ISSN:0975-9646

Moath Najeeb et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1336-1340

A Lexicon for Hadith Science Based on a Corpus

Moath Najeeb[#1], Abdelkarim Abdelkader[#2], Musab Al-Zghoul[#3], Abdelrahman Osman[#4]

[#]*Computer Science Department, Umm Al-Qura University*
College of Computer at Al-Qunfudh, Saudi Arabia

***Abstract*— The Lexicon is among the most important resources for written language processing and comprehension. In languages with rich morphology like classical Arabic language (The language of the Qur'an and classical literature), a careful design of the lexicon is crucial. This paper describes the building of a lexicon of Hadith science using the HPSG formalism. This formalism has seen for several years a great development in the field of NLP, especially in lexical and syntax analysis. One of the distinctive design features of HPSG is its integrated information. Information about phonology, morphology, syntax, semantics and all other components of the grammar is represented in a single structure named attribute value matrix (AVM), with the possibility of complex interactions.**

In order to ensure interoperability, reusability and the ease of exchange and transfer, we used the XML technology. Each part of hadith isnad like narrator, Telling tool, etc. is represented in a separate XML document with HPSG lexical features specifically targeted at morphological analysis. But it also specifies morphological, syntactic and semantics features (such as gender or number), which can later be used by parsers and other applications.

***Keywords*— Hadith Science, Isnad, Lexicon, Part of isnad, XML, HPSG.**

I. INTRODUCTION

Lexicons are important resources for Natural Language Processing (NLP) applications. During the last few years, researchers in the NLP domain took conscience of the stake that constitutes the electronic dictionaries and lexicons and of the importance of implementation technics. The choice of the formalism of linguistic knowledge representation must be taken into consideration in priority since it is a central element and structuring in the conception of NLP applications. So, the content, the volume and the format of the lexicon depend on the chosen formalism and encourage some linguistic treatments to others.

A lexicon must contain the maximum of knowledge on the language. According to the type of application aimed, this knowledge will be of different nature (phonological, morphological, syntactic, semantic, pragmatic, etc.).

The HPSG formalism (Head-Driven Phrase Structure Grammars) is a combination of various theories (BPSG, LFG, GB) [1] and [2]. It permits to integrate phonological, lexical, syntactic, semantic and pragmatic knowledge. It offers a rigorous framework for grammar development through its foundation on unification of feature structures. One very important part of the formalism is structure sharing.

In this paper, we used the HPSG formalism to build a lexicon for the Hadith science. This lexicon contains all necessary information for the design of a robust parser in hadith science field.

Thanks to the Internet, lexical data resources for each field are growing rapidely. Unfortunately, in hadith science field, despite numerous existing tools and resources like, Shamela [3], Aljamii [4], E-narrator [5], Dorar [6], etc. each resource has its own format and own structure. Furthermore, the existing lexicons or electronic dictionaries or databases are generally developed for a specific purpose and can't be reused easily in other applications.

In order for different tools and systems, for Arabic Natural Language Processing (ANLP) in general and Hadith science specially, to communicate and interoperate with each other, it is important to have a common language. Nowadays, the common language adopted by most organizations in different flelds is eXtensible Markup Language (XML), since XML can facilitate significant features such as personalization, interoperability, reusability, ease of exchange and flexibility.

The core idea with this work is that by organizing and disseminating the hadith science knowledges and contents into a uniform format, it is possible to achieve content reuse and interoperation between different organizations, expert scholars and students in hadith science.

The remainder of the paper is organized as follows. The second section gives an overview for hadith science. The third section describes the HPSG formalism. The fourth section defines the principle features of parts of hadith isnad. Finally, Section 5 gives a description of the HPSG based lexicon for hadith science that we implemented in XML.

II. OVERVIEW OF HADITH SCIENCE

The hadith science (علم الحديث) is a field of study in Islamic scholarship, dedicated to investigating and classifying the veracity of existing ahadith. This field includes the sayings of the Prophet Muhammad (may ALLAH bless him and grant him peace), his actions, his characteristics, stories and origins of the sayings. The narration of these Ahadith and the study of their origin and the terms of acceptance and meaning are the study of this science in addition to what benefits are extracted from them [7]-[10].

The term hadith (the plural is ahadith) is defined as follow : "Everything attributed to Prophet Muhammad (or the Companions or following generation), such as words,

deeds, explanations, or characteristics of his creation and character."[11]and [12].

