Encoding prototype of Al-Hadith Al-Shareef in TEI

Hajer Maraoui, Kais Haddar, Laurent Romary

Encoding prototype of Al-Hadith Al-Shareef in TEI

Hajer Maraoui[1], Kais Haddar[2], Laurent Romary[3]

[1] Faculty of Mathematical, Physical and Natural Sciences of Tunis
MIRACL Laboratory, University of Sfax
hajer.maraoui@gmail.com
[2] Faculty of Science of Sfax, MIRACL Laboratory
kais.haddar@yahoo.fr
[3] Inria, Team ALMAnaCH
laurent.romary@inria.fr

Abstract. The standardization of Al-Hadith Al-Shareef can guarantee the interoperability and interchangeability with other textual sources and takes the processing of Al-Hadith corpus to a higher level. Still, research works on Hadith corpora had not previously considered the standardization as real objective, especially for some standards such as TEI (Text Encoding Initiative). In this context, we aim at the standardization of Al-Hadith Al-Shareef on the basis of the TEI guidelines. To achieve this objective, we elaborated a TEI model that we customized for Hadith structure. Then we developed a prototype allowing the encoding of Hadith text. This prototype analyses Hadith texts and automatically generates a standardized version of the Hadith in TEI format. The evaluation of the TEI model and the prototype is based on Hadith corpus collected from Sahih Bukhari. The obtained results were encouraging despite some flaws related to exceptional cases of Hadith structure.

Keywords: Hadith text, TEI model, standardization, prototype.

1 Introduction

The processing of Al-Hadith Al-Shareef has always been a center of academic interest. On the one hand, this interest is due to the importance of Al-Hadith Al-Shareef in Islamic law. It is the second fundamental source, after the Quran, of Islamic legislation. On the other hand, linguistic researches on Arabic language define Al-Hadith corpus as one of references for classical Arabic from the pre-Islamic era. Many studies carried out on the Al-Hadith corpus especially in linguistic analyses and information retrieval. However, the representation of such corpora poses serious structuring and text unification problems, which require their standardization (or normalization). This could lead to a new presentation of the Text based upon a descriptive and detailed annotation for each of its parts. A normalized corpus allows also the compatibility and the interchangeability between NLP applications. Indeed,

the normalization of Al-Hadith Al-Shareef can bring the automatic processing of such corpus to another level.

However, normalizing Al-Hadith corpus is a complex task. In fact, it requires a specific model customized for standardizing the Hadith text. This task requires the selection of the standardization model that harmonizes with Arabic language in general and Al-Hadith structure specifically. Moreover, to realize a normalized Hadith corpus, an automatic encoding process can facilitate this task.

Our main objective is the normalization of Al-Hadith Al-Shareef. To realize this, we start with a deep study on Hadith text structure. Furthermore, to reach the normalization of Al-Hadith Al-Shareef, we apply the Text Encoding Initiative (TEI) (http://www.tei-c.org/) guidelines. We elaborate a TEI customized model for encoding Hadith text. Also, to create a normalized Hadith corpus, we make a prototype to create automatically an encoded version of Al-Hadith text with TEI structure.

In the present paper, we begin with a state of the art on Al-Hadith Al-Shareef. Second, we continue with an overview on TEI guidelines. Third, we present our model for encoding Hadith text. Then, we present our prototype for constructing a normalization of Hadith texts with TEI structure. This section is followed by an evaluation step. We cloture our paper with a conclusion and some perspectives.

2 State of the Art on Al-Hadith Al-Shareef

Hadiths (or prophetic traditions) are narrations on the life and deeds of Prophet Muhammad (peace and blessing upon him), which report what he said or did, or of his implicit approval of something said or done and by itself define what is considered good, by providing details to regulate all aspects of life in this world and to prepare people for the beyond, clarifying the Qur'anic shades. The traditional Muslim schools of jurisprudence regarding Hadith constitute an important tool for understanding Qur'an and in all matters related to jurisprudence [1][2].

