Segmentation tool for hadith corpus to generate TEI encoding

Hajer Maraoui, Kais Haddar, Laurent Romary

Segmentation tool for hadith corpus to generate TEI encoding

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

[1] Faculty of Sciences of Tunis, University of Tunis El Manar, MIRACL Laboratory, Tunisia
hajer.maraoui@fst.utm.tn
[2] Faculty of Science of Sfax, University of Sfax, MIRACL Laboratory, Tunisia
kais.haddar@yahoo.fr
[3] Inria, Team ALMAnaCH, Germany
laurent.romary@inria.fr

Abstract. A segmentation tool for a hadith corpus is necessary to prepare the TEI hadith encoding process. In this context, we aim to develop a tool allowing the segmentation of hadith text from Sahih al-Bukhari corpus. To achieve this objective, we start by identifying different hadith structures. Then, we elaborate an automatic processing tool for hadith segmentation. This tool will be integrated in a prototype allowing the TEI encoding process. The experimentation and the evaluation of this tool is based on Sahih al-Bukhari corpus. The obtained results were encouraging despite some flaws related to exceptional cases of hadith structure.

Keywords: Hadith corpus, hadith structure, segmentation tool, TEI encoding.

1 Introduction

A segmentation phase is necessary to analyse and standardize corpora especially for hadith corpora. In fact, the automatization of the segmentation of large hadith corpora guarantee the treatment optimization in term of time and results precision. Moreover, it facilitates the processing and the manipulation for some tools that analyze and standardize each part of the text in hadith corpus.

However, this phase imposes some problems especially with the structure of large hadith corpora, which characterised with text particularity and variant features. Furthermore, these corpora are originally in Arabic language, which impose some difficulties that related with language specificities.

In this context, to segment a large corpus of hadith, such as Sahih al-Bukhari, we must first start with a deep study on hadith corpora specificities and identify the features of hadith text. Then, we must classify the separator terms that recognise the different units in hadith corpus and develop the necessary resources to realise the

segmentation process. After that, we need to develop a tool for the segmentation of hadith corpus.

In the present paper, we begin with a state of art on hadith in Sahih al-Bukhari corpus. Then, we illustrate the features of hadith text. After that, we present our segmentation tool for hadith text from sahih al-Bukhari. Then, we continue with by an evaluation step. We cloture our paper with a conclusion and some perspective.

2 Stat of art on hadith in Sahih al-Bukhari corpus

Sahih al-Bukhari (or صحيح البخاري in Arabic) is one of the six major hadith collections of Sunni Islam. The Arabic word sahih (صحيح) translates as authentic or correct. The prophetic traditions (or hadiths), were collected by the Muslim scholar Muhammad al-Bukhari, after being transmitted orally for generations [4]. Al-Bukhari went to Mecca when he was 16 years old and learned hadith there. He traveled around other Islamic countries to collect hadiths. He had collected 600,000 hadiths of which he only considered 7,275 ones as authentic in his well-known work, Sahih al-Bukhari. He finished his work around 846/232 AH [5]. The book is considered as one of the two most authentic and trusted collections of hadith along with Sahih Muslim.

Sahih al-Bukhari covers almost all aspects of life in providing proper guidance of Islam such as the method of performing prayers and other actions of worship directly from the Prophet Muhammad. In Islamic terminology, the term hadith refers to reports of statements or actions of the Prophet Muhammad, or of his implicit approval or criticism of something said or done in his presence. Classical hadith specialist Ibn Hajar al-Asqalani says that the intended meaning of hadith in religious tradition is something attributed to Muhammad but that is not found in the Quran. The two major aspects of a hadith are the *Matn* which is the actual narration, and *Sanad* or *Isnad* which is the chronological list of narrators. Each *Isnad* reporter mentions the person from whom he heard the hadith all the way to the prime reporter of the *Matn* itself. The *Isnad* was an authentication of the hadith to verify that it is actually come from the Prophet Muhammad. The *Isnad* means literally 'support', so it is the support in determining the hadith authenticity or weakness (more details in [4] and [6]).

