Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3390–3398
Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Constructing a Bilingual Hadith Corpus Using a Segmentation Tool

Shatha Altammami[1,2], Eric Atwell[2], Ammar Alsalka[2]
King Saud University[1], University of Leeds[2]
Saudi Arabia, UK
Shaltammami@ksu.edu.sa[1]
{Scshal, E.S.Atwell, M.A.Alsalka}@leeds.ac.uk[2]

Abstract

This article describes the process of gathering and constructing a bilingual parallel corpus of Islamic Hadith, which is the set of narratives reporting different aspects of the prophet Muhammad's life. The corpus data is gathered from the six canonical Hadith collections using a custom segmentation tool that automatically segments and annotates the two Hadith components with 92% accuracy. This Hadith segmenter minimises the costs of language resource creation and produces consistent results independently from previous knowledge and experiences that usually influence human annotators. The corpus includes more than 10M tokens and will be freely available via the LREC repository.

Keywords: Hadith, parallel corpus, NLP, language resource, Arabic, English.

1. Introduction

Current advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) led to attempts at computerising a range of tasks that require domain experts. One of the research areas that caught 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 text, 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 Muslims holy book), most laws and legislations are obtained from the Hadith due to its larger scope and incorporated details. Yet, Islamic computational studies has 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.

Research in the area of Hadith computation is still in its infancy (Bounhas, 2019). Yet, there is an annual increase in the number of published papers indicating it is gaining wider attention from multi-disciplinary researchers (Azmi et al., 2019). In such work, researchers gather their own dataset from different sources and sometimes manually process them (Luthfi et al., 2018). This indicates 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 unfeasible to establish benchmarks, compare results or set evaluation measures (Guellil et al., 2019), which makes establishing a Hadith Common Dataset Initiative an imperative.

Through this project, we initiate the Hadith Common Dataset by introducing a well-structured parallel corpus of Hadith in its original classical Arabic text and 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 constitutes only 11% of the available Arabic resources (Guellil et al., 2019).

We name this language resource the *Leeds University and King Saud University (LK) Hadith corpus* to represent the collaboration between the two universities. The corpus will be released via the LREC data repository. Hence, our contribution is twofold:

1. An improvement on a previously created Hadith segmentation tool that automatically identifies and annotates the two components of the Arabic Hadith (Altammami et al., 2019). Segmentation in Hadith has a special meaning different from the standard NLP segmentation where words are segmented to their morphological components. Instead, Hadith segmentation aims to split the Hadith text into its two main components (*Isnad* and *Matn*), where each component consists of several words.
2. A well-structured Arabic-English parallel Hadith corpus that can be used by a broad range of audience. It is particularly useful for those working on Hadith computational studies to test their systems using a common dataset. The corpus is structured to allow research focus on any component of the Hadith. For example, to build ontologies that support Hadith authenticity by focusing on the *Isnad*.

In the next section, we give a brief overview of Hadith and its structure. Then we discuss related work of existing corpora and compare it to our corpus. After that we describe the data source of our corpus and the methodology used to collect the data. Then we explain the construction and evaluation of the Hadith segmenter which was used to annotate the *Isnad* and *Matn* components in the corpus. Finally, we discuss limitations and future directions.

2. What is Hadith

Muslims believe the Quran is God's divine words, which enjoined them to follow the guidance of Prophet Muhammad in their laws, legislations, 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. For example, Muslims prayer, is obtained from the prophet's reported actions; since it is stated in the Quran as an obligation without the exact details of practice.

The 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, it could be a short sentence or long paragraph describing what the prophet said in a specific incident, a dialogue of 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 have deliberately 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.

2.1 Hadith Structure

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 enable researchers access the different component.

Ali bin Mohammed told us, Wakia told him that Younis bin Abi Ishaq heard Mujahid, heard Abu Hurayrah said the Messenger of Allah peace be upon him (PBUH) said: "Jibra'il kept enjoining good treatment of neighbours until I thought he would make neighbours heirs."