The hadith science is divided into two categories:

A. Hadith science riwayah علم الحديث رواية

This category deals with the Matn's meanings and with explaining it. The subject matter of this science is related to the sayings of the Prophet (pubh), his actions, his approvals, or his natural or ethical description.

B. Hadith science dirayah علم الحديث دراية

The specialty of Science of Hadith Dirayah is the collection of rules and other affairs which through will be known the position of reporter of hadith and what they reported from aspect of acceptance and rejection. So, This category deals with the sanad, and the states of the narrators whether they are trustworthy or liars etc. [13]

The advantage of the science of hadith dirayah is to know the acceptance from its rejection. On the basis of this, these two sciences are inseparable from one to another, rather the science of hadith riwayah would be fruitless whenever it is not accompanied by the science of hadith dirayah to make it possible in knowing the acceptance from rejection [14].

C. Components of Hadith

Each hadith is composed of three parts as is shown in Fig. 1.

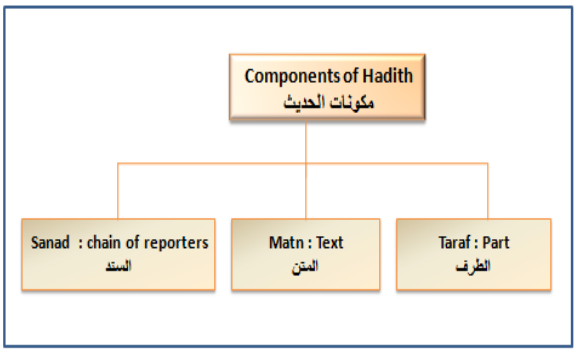
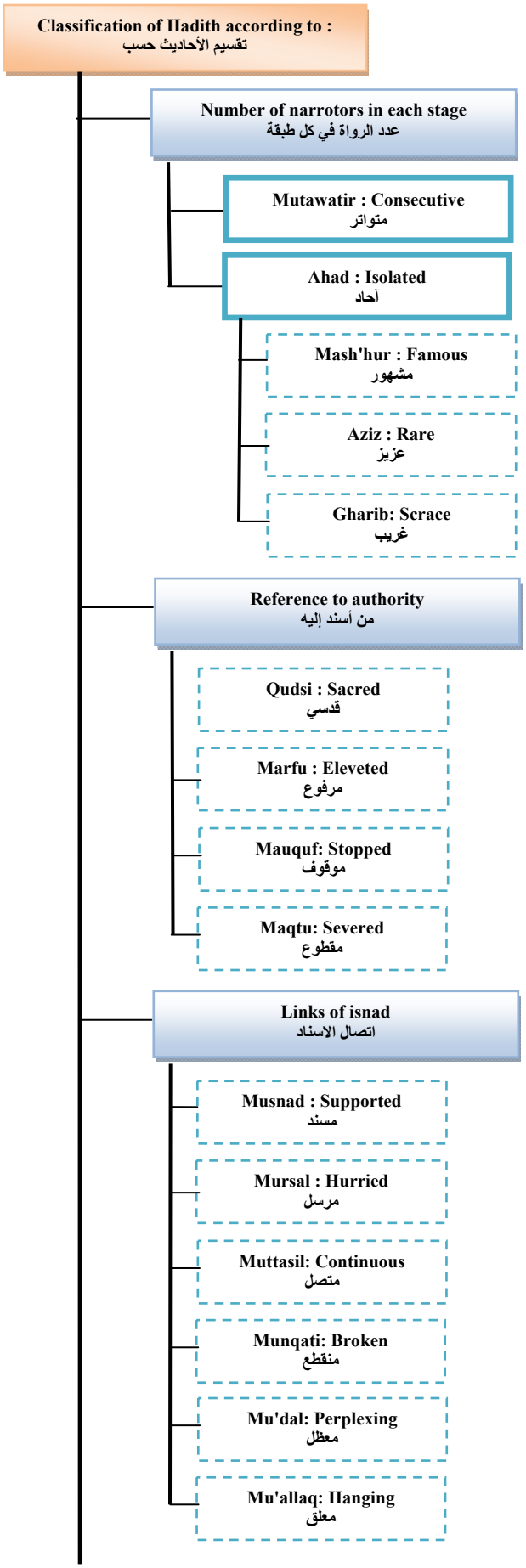


Fig. 1 Components of Hadith

The first component is the chain of narrators (reporters), also called silsila isnad or sanad (سند). This chain includes the start narrator or the 'originator' and the final narrator of the hadith. Between the start and the final narrator, there are any numbers of transmitters (narrators) who have passed on the hadith orally from one to the other. The second essential component is the text known as the 'matn' and is carried from the originator. The third part is named taraf (the part, or the beginning sentence, of the text which refers to the sayings, actions or characteristics of the Prophet)[15] and [16].

