ISSN No. 0976-5697

Volume 2, No. 3, May-June 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Computational Recognition of literary forms and its Transformation to a Literature Ontology: an Experiment on Quran

Aliabbas Petiwala*
Dept. of Computer Science,
Pondicherry University
Pondicherry, India
aliabbas_aa@yahoo.com

S.SivaSathya
Dept. of Computer Science,
Pondicherry University
Pondicherry, India

***Abstract*:** The Quran being a literary masterpiece in Arabic, employs a versatile gamut of rhetorical features, literary forms and figures of speech not present in any other rhymed prose, past or present. Linguistic Scholars have meticulously enumerated and classified hundreds of figures of speech. The degree and count of figures of speech employed in a piece of natural language text can give a useful insight to the degree of literary excellence of the given text. Thus it is required to first computationally identify an assortment of figures of speech present in the literature text. Computational recognition of various figures of speech can help us quantify the literary excellence score as a function of the absolute counts of the literary features and figures of speech. This paper presents various Lexico-syntactic patterns to recognize important literary forms. The English translation of the Quran was chosen as a gold standard in the experiments owing to its rich literary composition. These literary forms are then modeled as a literary ontology in OWL web ontology language and can be harnessed by semantic web applications.

Keywords: Computational Linguistics, Literary Forms, Ontology, Semantic Web, Quran, Islam, Text Mining.

I. INTRODUCTION

During the Renaissance, scholars meticulously enumerated and classified various figures of speech. Henry Peacham, for example, in his "The Garden of Eloquence (1577)", enumerated 184 different figures of speech. The degree and count of figures of speech employed in a piece of natural language text can give a useful insight to the degree of literary excellence of the given text.The Quranic text is proposed to be used as gold standard in the experiments owing to its rich literary composition.Following is an incomplete list of the multitude of figures of speech and literary forms found in the Quran:

• Analogy (see 88:15–16 & 93:9-10) .
• Alliteration (see 33:71 & 77:20)
• Antiphrasis (see 44:49)
• Antithesis (see 35:7 & 9:82)
• Asyndeton (see 13:2)
Chiasmus (see 3:27).
• Homonymy (see 2:14-15 and 3:54) .
• Hyperbole (see 7:40, 33:10 and 39:71-72).
• Isocolon (see 65:7-10) .
• Metaphor (see 19:4 and 21:18) .
• Metonymy (see 54:13 and 6:127)
• Polypton (see 80:25-26).
• Rhetorical Questions (see 55:60 and 37:91-92)
• Stress (see Quran 29:62 and 3:92).
• Narratives (82:1-4).
• Wordplay & Ambiguity (2:61).
• Dramatic Dialogue(dialogue of Moses and Pharaoh in 26:16).

It is to be noted that the many literary forms that are less rhetorical in nature are not dependent on the language of the text since they convey literary forms like metaphors, parables, oaths etc. these are even preserved when we translate from one language to the other . This justifies our use of English translations.

II. STATE OF ART

Much work has already been done on learning domain specific ontologies from text [1][2-4].Various statistic, natural language processing techniques [5-7] have been used to represent technical domain specific knowledge as an ontology. Thetechnical documents used to learn the ontologies are themselves semantically structured and follow a consistent format and set of implicit rules for representing the domain knowledge. For example, in the biological domain which is the domain where most of the ontology learning techniques were used, there is a well-defined set of terminologies and jargon used in the biomedical literature. However, they are unable to capture rich literary forms that are present in other literatures like story books, epics, poetry etc. Work has already been done for morphological analysis [8-11] of the Arabic Quranic text and the language research group at University of Leeds are currently working to develop an ontology for the Quran which is yet not complete. Their ontology does not reflect the literary forms used in the Quran and is encoded in the now obsolete KIF format since the semantic web uses the RDF\OWL ontologies and is expert engineered. A similar approach using natural language processing without taking into consideration the various literary forms of the Quran has also been done. [ ]. The approach proposed in this paper differs significantly from the above approaches in the following lines

- Firstly, a semi-automated methodology to construct the literature ontology recognizing the various literary forms in the Quran has been used.
- An initial expert prepared ontology with easily available index terms of Quran is used as a starting point for the ontology construction process.
- WordNet based algorithm[12], [13]is also employed in developing the ontology along with Lexico

© 2010, IJARCS All Rights Reserved

Syntactic Patterns to model the various literary forms in the Quran.

III. THE UNIQUE LITERARY FORM OF THE QURAN

Although the Quran was revealed to Prophet Muhammad in Arabic speech ,It is well known amongst Muslim and Non-Muslim scholars that the Quranic discourse cannot be described as any of the known forms of Arabic speech; namely Poetry and Prose [14].Prophet Muhammad had received a book that would deal with matters of belief, legislation, international law, politics, ritual, spirituality, and economics in an 'entirely new literary form'[15]The inability of any person to produce anything like the Quran, due to its unique literary form is the essence of the Quranic miracle[14] Our work uses the interpreted meanings of the English translation of the noble Quran because an ontology inherently represents the semantics of the domain knowledge hence meanings of the Quran conveyed in the translation correspond to the actual concepts and relationships delineated in the Quran.