حَدَّثَنَا عَلِيُّ بْنُ مُحَمَّد، حَدَّثَنَا وَكِيعٌ، حَدَّثَنَا يُونُسُ بْنُ أَبِي إِسْحَاقَ، عَنْ مُجَاهِد، عَنْ أَبِي هُرَيْرَةَ، قَالَ قَالَ رَسُولُ اللَّهِ صلى الله عليه وسلم " مَا زَالَ جِبْرَائِيلُ يُوصِينِي بِالْجَارِ حَتَّى ظَنَنْتُ أَنَّهُ سَيُوَرِّثُهُ " .

Figure 1: Hadith: Isnad in Bold Followed by Matn

2.2 Existing Hadith Books

In the Islamic literature, there are six canonical Hadith books which are considered authentic. They are a hybrid of two book genres, *Musanaf* and *Musnad* (Brown, 2009). The former includes books that categorizes Hadiths into topics and does not emphasise on the authenticity. On the other hand, Musnad books organizes Hadith based on chain of narrators to place more emphasise on authenticity. This hybrid genre became known as *Sahih* or *Sunan*, where authentic Hadiths are organized under subtitles that indicate the legal implication or ruling the reader should derive from the subsequent Hadiths.

Nowadays, these canonical books are collectively called 'Al-Sihah al-Sittah', which translates as 'The Authentic Six', and they include Sahih Bukhari, Sahih Muslim, Sunan Abu Daud, Sunan Tirmizi, and Sunan Ibn Maja, Sunan Nesa'i [1], and they form the base for Islamic Hadith books. It is worth noting that these books are named after the scholars who compiled them. For example, Sahih Bukhari was compiled by Muhammad al-Bukhari who dedicated years of his life studying Hadiths authenticity before adding them to his book.

Despite their collective name 'The Authentic Six', not all incorporated Hadiths possess the same degree of authenticity. Rather, they were named based on the dominance of authentic Hadiths incorporated (Khan, 1987).

3. Related Work

There are many existing Arabic corpora (Atwell, 2019). 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

[1]We adapted the English spelling of the books names as mentioned in previous survey (Azmi et al., 2019).

incorporate Hadith (Al-Thubaity, 2015). The KSU 50 million words corpus of classical Arabic is designed to help researchers understand the use of words during the period of Quran revelation (Alrabiah et al., 2013). That is to understand its resemblance to the language of Arabs at that time. Another corpus which incorporate Hadith books is the Historical Arabic Corpus or HAC (Hammo et al., 2016), which contains 45 million words from different time periods. Moreover, Tashkeela (Zerrouki and 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 used an open-source OCR called Kraken ibn Ocropus to turn Arabic books into digital form. They aim to incorporate Persian and other languages forming a very large Islamic corpus (Belinkov et al., 2018).

A smaller Hadith corpus is presented in Alosaimy's PhD work as an annotated linguistic resource which comprises 144,000 words extracted from the Riyadu Assalihin Hadith book. The process of developing this corpus went through phases, as the text was diacritized by borrowing diacritics of the same text cited in different resources by relying on word n-grams concordance. Moreover, this corpus provides tagging at the morphological level and several Hadith translations aligned at the narrative level (Alosaimy and Atwell, 2017).

In another study, a survey was conducted to enumerate the freely available Arabic corpora and stated the existence of one Hadith corpus. However, it was not accessible, 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 obsolete after time. Therefore, we attempt to mitigate that by sharing our corpus on LREC repository.

Recently, a Hadith corpus was created by scrapping different Hadith websites that cover several languages including Arabic, English and Urdo (Mahmood et al., 2018). We aim to investigate merging it with our corpus by applying AI methods to align the different translations.

Our corpus is different from the existing ones since it provides Arabic Hadith with its English Translation aligned at the narrative level. This cannot be accomplished by extracting the Hadith from the existing corpora since they include Arabic texts only. Hence, we found *sunnah.com* website where tremendous human efforts were devoted to structure and align the Arabic Hadiths with its English translations. Hence, we extracted the data, then applied our Hadith segmentation tool to label the *Isnad* and *Matn* components of the Hadith. Table 1 highlights the difference between our corpus and existing ones.