D. Classifications of Hadith

The scholars proposed seven classifications of hadith. these classifications are shown in the Fig.2 [17].



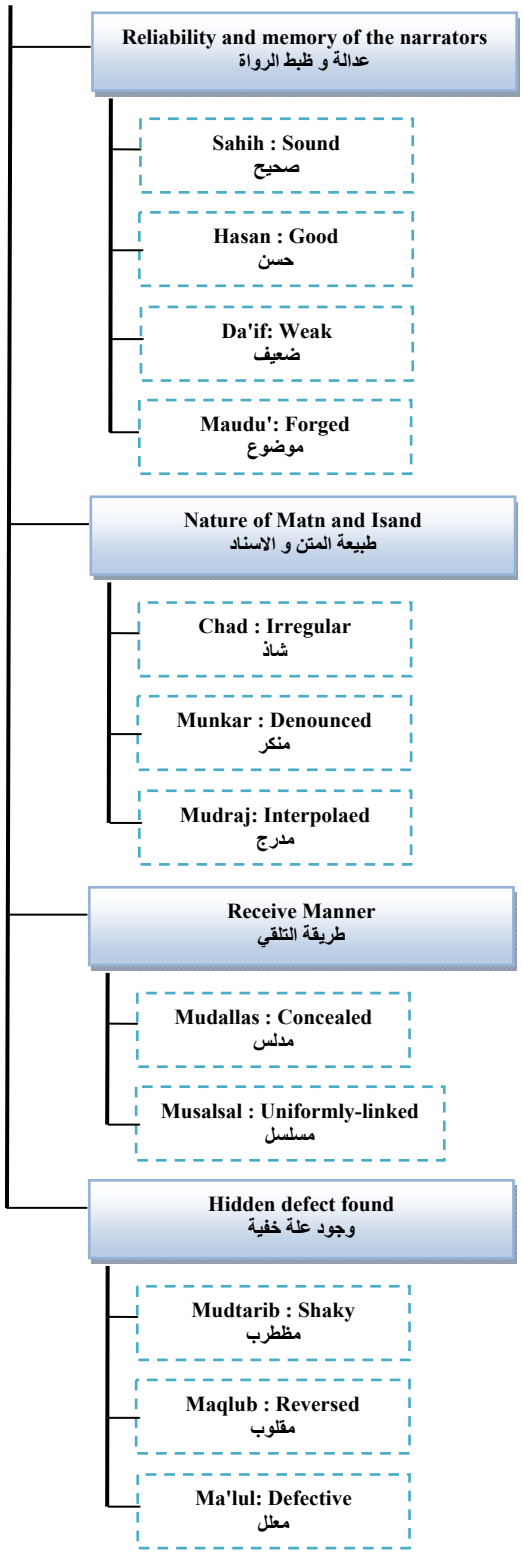


Fig. 2 Classification of Hadith

III. THE HPSG FORMALISM

Head-Driven Phrase Structure Grammars (HPSG) were first introduced in [1], while the complete theoretical framework was published later in [2]. The HPSG formalism was designed as a refinement of Generalized Phrase Structure Grammar, and similarly to GPSG, describes a linguistic theory [18].

There are several reasons for the continually increasing popularity of HPSG. Even if HPSG - as linguistic formalism - is a combination of various theories, it offers a rigorous framework for grammar development through its foundation on unification of feature structures. The support for sub-categorization and semantic representation are important qualities of HPSG. Furthermore, the founders of this theory list key components of the HPSG formalism that explain the impact it had on grammar development, such as:

- its foundation on typed feature logic,
- using the same formalism for universal and particular generalizations,
- allowing for language-specific analyses, that are less natural described in other formalisms

The Head-Driven component of the HPSG name reveals one of the underlying principles of this formalism. The most important element of a phrase is its lexical head [19]. The lexical head incorporates both syntactic information (such as part of speech and dependency relations with other constituents) and semantic information. Lexical entities, in general, are information-rich structures, with feature values from the type signature. This lexical-centered organization of HPSG eliminate redundancies, therefore the amount and complexity of phrasal rules is greatly reduced. Typed feature structures (TFS) are widely used in natural language processing and computational linguistics to enrich syntactic categories. They are very similar to frames from knowledge representation systems or records from various programming languages like C.