Research in the *Isnad* is very important in the science of hadith. To define the hadith authentication, the traditional methods consist of following clear steps in the judgment on the *Isnad*. Currently, software tools allow hadith judging like electronic hadith encyclopedias and some websites. Moreover, information retrieval and search engines that related to semantic web can be used to serve in deciding the authenticity of hadith.

Many projects worked on hadith corpus. Indeed, these projects focused on several branches of researches such as hadith ontology, linguistic analyzing, Hadith segmentation, classification and the mining of information.

In [7], the researchers proposed a model named SALAH for the unsupervised segmentation and the linguistic analysis of the hadith texts. The model automatically segments each text unit in *Isnad* and *Matn*. After that, a personalized augmented version of the AraMorph morphological analyzer (RAM) examines and annotates

lexically and morphologically the text content. The system generates as final output a graph with relations among transmitters and a lemmatized text corpus in XML format.

In [8], the author constructed an ontology-based Isnad Judgment System (IJS) that automatically generates a suggested judgment of Hadith Isnad. This is based on the standard instructions followed by the Hadith scholars to judge hadith Isnad. A prototype of the approach implemented to provide a proof of concept for the requirements and to verify its accuracy.

Authors of paper [9] 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.

In [10], 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.

3 Arabic hadith features: Sahih al-Bukhari corpus

Sahih al-Bukhari is arranged like books in Islamic jurisprudence (or " فِقه ", fiqh). It contains also some additional sections focusing on different topics such as the origins of creation, the exegesis of the Quran and the Holy prophets. Al-Bukhari also expressed his own views of the issues in fiqh by classifying the sections of his book and assigned for each section a significant title which express the fiqh of al-Bukhari. These titles involve all the interpretations and explications of al-Bukhari to clarify the meaning of hadiths which are difficult to understand.

Sahih al-Bukhari corpus is structured into 97 chapters and 3450 sections. The chapters contain 7563 hadith in totality. This number includes repeated hadiths. It confirmed contains 7275 ones, and setting aside the repeated ones, as Nawawi says, it includes 4,000 hadiths, though Ibn Hajar says that on this count, it contains 2,761 hadiths. Moreover, the number of its hadiths is different in different versions of the book; for example, Firabri has cited 300 hadiths more than those cited by Ibrahim b. Ma'qil al-Nasafi, and the latter cited 100 hadiths fewer than those cited by Himad b. Shakir al-Nasawi.

3.1 Hadith features

Each hadith in Sahih al-Bukhari is cited in a relative section, titled by the author to express his interpretation of the meaning of hadith, under a chapter combine the sections related with one topic. Moreover, each hadith is quoted with an order number and finish with the references. This structure in accurately respected in Sahih al-Bukhari. As an example, Fig. 1 present hadith number 3209 from chapter 59 “beginning of creation”, section 6 “the reference to angels”, as it occurs in sahih al-Bukhari.

59 - كتاب بدء الخلق (...)
6 - باب ذِكْرِ الْمَلاَئِكَةِ (...) 3209 - حَدَّثَنَا مُحَمَّدُ بْنُ سَلاَمٍ، أَخْبَرَنَا مَخْلَدٌ، أَخْبَرَنَا ابْنُ جُرَيْجٍ، قَالَ: أَخْبَرَنِي مُوسَى بْنُ عُقْبَةَ، عَنْ نَافِعٍ، قَالَ: قَالَ أَبُو هُرَيْرَةَ رَضِيَ اللَّهُ عَنْهُ، عَنِ النَّبِيِّ صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ، وَتَابَعَهُ أَبُو عَاصِمٍ، عَنِ ابْنِ جُرَيْجٍ، قَالَ: أَخْبَرَنِي مُوسَى بْنُ عُقْبَةَ، عَنْ نَافِعٍ، عَنْ أَبِي هُرَيْرَةَ، عَنِ النَّبِيِّ صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ، قَالَ: " إِذَا أَحَبَّ اللَّهُ العَبْدَ نَادَى جِبْرِيلَ: إِنَّ اللَّهَ يُحِبُّ فُلاَنًا فَأَحْبِبْهُ، فَيُحِبُّهُ جِبْرِيلُ، فَيُنَادِي جِبْرِيلُ فِي أَهْلِ السَّمَاءِ: إِنَّ اللَّهَ يُحِبُّ فُلاَنًا فَأَحِبُّوهُ، فَيُحِبُّهُ أَهْلُ السَّمَاءِ، ثُمَّ يُوضَعُ لَهُ القَبُولُ فِي الأَرْضِ ". [طرفاه في 6040, 7435].