4. Data collection

The six canonical Hadith books are well structured, they follow the Arabic naming conventions used in the 7th century where a book is divided into books and each consists of chapters. However, to design our corpus, we converted that to the modern naming conventions where a book is divided into chapters which include sections that incorporate Hadiths.

The Hadiths are organized into a topology of topics where each chapter is dedicated to one theme. Within the chapter, there are several sections that the scholar 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 incorporate a comment by the scholar, usually regarding the authenticity of the Hadith.

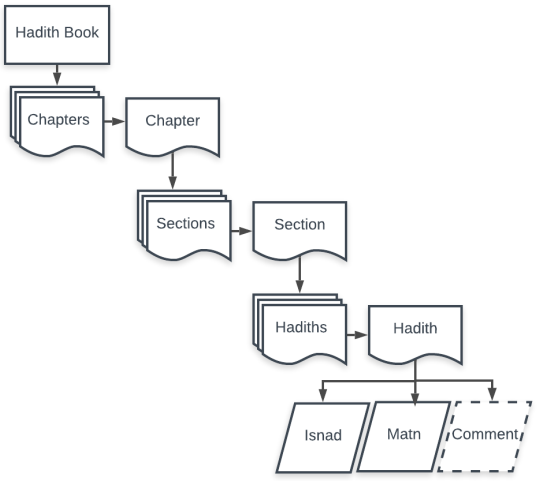

Figure 2: Canonical Books Structure

We intend to maintain this structure in our corpus. Hence, we sought electronic sources of Hadith that followed this structure. Several websites hosts Hadith books; however, they did not meet our requirements. For example, *ahadith.co.uk* contains the English translation 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 satisfy building a parallel corpus of English-Arabic aligned Hadiths since the English version is provided as a downloadable PDF file for few books.

Only *sunnah.com* met our requirements. It maintained the structure of the books and the English translation is aligned in parallel with the Arabic Hadith at the narrative level. Moreover, every component including

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

Table 1: Corpora Comparison

chapter, section, *Isnad* and *Matn* are allocated a unique HTML tag.

5. Corpus Creation

We developed a software to scrape *sunnah.com* pages and extracted the information from every Hadith. However, the *Isnad* was not annotated consistently. For example, the Arabic *Isnad* is not separated from the *Matn* in most Hadiths, despite the existence of an HTML tag dedicated to *Isnad*. In other cases, only the prophetic words are considered *Matn*, while the narration of the incident is incorporated within the Isnad. This could be due to the website being built by a group of web developers. To overcome the inconsistent annotation, a Hadith segmentation tool was developed to automatically segment *Isnad* from *Matn*.

5.1 Hadith Segmentation Tool

Building a Hadith segmentation tool is a non-trivial task that possesses key challenges associated with Hadith structure that requires novel methods to overcome them. In fact, recognizing sentence boundaries in a running text is a difficult task in languages such as Arabic, especially in the absence of strict punctuation rules and the lack of capitalization. Moreover, segmenting Hadith components is a domain-specific task that can be even tricky for the non-specialist. Therefore, automating it ensures consistency in segmentation.

In a previous work (Altammami et al., 2019), we created the first version of the Hadith segmenter which uses look-up lists and applies a back-off algorithm to segment *Isnad* from *Matn*. Although it produced acceptable results, we improve its performance by incorporating a machine learning (ML) model into the pipeline, and modified the algorithm to deal with irregular Hadith structures. Moreover, we doubled the evaluation data to 500 Hadiths extracted from the six canonical books where Hadiths with irregular structures were manually chosen. This is to ensure the evaluation data is representative of the whole corpus. Using this approach, 464 Hadiths were correctly segmented producing 92% accuracy.