This formalism is based on unification. The unification of two typed feature structures is an operation that produces a feature structure that is their least upper bound, with respect to the assumption ordering. Unification fails when the two feature structures provide inconsistent information. Intuitively, unification is accomplished by first unifying the types of the root nodes of the two feature structures, and replacing them with the result of the type unification. By recursion, nodes that are values of identical features are then unified, and replaced with the result of the unification. Failure occurs when an inconsistency between nodes occurs (i.e., when two types can not unify).

Agreement phenomena involve morpho-syntax, semantics, and pragmatics, and so it is not surprising that this is another domain in which the sign-based formalism of HPSG has yielded significant results [20].

The central concept of the HPSG theory of agreement is the INDEX, which unifies some of the properties of constants and variables from logical formalisms. In the simplest cases an index is an abstract linguistic entity that is referentially linked to some object in the interpretation domain. Indices are also used with quantification, in which case they behave much like variables. Unlike constants and variables in logic, however, indices have an internal organization that reflects properties of the associated linguistic entities or referential objects. This information includes number, gender, and person. This makes it possible to straightforwardly account for a number of agreement phenomena. For instance, if we assume that person/number/gender information is encoded on indices and that the relation between reflexive pronouns and their

antecedents involves INDEX identity, then the distribution of forms in follows immediately (show Fig.3).

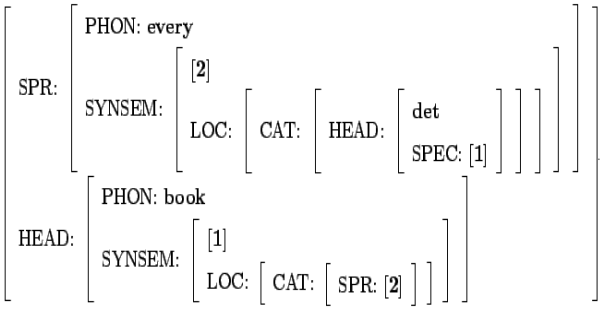


Fig. 3 Example of AVM

IV. BASIC FEATURES OF PART OF ISNAD

As we mentioned earlier, Muslims are interested of Isnad since the early centuries because it helps differentiate between the sound (accepted) and weak (rejected) Hadith [21]. The scholars of hadith judge it based on the chain of narrator and the individuals involved in the chain of narrators.

After a study isand based on discussion with experts on hadith science, we have identified the following main parts of isnad :

- Hadith : represents the Matn and Isnad of hadith
- Chain of narrators (silsila): represents a sequence of narrators.
- Narrator : represents a person who narrates (tells) Hadith.
- Author : represents the author of the book.
- Book : represents the book of Hadith.
- Chapter : represents the chapter in Hadith book
- Bab : represent a part of the chapter
- Transmission tool
- Transmission tool prefix : represents a determinant that above the transmission tool.
- Konia : represents a nickname of the narrator
- Name prefix : represents a noun that precede the narrator name.
- Prophet name : represents a name of Prophet Mohammed (Peace be upon him).

Fig. 4 shows a part of the Context Free Grammar (CFG). This grammar contains all terminals and non terminals and the production rules used to form the hadith isnad.

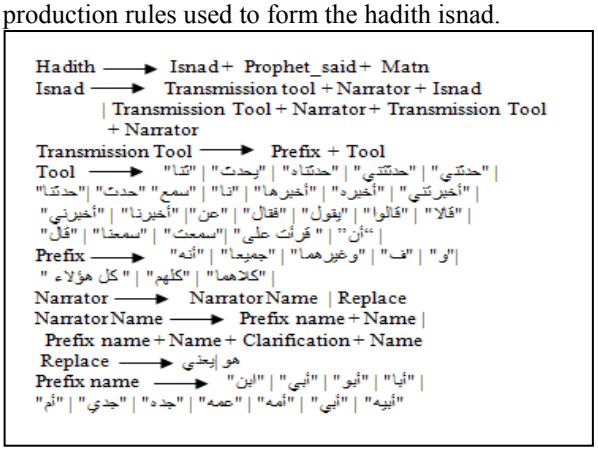


Fig. 4 A part of CFG isnad

For each part of isnad, all information about phonology, lexical, morphology, syntax and semantics is represented in a single structure named attribute value matrix (AVM). The part of isnad is called lexical entry. Fig. 5 shows an example of AVM for narrator lexical entry according to the HPSG formalism.