The Hadith consists of two parts: the actual narration, which called *Matn* (المتن); and the chain of narrators who has transmitted the narration, known as *Isnad* (إسناد). The *Isnad* consists of a short or long chronological list of the narrators, each mentioning the one from whom he heard the Hadith all the way to the prime narrator of the *Matn* followed by the *Matn* itself [1].

Research in the *Isnad* is very important in the science of Hadith. Islamic scholars have agreed that *Isnad* is required to prove the accuracy and the soundness of the Hadith which means that any flaw in the chain of transmitters lead to the negation of the Hadith. In order to know whether the Hadith is authentic or not, the Hadith scholars follow clear steps in the judgment on the Hadith *Isnad* that considered as traditional methods. Sahih Bukhary and Sahih Muslim are the recognized collection of authentic assortment of the *Sunna* [1] [3-5].

Nowadays, software tools help judge the Hadith *Isnad* like electronic Hadith encyclopedias and some websites. Additionally, information retrieval and search engines that related to semantic web can used to serve in deciding the degree of the

Hadith *Isnad*. Scholars such as Al-Albani have agreed and encouraged using computers and programs in serving religion and Hadith [1].

There are many projects handled with Hadith corpus. Indeed, these projects focused on several branches of researches such as Hadith ontology, linguistic analyzing, Hadith segmentation, authorship attribution, classification and the mining of information.

In [5], the researchers proposed a model for the unsupervised segmentation and the linguistic analysis of the Arabic texts of Hadith. This model is named SALAH. The model automatically segments each text unit in a transmitter chain *Isnad* and text content *Matn*. A tailored, augmented version of the AraMorph morphological analyzer (RAM) analyzes and annotates lexically and morphologically the text content. A graph with relations among transmitters and a lemmatized text corpus, both in XML format, are the final output of the system.

In [1], the autor constructed an ontology-based *Isnad* Judgment System (IJS) that automatically generates a suggested judgment of Hadith *Isnad*. It based on the rules that Hadith scholars follow to produce a suggested judgment. A prototype of the approach implemented to provide a proof of concept for the requirements and to verify its accuracy.

Authors of paper [6] built a domain specific ontology (Hadith Isnad Ontology) to support the process of authenticating *Isnad*. They evaluate the ontology through Hadith example and DL-Queries.

Author of paper [7] compared the effectiveness of four different automatic learning algorithms for classifying Hadith corpus into 8 selective books depending on Sahih Bukhary. The automatic learning algorithms are Rocchio algorithm, K-NN algorithm (K- Nearest Neighbor), Naïve Bayes algorithm and SVM algorithm (Support Vector Machines).

In [2], the authors reported on a system that automatically generates the transmission chains of a Hadith and graphically display it. They involve parsing and annotating the Hadith text and identifying the narrators' names. They use shallow parsing along with a domain specific grammar to parse the Hadith content.

In [8], the author experimented author discrimination techniques between the Qur'an and the Hadith. The Qur'an is taken in its entirety, whereas for the Prophet's statements, the researcher chose only the certified texts of Sahih Bukhari. Three series of experiments are done and commented on. The author's investigation sheds light on an old enigma, which has not been solved for 14 centuries: in fact, all the results of this investigation have shown that the two books should have two different authors.

In [9], the researchers reimplemented and evaluated the methods of artificial intelligence using a single dataset. The result of the evaluation on the classification method reveals that neural networks classify the Hadith with 94% accuracy. The Hadith mining method that combines vector space model, Cosine similarity, and enriched queries obtains the best accuracy result.

3 The Text Encoding Initiative

The Text Encoding Initiative (TEI) is an international project aiming at the development of a set of standards for the preparation and the exchange of electronic texts. The TEI was founded in November 1987 by a group of international text database leaders. TEI was created officially in 1988 under the aegis of the ACH[1], the ACL[2] and the ALLC[3] [1].