The Hadith segmenter pipeline is shown in Figure 3, where it applies the following steps:

1. First, it takes the Hadith input and pre-processes it to remove diacritics, punctuations and extra white spaces. Diacritics were removed to overcome data sparseness and enhance term weighting.
2. Then it tokenizes the pre-processed Hadith into bigrams of words. Bigrams were chosen based on its best performance compared to the other n-gram features as explained in section 5.1.4.
3. After that it labels every token as '*Isnad*' or '*Matn*' by using a Naive Bayes Classifier which we describe in section 5.1.1.
4. Once every token is labelled, a rule-based approach is applied to find the exact segmentation point as detailed in section 5.1.2.
5. Finally, the segmentation point is applied on the original Hadith to produce Isnad and Matn segments with diacritics and punctuations.

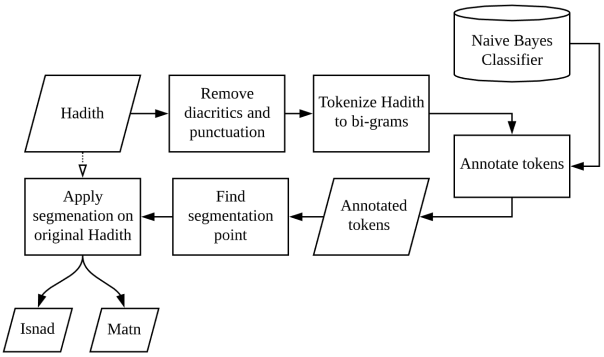

Figure 3: Pipeline of Hadith Segmenter

5.1.1 Building a Classifier Model

We used ML to ensure the Hadith segmenter scales well with new data extracted form various Hadith

books. Hence, we built a Naïve Bayes Classifier that takes in Hadith bigrams and classifies each as '*Isnad*' or '*Matn*'. It is trained on 4,686 segmented Hadiths extracted from Sahih Bukhari book. This training data includes 314,340 bigrams instances divided into *Matn* and *Isnad* as shown in Figure 4.

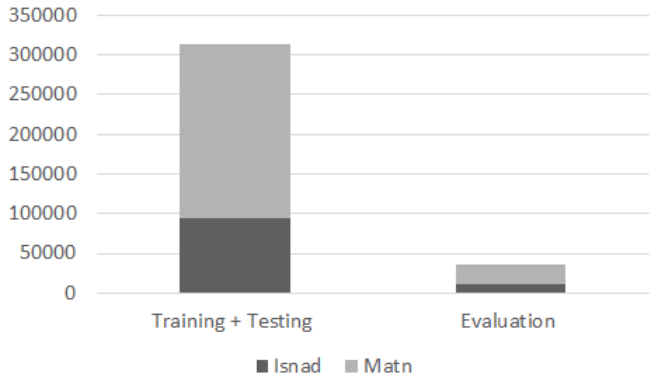

Figure 4: Data Distribution: Bigrams Tokens

In the training phase, the data is divided into 70% training and 30% testing which produced 94.9% accuracy. Since we were satisfied with its performance on testing data, we used it to classify our evaluation data which includes 36,218 bigrams and got 88.6% accuracy. Figure 5 shows that although *Isnad* training data was relatively small, it has the best performance. This could be due to the fact that *Isnad* usually consists of words that are either proper names or transmission methods e.g. 'said', 'heard'.

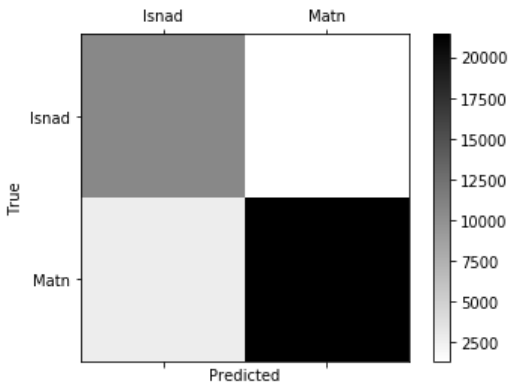

Figure 5: Confusion matrix of classifier performance to annotate bigrams tokens.