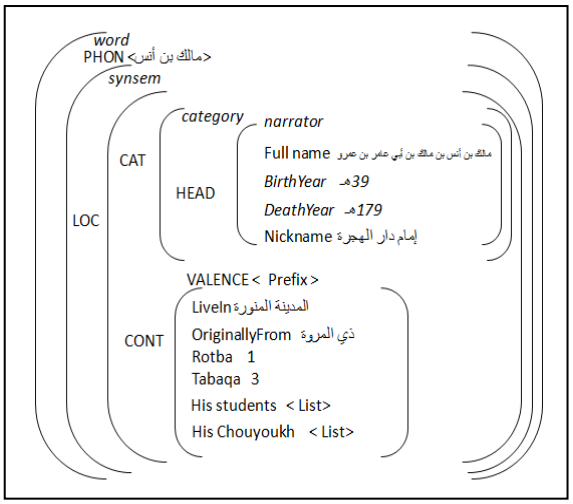


Fig. 5 Example of narrator AVM

V. IMPLEMENTATION

We used XML technology to developed our lexicon. We opted for XML in order to be able to structure, to standardize and to normalize the information used by the Hadith isnad processing or other NLP application, because it is a standard for describing how information is structured. This makes it much easier to move structured information from place to place, or from one program to another and it gives the ability to manipulate the information easily and quickly.

The lexicon is constituted of a set of XML documents. Every document concerns a part of isnad like matn, isnad, narrator, transmission tool, prefix, etc. (Fig. 6 is a part of an example of XML document).

We extracted a corpus form Alboukhari and Muslim book [22] and [23]. This books contain more than 15000 hadith. The lexicon contains all parts of isnad of this ahadith .

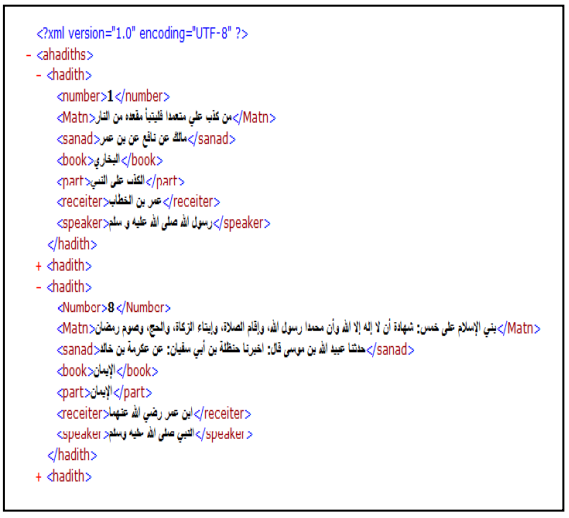


Fig. 6 Example of XML document

VI. CONCLUSIONS

This work has demonstrated that it is possible to build a reusable and interoperable lexicon for hadith science. This lexicon will be used in many applications and others systems to serve islamic studies. As perspectives, we consider, first, to extend the description of HPSG to cover other types of isnad buildings and we intend to enrich the lexicon by including other books.

ACKNOWLEDGMENT

The authors would like to thank Institute of Scientific Research and Revival of Islamic Heritage at Umm Al-Qura University (project # 34508026) for the financial support.

REFERENCES

[1] C. Pollard, and I. Sag. *Information-based syntax and semantics.* Volume 1. Fundamentals. CLSI Lecture Notes 13, 1987.

[2] C.Pollard and I.Sag, *Head-Driven Phrase Structure Grammars.* Chicago University Press, 1994.

[3] (2015)Overall Library website. [Online]. Available: http://shamela.ws/

[4] (2015)Hadith science collection. [Online]. Available: http://www.sonnaonline.com/

[5] M. Aqil Azmi and N. bin Badia, "E-narrator – an application for creating an ontology of hadiths narration tree semantically and graphically". Department of Computer Science, King Saud University, Riyadh, Saudi Arabia, 2011.

[6] (2015)website of addorar assania. [Online]. Available: http://www.dorar.net/.