59 – Chapter of beginning of creation (…)
6 – Section the reference to angels (…) 3209 - The Prophet (peace and blessing upon him) said, "If Allah loves a person, He calls Gabriel saying, 'Allah loves so and-so; O Gabriel! Love him.' Gabriel would love him and make an announcement amongst the inhabitants of the Heaven. 'Allah loves so-and-so, therefore you should love him also,' and so all the inhabitants of the Heaven would love him, and then he is granted the pleasure of the people on the earth". [Links in 6040, 7435].

Fig. 1. Hadith number 3209 from chapter 59, section 6 from sahih al-Bukhari.

In order to extract hadiths text from Sahih al-Bukhari, it is convenient to keep the relative coordination of each hadith to maintain the structure and authentication support.

3.2 Hadith Structure

The typical hadith structure consists of two parts, the transmitter chain (*Isnad*) and the actual narration (*Matn*). Since just close sets of words separate both *Isnad* from *Matn* and the various transmitters inside *Isnad* one from another, this explicit organization allows to detect and retrieve information with a relatively small amount of ambiguity.

Despite the regular structure of the majority of hadiths occurred in Sahih al-Bukhari, it still exists others with a different structure. These cases are related with implicit parameters which are referenced or mentioned in implicit way. For example, it common to find a hadith text with only the *Matn*. But also cited with an expression of the reference where to find the appropriate isnad or came as a continuation of the Matn of a previous hadith. As an Example, Fig. 2 illustrate this case in hadith number 654 from the section 32, chapter 10 “Call to prayers”.

10 - كتاب الأذان (...)
32 – باب فضل التَّهجير إلى الظّهر(...) 537 - « وَلَوْ يَعْلَمُونَ مَا فِي التَّهْجِيرِ لاَسْتَبَقُوا إِلَيْهِ ، وَلَوْ يَعْلَمُونَ مَا فِي الْعَتَمَةِ وَالصُّبْحِ لأَتَوْهُمَا وَلَوْ حَبْواً ». [أطرافه 615، 721، 2689 - تحفة 12570- 12577].

537 - “And if they knew the reward of offering the Zuhr prayer early (in its stated time), they would race for it and they knew the reward for 'Isha' and Fajr prayers in congregation, they would attend them even if they were to crawl”. [Links in 615, 721, 2689_ 12570 – 12577].

Fig. 2. Hadith number 654 from chapter 10, section 32 from sahih al-Bukhari.

The hadith in this case came to prove the previous one and occurs elsewhere with the reference indicated in the end of the *Matn*. In other context, other forms of hadith are more abstract and follow an irregular structure. The most of these hadith came to prove the previous hadith.

In Sahih al-Bukhari occurs also other form of hadiths. These traditions contain a complex structure in the part of mentioning the Isnad. Frequently, a hadith can have more than one chain of narrators as Isnad. These chains can occur in the front of the text cited after each other or divided in two sections before and after the Matn. As Example of this composition, Fig. 3 present hadith number 6 from section 1 "How was the revelation on the Prophet", chapter 1 "Revelation".