The publication of the works of the various committees' results was in the form of "Guidelines" which have been developed and are maintained by the Text Encoding Initiative Consortium [10]. The TEI guidelines recommend suitable ways to represent the features of textual resources using a set of XML elements in order to elicit the text structure and simplify its digital processing. Indeed, these guidelines present conventions of usable coding in several domains and can be applied to texts in any natural language, of any date, in any literary genre or text type [10]. Moreover, they can be applied as well to create new information that to exchange existing information.

TEI annotation is based on a "patrimonial" transcription of the text, interested in giving as much information as possible while allowing automatic processing for language searches, historical data, different versions of the document and variations level of accuracy that can be adapted to the desired searches [11]. TEI offers the possibility of multilingual uses of the structural description. The universality of the elements allows the compatibility of the analyzes, and the digitized sources within such frameworks as the BVH[4] and belong to all encyclopedic domains, in several ancient and modern languages.

The rules and recommendations made in these guidelines are expressed in terms of the extensible Markup Language (XML) so a TEI document has to comply with XML coding rules. One of the main advantages of the XML language is that it is possible to solve by encoding the very large typographic and textual variation of the documents: this flexibility should not, however, be a hindrance to the acquisition and encoding of textual corpora.

The fundamental structure of a TEI document describes the textual part of the text. In what follows, we give overviews onto the structure and the TEI representation of documents. The basic structure of a document encoded with TEI is represented as follow:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <!-- Header properties; meta-data -->
  </teiHeader>
  <text>
```

1 Association for Computers and the Humanities
2 Association for Computational Linguistics
3 Association for Literary and Linguistic Computing
4 The laboratory LI-RFAI (director Jean-Yves Ramel) and the consortium Navidomass (ANR project 2007-2009).

