, E. (2017). Tagging Classical Arabic Text using Available Morphological Analysers and Part of Speech Taggers. *Journal for Language Technology and Computational Linguistics, 32*(1), 1–26. Retrieved from http://www.jlcl.org/2017_Heft1/01-MorphosyntacticTagging.pdf

Alrabiah, M. (2013). The design and construction of the 50 million words KSUCCA.

Proceedings of WACL'2 …. https://doi.org/https://doi.org/10.1016/j.iac.2007.09.004

Altammami, S., Atwell, E., & Alsalka, A. (2019). Text Segmentation Using N-grams to Annotate Hadith Corpus. *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, 31–39.

Atwell, E. (2018). *Using the Web to model Modern and Qurʾanic Arabic*. Edinburgh University Press.

Azmi, A. M., Al-Qabbany, A. O., & Hussain, A. (2019). Computational and natural language processing based studies of Hadith literature: a survey. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-019-09692-w

Belinkov, Y., Magidow, A., Barrón-Cedeño, A., Shmidman, A., & Romanov, M. (2018). Studying the History of the Arabic Language: Language Technology and a Large-Scale Historical Corpus. Retrieved from http://arxiv.org/abs/1809.03891

Bounhas, I. (2019). On the Usage of a Classical Arabic Corpus as a Language Resource : Related Research and Key Challenges. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, *18*(3), 1–45.

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (2019). Arabic Natural Language Processing: an overview. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.02.006

Hammo, B., Yagi, S., Ismail, O., & AbuShariah, M. (2016). Exploring and exploiting a historical corpus for Arabic. *Language Resources and Evaluation*, *50*(4), 839–861. https://doi.org/10.1007/s10579-015-9304-9

Khan, S. H. (1987). *Al-Hitta Fi Dhikr Al-sihah Al-sitta*. Beiru.

Luthfi, E. T., Suryana, N., & Basari, A. H. (2018). Digital Hadith authentication: A literature review and analysis. *Journal of Theoretical and Applied Information Technology*.

Mahmood, A., Ullah, H., K., F., Ramzan, M., & Ilyas, M. (2018). A Multilingual Datasets Repository of the Hadith Content. *International Journal of Advanced Computer Science and Applications*, *9*(2), 165–172. https://doi.org/10.14569/IJACSA.2018.090224

Zaghouani, W. (2017). Critical Survey of the Freely Available Arabic Corpora. *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme,LREC*, 1–8. https://doi.org/10.13140/RG.2.1.1362.1284

Zerrouki, T., & Balla, A. (2017). Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems. *Data in Brief*, *11*, 147–151. https://doi.org/10.1016/j.dib.2017.01.011