It is worth noting that the typical out of vocabulary (OOV) problem which is usually associated with proper nouns does not apply to Hadith. This is because the names of scholars in the Hadith literature are a closed set.

5.1.2 Segmentation Algorithm

Once the bigrams are annotated, the segmenter finds the exact segmentation point. Hadith segmentation is a domain specific task, and as shown in a previous survey, rule-based approaches produced the highest accuracies (Altammami et al., 2019). Our rule-based Hadith segmentation approach is simplified in Algorithm 1. This algorithm was able to identify *Isnad* with irregular patterns, and Hadiths that contain parallel *Isnad* which was one of the limitations in version one.

Algorithm 1 : Find Segmentation Point

```
for every token do
    if token is Matn then
        if next three tokens are Matn then
            Segmentation point A found
        end if
    end if
end for
if Segmentation point A is found then
    Check if another Isnad exists
    for every token after segmentation point A do
        if five tokens labelled Isnad follows then
            Find next set of tokens labelled Matn
            Segmentation Point B found
        end if
    end for
end if
if no segmentation point found then
    Hadith does not contain Matn
end if
```

5.1.3 Segmentation Result

The Hadith below is an example with parallel *Isnad* where the first chain of narrators is followed by the prophet's name, which is followed by another chain of narrators that ends with the prophet's name as well. These two chain of narrators are followed by the *Matn*, where the segmentation point should be detected.

To segment this Hadith, the tool uses the classifier to label the first 13 tokens as *Isnad*, followed by 5 tokens as *Matn*, then another set of 6 tokens as *Isnad*, and finally 7 tokens as *Matn*. Then it finds the segmentation point by detecting that the first set of *Matn* tokens is followed by another set of *Isnad* tokens. Therefore, it segmented the Hadith after the second set of *Isnad* tokens as indicated by '||||'in the text.

Example 1:

> 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".
> 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.

Another limitation in the first version was dealing with *Isnad* that contains irregular patterns which is addressed in this work as illustrated in Example 2. However, Example 3 demonstrates that there is room for improvement especially to deal with Hadiths that contain vague segmentation points. Note that for space issues Hadiths in the examples were truncated as indicated by (...).

Example 2:

Nasser bin Ali Juhadhmi and Abu Ammar told us and the meaning is the same but the words are of Ammar they said, Sufian bin Aayneh from Alzahri from Hamid bin Abdul Rahman on the authority of Abu Hurayrah said |||| a man came and said, "O Allah's Apostle! I have been ruined." ...
حدثنا نصر بن علي الجهضمي وأبو عمار والمعنى واحد واللفظ لفظ أبي عمار قالا أخبرنا سفيان بن عيينة عن الزهري عن حميد بن عبد الرحمن عن أبي هريرة قال أتاه |||| رجل فقال يا رسول الله هلكت...

Example 3:

Muhammad bin Mansour told us, that Sufian said Yahya bin Said told us about Muslim bin Abi Maryam |||| a Sheikh from Madinah then I met the Sheikh and he said he heard Ali bin Abdul Rahman say I prayed beside Ibn Omar, while I turned the gravel he said Do not fluctuate the gravel, turning the gravel is from the devil and do as I saw the Messenger of Allah peace be upon him do...
أخبرنا محمد بن منصور قال حدثنا سفيان قال حدثنا يحيى بن سعيد عن مسلم بن أبي مريم |||| شيخ من أهل المدينة ثم لقيت الشيخ فقال سمعت علي بن عبد الرحمن يقول صليت إلى جنب ابن عمر فقلبت الحصى فقال لي ابن عمر لا تقلب الحصى فإن تقليب الحصى من الشيطان وافعل كما رأيت رسول الله صلى الله عليه وسلم يفعل قلت وكيف رأيت رسول الله صلى الله عليه وسلم يفعل قال هكذا ...