```
    <front>
      <!-- front information -->
    </front>
    <body>
      <!-- main body -->
    </body>
    <back>
      <!-- back information -->
    </back>
  </text>
</TEI>
```

Any textual document encoded with TEI includes a document header, with the <teiHeader> element, and a text part within the <text> element. TEI header contains all the information analogous to that provided by the title page of a printed document. It has four parts: a bibliographic description of the machine-readable text, a description of the way it has been encoded, a text profile, and a revision history [11-13].

A TEI document can contain five textual elements: <text>, <front>, <group>, <body> and <back>. Only <text> and <body> are the obligatory elements. The use of the <front>, <group> and <back> is optional. The <front> element is defined to cluster together the pieces located before the beginning of the text itself. The <back> element is used in case the document contains an annex in the back of the text. The <group> element is specified to include several texts in collections. The <body> is included in the text and it contains the core text of the document [11].

Personalization is a central aspect of TEI using. There are three methods of customization in TEI: TEI Lite, web application Roma and TEI ODD. TEI Lite was originally designed as a demonstration of the customization mechanism. The Roma web application was introduced to select TEI modules that manipulate the elements. As for TEI ODD language, it essentially allows the manual specification of the TEI models, allowing the modification or addition of new elements [10].

Encoding quotation with TEI

Quotation marks are conventionally used to indicate certain elements appearing in a text, the most frequent case is for the quotation. However, the marking of the underlying logical element (for example, a quotation or a piece of direct speech) in the text is recommended, rather than just recording quotation marks in the text [10-12]. The following TEI elements are specified for the encoding of quotation and narration which can be adaptable with the structure of Al-hadith *Matn*:

- **<q>** (or quoted) contains material which is distinguished from the surrounding text using quotation marks or a similar method. It contains a citation or an apparent citation - the representation of a speech or thought, marked out to indicate that it is a quotation. Among the possible attributes there are:

  - The @type attribute: it can be used to indicate whether the quoted passage is pronounced or simply thought, or to characterize it more finely: possible values are: *spoken* (for the representation of direct speech, usually marked by quotation marks) and *thought* (for the representation of thoughts, for example, internal monologue).
  - The @who attribute: it identifies the speaker in the case of a direct speech passage.
- **<said>** (speech or thought) indicates passages thought or spoken aloud, whether explicitly indicated in the source or not, whether directly or indirectly reported, whether by real people or fictional characters.
- **<quote>** (quotation) contains a phrase or passage attributed by the narrator or author to some agency external to the text.

The <q> element may be used if no further distinction beyond this is judged necessary. If it is felt necessary to distinguish such passages further, for example to indicate whether they are regarded as speech, writing, or thought, either the type attribute or one of the more specialized elements discussed in this section may be used. For example, the element <quote> may be used for written passages cited from other works, or the element said for words or phrases represented as being spoken or thought by people or characters within the current work. If the distinction among these various reasons why a passage is offset from surrounding text cannot be made reliably, or is not of interest, then any representation of speech, thought, or writing may simply be marked using the q element. Quotation may be indicated in a printed source by changes in type face, by special punctuation marks (single or double or angled quotes, dashes, etc.) and by layout (indented paragraphs, etc.), or it may not be explicitly represented at all.

Encoding person name with TEI

The TEI guidelines present several models for encoding a set of types of named entities. One of these models aims to annotate person names. This can cover the annotation of all the information related with the person name (such as first name, family name, additional name, etc.). To conduct nominal record linkage or even to create an alphabetically sorted list of personal names, it is important to distinguish between a family name, a forename and an honorary title. Similarly, when confronted with a string such as 'أمير المؤمنين أبي حفص عمر ابن الخطّاب' ('*Prince of the Believers Abu Hafs Umar Ibn Al-Khattab*'), the analyst will often wish to distinguish amongst the various constituent elements present, since they provide additional information about the status, the occupation, or the residence of the person to whom the name belongs. The following TEI elements are provided for encoding person name and related purposes.