1- كتاب بدء الوحي (...)
1 - باب كَيْفَ كَانَ بَدْءُ الْوَحْيِ إِلَى رَسُولِ اللَّهِ (...) 6 - حَدَّثَنَا عَبْدَانُ قَالَ أَخْبَرَنَا عَبْدُ اللَّهِ قَالَ أَخْبَرَنَا يُونُسُ عَنِ الزُّهْرِيِّ ح وَحَدَّثَنَا بِشْرُ بْنُ مُحَمَّدٍ قَالَ أَخْبَرَنَا عَبْدُ اللَّهِ قَالَ أَخْبَرَنَا يُونُسُ وَمَعْمَرٌ عَنِ الزُّهْرِيِّ نَحْوَهُ قَالَ أَخْبَرَنِى عُبَيْدُ اللَّهِ بْنُ عَبْدِ اللَّهِ عَنِ ابْنِ عَبَّاسٍ قَالَ كَانَ رَسُولُ اللَّهِ صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ، أَجْوَدَ النَّاسِ ، وَكَانَ أَجْوَدُ مَا يَكُونُ فِى رَمَضَانَ حِينَ يَلْقَاهُ جِبْرِيلُ ، وَكَانَ يَلْقَاهُ فِى كُلِّ لَيْلَةٍ مِنْ رَمَضَانَ فَيُدَارِسُهُ الْقُرْآنَ ، فَلَرَسُولُ اللَّهِ صَلَّى اللهُ عَلَيْهِ وَسَلَّمَ، أَجْوَدُ بِالْخَيْرِ مِنَ الرِّيحِ الْمُرْسَلَةِ. [أطرافه 1902، 3220، 3554، 4997 - تحفة 5840 - 1/5].

1 – Chapter of revelation (…)
1 – Section of how the revelation was for the prophet (…) 6 - Allah's Messenger (peace and blessing upon him) was the most generous of all the people, and he used to reach the peak in generosity in the month of Ramadan when Gabriel met him. Gabriel used to meet him every night of Ramadan to teach him the Qur'an. Allah's Messenger (peace and blessing upon him) was the most generous person, even more generous than the strong uncontrollable wind (in readiness and haste to do charitable deeds). [Links in 1902, 3220, 3554, 4997 _

Fig. 3. Hadith number 6 from chapter 1, section 1 from sahih al-Bukhari.

Based on sahih al-Bukhari corpus study, we determinate that to realize a hadith normalization system, we must start with the prime unite of sahih al-Bukhari corpus which is the hadith text. In fact, to reach the corpus normalization, a hadith extraction and segmentation phase is essential to prepare to the succeeding phase for encoding hadith text with TEI. This lead as to investigate the development of an extraction and segmentation tool integrated in our Hadith encoding prototype to split out the hadith text for the next encoding step which is presented in the next section.

4 Hadith segmentation tool

To implement the segmentation method of hadith text from Sahih al-Bukhari corpus, we developed a tool for this process. This tool is made to prepare the hadith text for the hadith encoding prototype with TEI (more details in [2] and [3]). The implementation is based on JAVA language and the API JDOM Library. Fig. 4 presents the general architecture of this program. This tool is developed to extract automatically each hadith text from Sahih al-Bukhari then segment each one to *Isnad* and *Matn* which goes as input to the hadith encoding process.

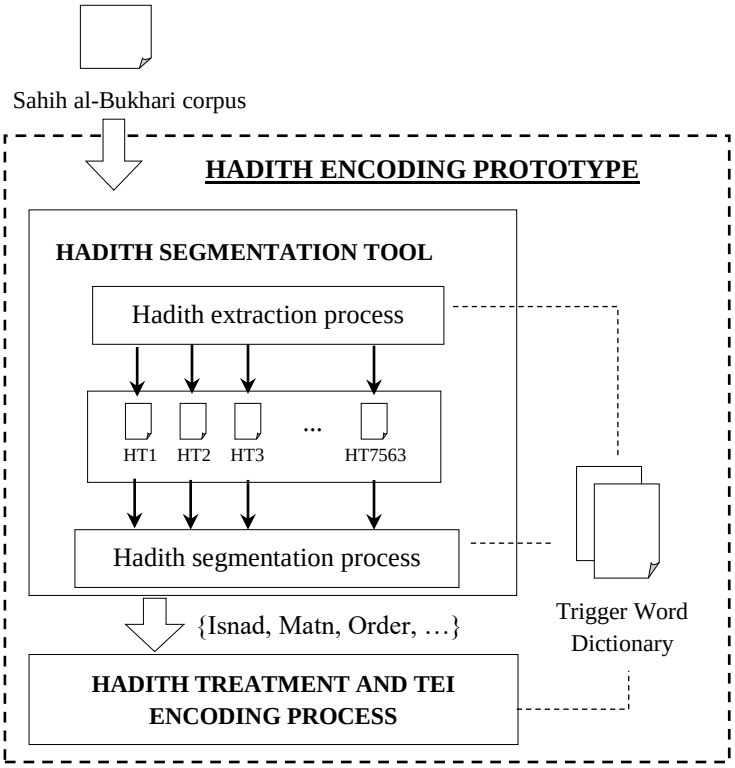

Fig. 4. General architecture of the segmentation tool of hadith corpus.