5.1.4 Segmenter Analysis

In this section we compare the performance of the segmenter versions. Version 1 represent the previous work (Altammami et al., 2019), and version 2 as discussed in this paper. Figure 6 shows their performance on the evaluation data of 500 Hadiths. It is clear that among the different n-gram models, bigrams scores the highest accuracy. Although the overall performance is not significantly different, version 2 accuracy is slightly higher specifically in segmenting Hadiths with irregular patterns.

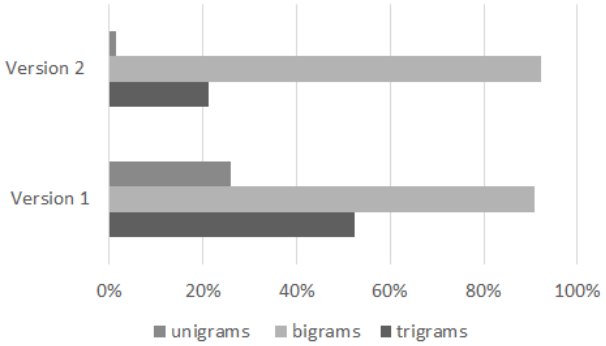

Figure 6: Segmenter performance with unigram, bigram, and trigram models.

5.2 Hadith Annotation

Once we were satisfied with the segmentation tool, we started the process of constructing and annotating the corpus. We applied the Hadith segmenter to extract *Isnad* and *Matn* of every Arabic Hadith. Then we captured Hadith meta data including, chapter, section, Hadith number, and saved it in a record where they are separated by commas. Hence, the CSV (comma separated values) files are 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 listed in Table 2 and an example of how one Hadith record is represented in a CSV file is broken down for readability in Figures 7, 8 and 9.

An illustration of the LK Corpus structure is shown in Figure 10. It is a simple structure that corresponds to the original structure of the books. 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 Sahih Bukhari folder which represent the number of chapters in Sahih Bukhari book. The first CSV file is named 'Chapter1.csv' and it contains seven Hadith records. Table 3 shows the number of Hadiths in each book.

5.3 Corpus Evaluation

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 Hadiths is larger than the Arabic version. However, the Arabic Hadiths are richer in vocabulary as it contains more unique words than the English version as shown in Table 4.

It is worth noting that actual number of Hadith teachings is few thousands, but it exploded to a very large number since the Hadith scholars count

Annotation	Description
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.

Table 2: Corpus Annotation

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 7: Example of Hadith Record Extracted from Sunan Tarmizi – 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 8: 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 خَطِيئَةٌ . 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 9: Continued Example of Hadith Record – Part 3

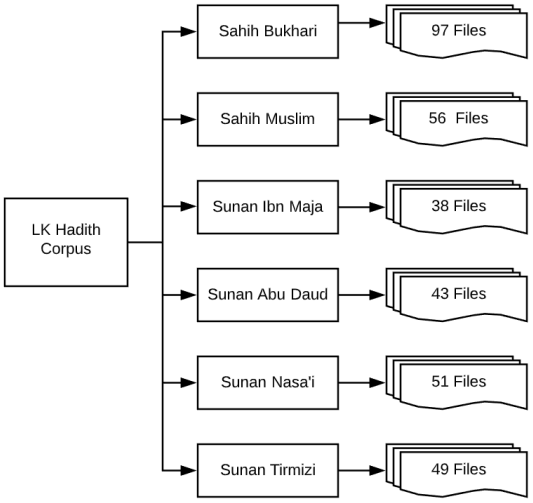

Figure 10: LK Hadith Corpus Structure

Book	Number of Hadiths
Bukhari	6,633
Muslim	7,293
Nesa'i	5,680
Ibn Maja	10,082
Tirmizi	4,209
Abu Daud	5,141
Total	39,038

Table 3: Number of Hadiths From Each Book

each transmission channel as a unique Hadith when they compiled them into their collections. In other words, some *Matn* are narrated by different chain of narrators which makes them different Hadiths. Example 4 illustrates this kind of Hadith that consists of *Isnad* only.