- **<persName>** (personal name) contains a proper noun or proper-noun phrase referring to a person, possibly including one or more of the person's forenames, surnames, honorifics, added names, etc.
- **<surname>** contains a family name, as opposed to a given, baptismal, or nick name.
- **<forename>** contains a forename, given or baptismal name.
- **<roleName>** contains a name component which indicates that the referent has a particular role or position in society, such as an official title or rank.
- **<addName>** (additional name) contains an additional name component, such as a nickname, epithet, or alias, or any other descriptive phrase used within a personal name.
- **<nameLink>** (name link) contains a connecting phrase or link used within a name but not regarded as part of it, such as "*de"* or *"بن".*
- **<genName>** (generational name component) contains a name component used to distinguish otherwise similar names on the basis of the relative ages or generations of the persons named.

We were inspired by TEI models for encoding quotation and person information to build a model for Hadith text encoding. The following section illustrates this model who was improved to encode *Matn* and *Isnad* of Hadith text.

4 Proposed model for Hadith encoding with TEI

In this section, we present our proposed encoding model shaped for Hadith text and inspired from the TEI guidelines. To realize the standardization of Hadith corpus with TEI, we start by the selection of the necessary data categories that harmonized with Hadith text specification. The structure of the Hadith text characterized with the imbrication of the different part of the text: Hadith include *Isnad* and *Matn*, and each one of them include other parts. Also, TEI allows the imbrication and the restructuring of several models and elements.

We start with the concept of the similarity between the structure of Hadith text is similar to the basic structure of quotations and named entities: considering the *Matn* text of the Hadith as a quotation and the *Isnad* as an enchained list of person names. To achieve our initial TEI model for Hadith encoding, we select the adaptable TEI elements for *Isnad* and *Matn* from TEI model for encoding quotations and named entities [14].

Starting with the *Isnad* encoding, we consider that the structure of TEI model for encoding the person name can be developed to create a TEI model conforming with the structure of an Arabic person name. Moreover, we based on the integration and imbrication of TEI elements to represent the chain of Al-Hadith transmitters. Returning to the TEI coding of complex structures in general, person names are presented at least by a single <persName> element. The <persName> element is the one that includes all the person name information. Thus, we use this element for transmitter name annotation.

The structure of the *Matn* is quite similar to the structure of quotations. In fact, many suggestions were proposed by TEI to encode quotations or narration of the author or transmitted quote mentioned inside the document. We used a TEI model to encode *Matn* text. Fig. 1 illustrates the basic elements model that we customized to adapt Al-Hadith text structure.

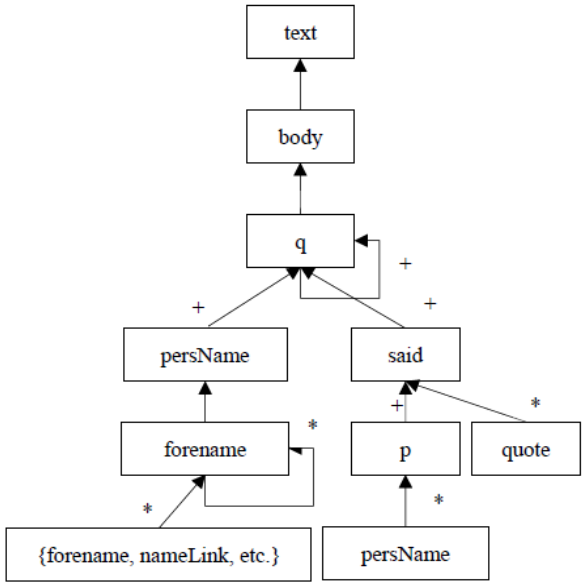

Fig. 1. Extraction from the basic TEI encoding model for Hadith text

Fig. 1 presents our adapted model respecting the main structure of an ideal Hadith text. The representation starts with the encoding of the header of the Hadith proprieties in <teiHeader> element. Then the main corpus in <body> element includes in <text> element. The body contains a <q> element to encode all the Hadith. This quotation element comprises the *Isnad* encoded in a succession of <persName> elements and the *Matn* in a <said> element. <persName> covers all the transmitter name information. The *Matn* also can include person names which will be encoded in <persName>. The *Matn* may contain some direct quotations encoded in <quote> elements.

The development of a specific TEI model for Al-Hadith texts aims to formulate a TEI structure for the encoding of Al Hadith texts. To this first objective, we select the necessary data categories and the adaptable TEI elements from TEI model for encoding quotations and named entities. Our second objective was the creation of a prototype to generate a normalized Hadith text encoded with TEI model. We present this prototype in the next section.

5 Elaborated prototype for the automatization of Hadith encoding in TEI