In the processing of our hadith encoding prototype, the system need as an input the *Isnad* and the *Matn* of each hadith in the used corpus. This input data is the output of the presented tool. To prepare hadith text for the encoding prototype, the system requests the user to choose an external file, with .txt extension which contains the hadith corpus in Arabic language, by selecting the file path. Then, the program reads the corpus and identify each hadith text. To detect hadith text in the corpus, we based on the fact that each hadith starts with a hadith number and a trigger word, as we name it, that express the beginning of the *Isnad*. We built for this reason a trigger word dictionary in XML format for the words that indicate the hadith beginning. Trigger word dictionary includes also the terms that separate the narrator names and indicates the commencement of the *Matn*. The trigger word identification allows the manipulation of hadith text. We classify these parts of speech to three classes: Trigger Word Before (TWB), Trigger Word After (TWA) and Trigger Word Between (TWT). Table 1 illustrates the different forms of these terms in the trigger words dictionary.

Table 1. Sammery table for the trigger words classification.

Trigger Word	Example
Trigger Word Before (TWB)	حدّثنا (He told us) – أخبرني (He told me) – أخبرنا (He told us) – سمعت (I heard) – عن (from) – أنّ (That) – سأل (He asked) – زادنا (he enhanced) – جاء (He came) – ذكر (He mentioned) – زعم (He claimed that) – لقيت (I found) – وتابعه (He continued after him) ...
Trigger Word After (TWA)	قال (He said) – يقول (says) – قالت (She said) …
Trigger Word Between (TWT)	أنّه (That he) – أنها (That she) – أخبرني (He told me) ...

The first-class TWB comprise the terms that appears before each narrator name such as "حَدَّثَنَا مُحَمَّدُ بْنُ سَلاَمٍ" (Mohamed Ben Salam told us) or a person indication "عَنْ نَافِعٍ" (Transmitted from Nafaa). The second-class TWA is for the words that occurs after a narrator name such as "عَبْدُ اللَّهِ قَالَ" (Abdu Allah said). The third-class TWT is for the trigger word that appears between two expressions of narrator names such as "أَنّه قَال" (That he said). The TWT class include all the terms that relate two sections of trigger word and narrator name. This dictionary allows the prototype, including the integrated programs, to detect the right parts and segment the hadith text. The tool connects to this dictionary and spot the hadith text.

The termination of each hadith is identified by the hadith references numbers and the word "أطرافه" or "تحفه" (which have as meaning, references and links) or a flexion form of them. This method allows the treatment of Sahih al-Bukhari Arabic corpus with no need to modify the input file or to delete any other data exists in the corpus else then hadith texts. After this step, the program extracts each spotted hadith text and save it in the workspace as text file. As well, the segmentation tool completes the reading and the separation of the *Isnad* and the *Matn* from each hadith text. To achieve the hadith segmentation, we start by analyzing Sahih al-Bukhari corpus and identifying the terms that separate the two parts. These terms are included in the trigger word dictionary. Also, the segmentation process of *Isnad* and *Matn* is based on trigger word.

Hadith corpus segmentation tool is an integrated program in the hadith encoding prototype which allows the encoding of all the hadiths of Sahih al-Bukhari in one running action. This automatic preparation of hadith text supported the optimization of hadith corpus encoding using prototype. We have evaluated this tool based on Sahih al-Bukhari corpus. The results are presented in the following section.

5 Evaluation and discussion

To evaluate the hadith corpus segmentation tool, we select as input a text file containing Sahih al-Bukhari corpus which include 7563 hadith. Consequently, the tool system creates respectively 7485 hadith text files and save them. Then the system

segments each hadith and sends the Isnad and the Matn to the encoding process in the prototype core. Table 2 illustrates the obtained results.