Example 4:

> I was told the same by Mohammed bin Muthanna who told us that Abdul Samad told him that Muthanna said this .
> وحدثنيه محمد بن المثنى حدثنا عبد الصمد حدثنا المثنى بهذا الإسناد.

Moreover, the above Hadith is translated by the first author since an English translation is not provided. This is because such Hadiths do not contain any teaching (*Matn*). Hence, the corpus contain a number of similar Hadiths which do not have a parallel English value.

	English	Arabic
Word Tokens	5,498,262	4,768,042
Word Types	70,714	147,547

Table 4: Word Frequencies

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 have found minor 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 *Sunnah.com*. So far, we have checked Sahih Bukhari against the PDF version of the book, and we are confident that it is the gold standard of our corpus even though the remainder of the corpus was not manually checked. Furthermore, since the segmenter produced an accuracy of 92% on evaluation data. the annotation of Isnad and Matn segments in the corpus has an error rate of 8%.

6. Conclusion and Future Work

We have presented the creation of LK Hadith parallel corpus using a domain-specific tool to segment and annotate Hadith components. This corpus is particularly useful for researchers in Hadith computational studies which is currently in its infancy. Moreover, the Arabic–English Hadith pair opens new avenues to other areas of research including machine translations of classical Arabic.

Currently this corpus is being exploited for experiments in unsupervised relation discovery between the Hadith and the Quran. Hence, the well-structured corpus facilitated focusing on the *Matn* component to study the actual Hadith teaching without the *Isnad* (chain of narrators) affecting the results.

In the future, we plan to extend this corpus to include Hadith commentaries aligned with Hadith at the narrative level and possibly include the translations of Hadiths in other languages. Additionally, we aim to develop the segmentation tool to detect Hadiths in a running text. Researchers in Islamic Digital Humanities are keen to have such tool that will enable the automatic detection and extraction of Hadith from electronic books. One of their current projects is studying forged Hadiths attributed to the prophet to understand the political views at a specific time in history. Therefore, forged Hadiths are being discovered and might keep emerging, which indicates a Hadith segmentation tool is not dealing with a closed set of data.

7. 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.

References

Al-Thubaity, A. O. (2015). A 700m+ arabic corpus: Kacst arabic corpus design and construction. *Language Resources and Evaluation*, 49(3):721–751.

Alosaimy, A. and 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.

Alrabiah, M., Al-Salman, A., and Atwell, E. (2013). The design and construction of the 50 million words ksucca. In *Proceedings of WACL'2 Second Workshop on Arabic Corpus Linguistics*, pages 5–8. The University of Leeds.

Altammami, S., Atwell, E., and Alsalka, A. (2019). Text segmentation using n-grams to annotate hadith corpus. In *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, pages 31–39.

Atwell, E. (2019). Arabic corpus linguistics. *Using the Web to Model Modern and Qurʾanic Arabic*, pages 100–119.

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*, pages 1–46.

Belinkov, Y., Magidow, A., Barrón-Cedeño, A., Shmidman, A., and Romanov, M. (2018). Studying the history of the arabic language: Language technology and a large-scale historical corpus. *arXiv preprint arXiv: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):23.

Brown, J. A. (2009). *Hadith: Muhammad's legacy in the medieval and modern world*. One world Publications.

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., and Nouvel, D. (2019). Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*.

Hammo, B., Yagi, S., Ismail, O., and AbuShariah, M. (2016). Exploring and exploiting a historical corpus for arabic. *Language Resources and Evaluation*, 50(4):839–861.

Khan, S. H. (1987). *Al-Hitta Fi Dhikr Al-sihah Alsitta*. Beirut.

Luthfi, E. T., Suryana, N., and Basari, A. H. (2018). Digital hadith authentication: A literature review and analysis. *Journal of Theoretical and Applied Information Technology*, 96(15):5054–5068.

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

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*, pages 1–8.

Zerrouki, T. and Balla, A. (2017). Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems. *Data in brief*, 11:147.