To implement the creation of the encoded Hadith corpus with our TEI model, we proposed a prototype for encoding Hadith text. This prototype is developed to generate automatically the encoded Hadith texts with TEI format. For the implementation, we used some tools and programs. First, we designed our system with UML. Then, we used Oxygen XML Editor to adapt a TEI structure for the encoded of Hadith text. After that, we developed the prototype using JAVA language and the API JDOM Library.

The creation of a normalized Hadith with our prototype can be divided into two steps. First, as an input file, the system requests the user to choose an external Hadith file path with .txt extension which contains a Hadith text. Then, the prototype reads the Hadith text and separate the *Isnad* from the *Matn*.

For the *Isnad*, the system identifies the narrators from the chains of transmitters in the *Isnad* and encodes each narrator name in a <persName> element. To keep the sequence of the transmitters in order, the system assigns for each <persName> element an "xml:id" attribute which contains as value the order of the narrator in the chains of transmitters. Furthermore, according to Hadith text, the narrator name can contain more than one forename. To organize them in order, the system attribute for each forename a "sort" attribute to sort them by number. This step allows the generation of the encoded chains of transmitters.

For the *Matn* of the chosen Hadith, the prototype identifies his structure and encloses it in a <said> element. To identify the prime narrator who is the first transmitter of the Hadith, the <said> element get a "who" attribute which contains the reference of the prime narrator as value. In Hadith text, the *Matn* is characterized by different structures, it could be narrative text with no quotations, or sort of conversations which can contain direct discourses or quotations. It can also contain other person names that need to be identified in the encoding phase. The prototype allows the identification of all these data information. The encoding of each part of the *Matn* goes line by line: the system identifies the structure of each part and encodes it in a corresponding element, for example, a narrative text identified as a narration an encoded in <p> (paragraph) element or a quotation encoded in <quote> element.

After that, the prototype generates the complete TEI encoded file and save it as an output of the system. The following XML code present an output file from our prototype covering the TEI encoding of the Hadith number 46 from the chapter of Belief from Sahih Al-Bukhari book.

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>…</teiHeader>
  <text>
    <body>
      <div xml:lang="ar">
        <q>
```

```
  <!--encoding of Isnad-->
  <p>حَدَّثَنَا</p>
  <persName xml:id="p5">إِسْمَاعِيلُ</persName>
  <p>قَالَ</p>
  <p>حدثني</p>
  <persName xml:id="p4">
  <forename sort="1">مَالِكُ</forename>
  <forename sort="2" type="nasab">
    <nameLink>بْنُ</nameLink>
    <forename>أَنَسٍ</forename>
  </forename>
  </persName>
  <p>عَنْ عَمِّهِ</p>
  <persName xml:id="p3">
    <forename sort="1" type="kunya">
      <nameLink>أَبِي</nameLink>
      <forename>سُهَيْلٍ</forename>
    </forename>
    <forename sort="2" type="nasab">
      <nameLink>بْنِ</nameLink>
      <forename>مَالِكٍ</forename>
    </forename>
  </persName>
  <p>عَنْ</p>
  <persName xml:id="p2">أَبِيهِ</persName>
  <p>أَنَّهُ</p>
  <p>سَمِعَ</p>
  <persName xml:id="p1">
    <forename sort="1">طَلْحَةَ</forename>
    <forename sort="2" type="nasab">
      <nameLink>بْنَ</nameLink>
      <forename>عُبَيْدِ الله</forename>
    </forename>
   </persName>
  <p>يَقُولُ</p>
<!--encoding of Matn-->
<said who="p1">
  <p>
  جَاءَ رَجُلٌ إِلَى رَسُولِ اللَّهِ صلّى الله عليه وسلّم مِنْ أَهْلِ نَجْدٍ، ثَائِرُ الرَّأْسِ، يُسْمَعُ دَوِيُّ صَوْتِهِ،
  وَلاَ يُفْقَهُ مَا يَقُولُ حَتَّى دَنَا، فَإِذَا هُوَ يَسْأَلُ عَنِ الإِسْلاَمِ فَقَالَ رَسُولُ اللَّهِ صلّى الله عليه وسلّم:
  </p>
  <quote>.«خَمْسُ صَلَوَاتٍ فِي الْيَوْمِ وَاللَّيْلَةِ»</quote>
  <p>:فَقَالَ هَلْ عَلَيَّ غَيْرُهَا؟ قَالَ</p>
  <quote>.«لاَ، إِلاَّ أَنْ تَطَوَّعَ»</quote>
```

```
            <p>قَالَ رَسُولُ اللَّهِ صلّى الله عليه وسلّم:</p>
            <quote>«وَصِيَامُ رَمَضَانَ».</quote>
            <p>قَالَ هَلْ عَلَيَّ غَيْرُهُ؟ قَالَ:</p>
            <quote>«لاَ، إِلاَّ أَنْ تَطَوَّعَ».</quote>
            <p>قَالَ وَذَكَرَ لَهُ رَسُولُ اللَّهِ صلّى الله عليه وسلّم الزَّكَاةَ. قَالَ هَلْ عَلَيَّ غَيْرُهَا قَالَ:</p>
            <quote>«لاَ، إِلاَّ أَنْ تَطَوَّعَ».</quote>
            <p>قَالَ فَأَدْبَرَ الرَّجُلُ وَهُوَ يَقُولُ وَاللَّهِ لاَ أَزِيدُ عَلَى هَذَا وَلاَ أَنْقُصُ. قَالَ رَسُولُ اللَّهِ صلّى الله
            عليه وسلّم:</p>
            <quote>«أَفْلَحَ إِنْ صَدَقَ».</quote>
          </said>
        </q>
      </div>
    </body>
  </text>
</TEI>
```