Table 2. Sammery table for the hadith extraction and segmenation tool results.

Evaluated corpus	Extracted hadiths	Correct extraction	Segmented hadiths	Correct segmentation	Incorrect segmentation
Sahih al-Bukhari 7563 hadith 97 Chapter	**7563**	**7563**	**7563**	**7259**	**304**

Table 2 shows that sometimes for particular Hadith texts, we can obtain erroneous segmentation. The total number of hadith text in Sahih al-Bukhari is 7563 hadith from 97 different Chapters. The hadith segmentation tool succeeded to extract 7563 Hadith correctly and save them as text files. Then, the segmentation process separate hadith fragments correctly for 7259 hadith texts. However, we found 304 Hadith that were not segmented properly. These hadith have particular structure which require some rectification to cover these cases. Indeed, we estimate the capacity of this tool manually. Table 3 illustrates the obtained values of precision, recall and F-score.

Table 3. Summary table of the precision, recall and F-score.

Hadith corpus	Precision	Recall	F-score
7563 Hadith	0,96	0,96	0,96

According to the value of precision, we conclude that the value of precision is worth 0.96. Also, the recall value is 0.96. These values provide an F-measure equal to 0.96.

Consequently, we conclude that the obtained results are encouraging. Besides, the unsupervised tool system does not require a pretreatment or execute modification on the input corpus file. However, we handled some problems. Some of them are related with the particular Hadith forms which need to integrate more specificity.

6 Conclusion and perspectives

In order to optimize the processing of hadith text, we aimed to developed a segmentation tool for hadith corpus. To achieve that, first, we started with a deep study of Hadith text structure from Sahih al-Bukhari. After that, we designed and created an integrated tool in our TEI encoding prototype for the segmentation of hadith corpus. Then, we tested this tool with a Sahih al-Bukhari corpus which englobe 7563 hadith texts from 94 chapters. As mentioned, the obtained values of measures show that the results obtained from the hadith segmentation tool are encouraging.

As perspective, the hadith segmentation tool can be used in others levels of analyses. Furthermore, we want to apply the proposed tool to other types of Hadith corpo-

ra such as Sahih Muslim. Also, we want to extend the set of rules to cover the exceptional structures of hadith text.

REFERENCES

1. 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).
2. Maraoui H., Haddar K., Romary L., Modeling of Al-Hadith Al-Shareef with TEI, ICEMIS Conference, Monastir, Tunisia, 978-1-5090-6778-7/17/$31.00 IEEE, (2017).
3. Maraoui H., Haddar K., Romary L., Encoding prototype of Al-Hadith Al-Shareef in TEI, ICALP Conference, Fez, Morocco, Springer, CCIS 782, pp. 1–13, (2017).
4. Abu Zaho M., الحديث والمحدثون, دار الفكر العربي, Riyadh, KSA, volume 1, (1984)
5. A.C. Brown, Jonathan, Hadith: Muhammad's Legacy in the Medieval and Modern World, Oneworld Publications, ISBN 978-1851686636, (2009).
6. Al-Ansari S., المقنع في علوم الحديث, Fawaz publishing house, volume 1, (1992).
7. 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).
8. 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).
9. 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, (2014).
10. Azmi A. and Badia N.: iTree – Automating the Construction of the Narration Tree of Hadith, IEEE, (2010).
11. Ben Ismail S. et al.: ALIF editor for generating Arabic normalized lexicons, ICICS Conference, 978-1-5090-4243-2/17/$31.00 IEEE, (2017).
12. Maraoui H. and Haddar K. : "Automatisation de l'encodage des lexiques arabes en TEI," In 2nd conference on CEC-TAL, Sousse, Tunisia, (2015).
13. Romary L. and Wegstein W.: "Consistent modeling of heterogeneous lexical structures," In Journal of Text Encoding Initiative, Issue 3, TEI and linguistics, (2012).
14. Alturki M.B.: البيان و التبين لضوابط ووسائل تمييز الرواة المهملين, King Saoud Univercity, (2007)
15. Alaskalani A.: تقريب التهذيب, Society AlRisala, Bayrout, Labunan, (2008).