[7] (2015) Official website of Khalifa Institute. [Online]. Available: http://islamic-world.net/

[8] Moath M. Najeeb, Abdelkarim A. Abdelkader and Musab B. Al-Zghoul, "Arabic Natural Language Processing Laboratory serving Islamic Sciences". International Journal of Advanced Computer Science and Applications (IJACSA), 5(3), 2014. http://dx.doi.org/10.14569/IJACSA.2014.050316 - See more at: http://thesai.org/Publications/ViewPaper?Volume=5&Issue=3&Code=IJACSA&SerialNo=16#sthash.s0cCJ4cy.dpuf.

[9] K. Bilal and S. Mohsin, "Muhadith: A Cloud based Distributed Expert
System for Classification of Ahadith", IEEE 10th International Conference on Frontiers of Information Technology, pp. 73-78, 2012. Bilal and S. Mohsin, "Muhadith: A Cloud based Distributed Expert
System for Classification of Ahadith", IEEE 10th International Conference on Frontiers of Information Technology, pp. 73-78, 2012.

[10] E. Dickinson, "An Introduction to the Science of Hadith". from the translator's introduction, pg. xiii, Garnet publishing, Reading, U.K., first edition, 2006.

[11] M. Ghazizadeh, M. Zahedi, M. Kahani, and B. Bidgoli, "Fuzzy Expert
system in determining Hadith validity", advances in computer and information sciences and engineering, PP.354-359, 2008.

[12] F. Harrag, E. El-Qawasmeh and A. Al-Salman, "Extracting Named Entities from Prophetic Narration Texts (Hadith)", ICSECS 2011, Part II, CCIS 180, pp. 289–297, 2011.

[13] M. Hyder and S. Ghazanfer, "Towards a database Oriented Hadith Research Using Relational, Algorithmic and Data-warehousing Techniques", The Islamic Culture, Quarterly Journal of Shaikh Zayed
Islamic Center for Islamic and Arabic Studies, Vol. 19, University of
Karachi, 2008.

[14] Darbala, I.M., Introductions about the graduation of the Hadith by electronic encyclopedias, 2011, Shamela Library.

[15] Azmi, A. and N. AlBadia, Mining and Visualizing the Narration Tree of Hadiths (Prophetic Traditions). Cross-disciplinary Advances in Applied Natural Language Processing: Issues and Approaches, 2012: p. 239.

[16] Laher, S. On Hadith Authentication. 2010 [cited 2015 30 Jan]; Available from: http://www.dewdropsweb.com/hadith-sahih/.

[17] Shuqairi, M. How to graduate Hadith using computer. 2009. company, R.A. Collector of Phrophet's Talks. 2012 [cited 2015 26 Jan].

[18] F.Popowich, C.Vogel, "Chart parsing head-driven phrase structure grammar". Technical Report CSS-IS TR 90-01, Simon Fraser University, 1990.

[19] Abdelkarim Abdelkader, Dalila Souilem Boumiza and Rafik Braham, "E-Assessment System Based on IMS QTI for the Arabic Grammar" International Journal of Advanced Computer Science and Applications(IJACSA), 5(10), 2014. http://dx.doi.org/10.14569/IJACSA.2014.051013 - See more at: http://thesai.org/Publications/ViewPaper?Volume=5&Issue=10&Code=IJACSA&SerialNo=13#sthash.tlqmcTaq.dpuf

[20] Dalila SOUILEM , Abdelkarim ABDELKADER , Rafik BRAHAM "E-Assessment System Based on Natural Language Processing for Arabic Language". International Journal of Computer Trends and Technology (IJCTT) V19(2):91-98, Jan 2015. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.

[21] Moath M. Najeebm, "Towards Innovative System for Hadith Isnad Processing". International Journal of Computer Trends and Technology (IJCTT) V18(6):257-259, Dec 2014. ISSN:2231-2803. www.ijcttjournal.org. Published by Seventh Sense Research Group.

[22] الإمام أبو عبد الله محمد بن إسماعيل البخاري الجعفي "الجامع المسند الصحيح المختصر من أمور رسول الله صلى الله عليه وسلم وسننه وأيامه المعروف بصحيح البخاري"، دار طيبة 2008.

[23] الإمام أبو الحسين مسلم بن الحجاج القشيري النيسابوري " المسند الصحيح المختصر من السنن بنقل العدل عن العدل إلى رسول الله صلى الله عليه وسلم المعروف بصحيح مسلم" تحقيق نظر بن محمد الفاريابي أبو قتيبة. دار طيبة 2006.