Our prototype allows as to generate encoded Hadith text with TEI format. To create encoded Hadith corpus, we made an evaluation phase to test the consistency of our prototype and the adaptability and the flexibility of the TEI model for Hadith text. The evaluation phase is presented in the following section.

6 Evaluation and discussion

To evaluate our prototype, we collected the first 1000 Arabic Hadith text from 14 chapters from Sahih Bukhari book. Consequently, the prototype generates respectively 1000 TEI files representing each Hadith encoded with TEI format. Table 1 illustrates the obtained results.

Table 1. Sammery table for the prototype results.

Evaluated hadith	Hadith chapter	Total encoded	Encoded correctly	Encoded incorrectly
1000 Hadith	**14 chapter of Sahih Bukhari**	**982**	**846**	**136**
007 Hadith	chapter of revelation of Sahih Bukhari	007	006	01
051 Hadith	chapter of belief of Sahih Bukhari	051	042	09
076 Hadith	chapter of knowledge of Sahih Bukhari	075	064	11
113 Hadith	chapter of ablutions (wudu') of Sahih Bukhari	112	096	16
046 Hadith	chapter of bathing (ghusl) of Sahih Bukhari	046	039	07
040 Hadith	chapter of menstrual periods of Sahih Bukhari	038	032	06
015 Hadith	chapter of ablution with dust of Sahih Bukhari	014	009	05
172 Hadith	chapter of prayer (salat) of Sahih Bukhari	164	150	14
082 Hadith	chapter of time of the prayer of Sahih Bukhari	081	066	15

273 Hadith	chapter of call to prayer of Sahih Bukhari	270	234	36
066 Hadith	chapter of Friday prayer of Sahih Bukhari	066	056	10
006 Hadith	chapter of fear prayer of Sahih Bukhari	006	005	01
042 Hadith	chapter of the two festivals of Sahih Bukhari	041	037	04
011 Hadith	chapter of the Witr prayer of Sahih Bukhari	011	010	01

Table 1 shows that sometimes for particular Hadith texts, we can obtain erroneous encoding. The total number of obtained encoded Hadith texts is 982. The program succeeded to produce 846 Hadith encoded correctly. However, we found 136 Hadith incorrect or incomplete encoded which require some rectification on our TEI model to obtains more encoding coverage for some particular part in some types of Hadith text. This problem of incorrect or incomplete encoding is related with some exceptions and particular Hadith forms such as irregularity in *Matn* and *Isnad* position or combination between two or more chain of transmitters referring a same *Matn*. Also, the prototype misses the encoding of some Hadith because of their exceptional forms: some Hadiths refer directly to prefix Hadiths and came without *Isnad* or at least they refer to the *Isnad* of their prefix Hadiths. Indeed, we estimate the quality of our work manually. Table 2 illustrates the obtained values of precision, recall and F-score.

Table 2. Summary table of the precision, recall and F-score.

Hadith corpus	Precision	Recall	F-score
1000 Hadith	0,86	0,85	0,85

According to the value of precision, we conclude that the value of precision is worth 0.86. Also, the recall value is 0.85. These values provide an F-measure equal to 0.85.

Consequently, we conclude that the obtained results are encouraging. Besides, we can say that this prototype is flexible and easy to maintain because it is based on an object-oriented programming language. However, we handled some problems. Some of them are related with the particular Hadith forms which need to integrate more specificity and to develop our TEI model to cover the encoding of the particular and the complex forms of Hadith text.

7 Conclusion and perspectives

The normalization of Al-Hadith Al-Shareef can take the automatic processing of such corpus to another level. In this work, to attain our main objective, we based on TEI guidelines to elaborate a TEI module customized for Hadith structure. To achieve that, first, we started with a deep study of Hadith text structure. Second, we identified the data categories from the TEI standard register which harmonized with the language specification. After that, we elaborated a prototype for the automatic processing of the encoding of Hadith text with our TEI model. Then, we tested our prototype with a 1000 Hadith text from 14 chapters from Sahih Bukhari book. The elaborated

prototype allowed us to generate normalized Hadith texts. As mentioned, the obtained values of measures show that the results obtained from our prototype are encouraging. These results can be used in others levels of analyses.

As perspectives, we want improve our TEI modeling of Al-Hadith Al-Shareef by incorporating other criteria and specifications for deeper encoding of the fundamental fragments of the Hadith. Also, we need to integrate more TEI elements in the Hadith model to reach a deep description for the exceptional part in Hadith structure. Besides that, we want to improve our prototype to generalize our method to cover complex Hadith structures.

REFERENCES

1. Dalloul Y. M.: An Ontology-Based Approach to Support the Process of Judging Hadith Isnad, Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Information Technology, March (2013).
2. Azmi A. and Badia N.: iTree – Automating the Construction of the Narration Tree of Hadith, IEEE, (2010).
3. Alrfaai S.: عناية العلماء بالاسناد وعلم الجرح والتعديل و أثر ذلك في حفظ السّنة النبوية, AlMadina Almonawara, (2004).
4. Alaskalani A.: تقريب التهذيب, Society AlRisala, Bayrout, Labunan, (2008).
5. Boella M. et al.: The SALAH Project: Segmentation and Linguistic Analysis of hadīṯ Arabic Texts, Proceeding of the seventh Asia Information Retrieval Societies Conference, Springer, Heidelberg, (2011).
6. Baraka R.: Building Hadith Ontology to Support the Authenticity of Isnad, International Jornal on Islamic Applications in Computer Science and Technology, Vol.2, Issue 1, 25-39, December (2014).
7. Alkhatib M.: Classification of Al-Hadith Al-Shareef using data mining algorithm, October (2010).
8. Sayoud H.: Author discrimination between the Holy Quran and Prophet's statements, Literary and Linguistic Computing, Vol. 27, No. 4, (2012).
9. Saloot M.A. et al.: Hadith data mining and classification: a comparative analysis, Published online: 8 January (2016).
10. Burnard L. and Sperberg-McQeen C.M.: TEI P5: Guidelines for Electronic Text Encoding and Interchange, Text Encoding Initiative Consortium, Version 3.0.0. revision 89ba24e, (2016).
11. Dufournau N., Demonet M-L. and Uetani T. : "Manuel d'encodage XML-TEI Renaissance et temps modernes Imprimés-manuscrits," Version Beta, UMR 6576, (2008).
12. Ide N. and Veronis J.: Une application de la TEI aux industries de la langue: le Corpus Encoding Standard, Cahiers GUTenberg, n°24, pp. 166–169, (1996).
13. Burnard L. and Sperberg-McQeen C. M.: La TEI Simplifiée : Une introduction au codage des textes électroniques en vue de leur échange, Cahiers GUTenberg, n°24, pp. 23–151, (1996).
14. Maraoui H., Haddar K. and Romary L.: Modeling of Al-Hadith Al-Shareef with TEI, ICEMIS Conference, (2017).
15. Bin Ismail S. et al.: ALIF editor for generating Arabic normalized lexicons, ICICS Conference, (2017).
16. Maraoui H. and Haddar K. : "Automatisation de l'encodage des lexiques arabes en TEI," In 2nd conference on CEC-TAL, Sousse, Tunisia, (2015).

17. Alturki M.B.: البيان و التبين لضوابط ووسائل تمييز الرواة المهملين, King Saoud Univercity, (2007)
18. Alaskalani A.: تقريب التهذيب, Society AlRisala, Bayrout, Labunan, (2008).
19. Romary L. and Wegstein W.: “Consistent modeling of heterogeneous lexical structures,” In Journal of Text Encoding Initiative, Issue 3, TEI and linguistics, (2012).
20. Azmi A. and Bin Badia N.: e-NARRATOR -An Application for Creating an Ontology of Hadiths Narration Tree Semantically and Graphically. Arabian Journa for Science and Engineering, 35(2 C), 51-68, (2010).