A closer study of the Quran can highlight a wide range and frequency of figures of speech and rhetorical features and which continue to exist in the English translations of the Quran because the semantics and the intended meaning of the verses is preserved and hence all translations of a particular verse are cross lingually semantically preserving. The following partial list referring the verses numbers has been provided to show that the Quran employs a versatile gamut of rhetorical features and literary forms not present in any other rhymed prose, past or present [14].

- Analogy (see 88:15–16 & 93:9-10).
- Alliteration (see 33:71 & 77:20).
- Metaphor (see 14:18)
- Motif (see 88:15–16 & 93:9-10).
- Oaths (see 92:1-3).
- Narratives (82:1-4).[16]
- Simile (see 2:74).
- Wordplay & Ambiguity (2:61)[16] .
- Dramatic Dialogue (dialogue of Moses and Pharaoh in 26:16) [16].
- Characterization (74:18-25)[16] .

A. Quranic Literature Ontology (QLO)

Quranic literature ontology is an ontology constructed from Quran considering all or some of the following features:

a. Similes " A figure of speech that expresses a resemblance between things of different kinds usually formed with `like' relationship or `as' relationship. The Quran has very frequently used this figure of speech to convey its message.

Example 1. "Then, even after that, your hearts were hardened and **became as rocks**, or worse than rocks, for hardness ..." (Quran 2:74)

Here the literal meaning of the word 'rocks' is negated by prefixing with the word 'like' or 'as' depending upon the translation. Henceasimile can be detected from theoccurrences of words'as', 'like', 'similitude', 'similar', 'same as'.

Hence we can define a relation:

Simile (heart:6,rock:12,2:74)

b. Metaphors: A figure of speech in which an expression is used to refer to something that it does not literally denote in order to suggest a similarity. It is an implied simile. It does not, like the simile, state that the thing is like another or acts as another, but takes that for granted and proceeds as if the two things were one.

The Quran has frequently used this figure of speech to convey its message as evident from the following verse:

Example 2. "A **similitude** of those who disbelieve in their Lord: Their works are as **ashes**, which the wind blows hard upon a stormy day. They have no control of anything that they have earned. That is the extreme failure."(Quran 14:18)

The word similitude is used throughout the Quran to delineate metaphors and parables. The distinguishing feature between a metaphor and a parable is that a metaphor always represents a concrete thing where as a parable or an allegory is represents an extended metaphor in terms of events, stories or incidents.

Here the relation :

Metaphor(ashes:14)

c. Motif: A unifying idea that is a recurrent element in a literary or artistic work. In the Quran the idea of the day of resurrection is repeated more than a hundred times.

d. Oath: Many of the early Quranic surah contain oaths, typical of which are the oaths sworn by the sun and the moon, day and night, and light and darkness. The purpose of an oath is to confirm a statement and place emphasis upon it .

Example 4. The following verses show that god takes oaths to confirm a statement and place emphasis:

(a) By the night as it envelops;

(b) And by the day as it appears in brightness;
(c) And by Him Who created male and female;

Here the keyword "By" can be used to identify the oaths
So (a) can be represented as
Oath(night:92:1)
Oath(day:92:2)

Table I: Candidate Entity Set At Different Layers of Conventional Ontology Learning Cake

Candidate Entity Set	Description
Terms	The intial input data source containing a set of key terms corresponding to the text corpus for the ontology learning process..
Synonyms	Related terms from the lower layer are clustered into Synonym set.
Concepts	Synonyms are mapped to concepts relevant to the text corpora.
Taxonomy	A Concept Hierarchy is created from the concepts.

© 2010, IJARCS All Rights Reserved

Relation Set	Relations among the concept are identified and stored in a set.

a. Dramatic Dialogue\Debate: The Quran beautifully delineates the fast paced dialogue between two parties as in the case between Pharaoh and Moses (Quran 26:16) giving valuable insight into the arguments of both Moses on one hand and Pharaoh on another.

Example 5.

Pharaoh said: And what is the Lord of the Worlds?

(Moses) said: Lord of the heavens and the earth and all that is between them, if ye had but sure belief.

Here the word "said" is used to identify the dialogue between two people.

It can be represented as

Dialogue(Pharoh:26:23)

Dialogue(Moses:26:24)

a. Narrative: Another common literary form in Quran which gives a graphic description of events both of the past and the future, the past events give the accounts of the stories of prophets like Abraham, Joseph(Yusuf) (Quran 12:1) etc. the description of future events concern with the portrayal of the day of resurrection and life after death (Quran 82:1-4).

Example 6.

We narrate unto thee (Muhammad) the best of narratives in that we have inspired in thee this Quran, though aforetime thou wast of the heedless…

Verily in Joseph and his brethren are signs (of Allah's Sovereignty) for the inquiring.

Here the word narratives\stories can be used to detect stories of the past. Which can be represented as?

Narrative(Joseph:12:1)

IV. ONTOLOGY LEARNING CAKE FOR QLO

The conventional ontology learning cake is shown in figure 1 [17]. Figure-2 shows the extended version of it used in this proposed work to accommodate the needs of recognizing the literary forms in the Quran. The **c**andidate Entity Set(CSE) used in this work is shown in Table-1, which is defined as the set of recognized terms at the n^{th} layer of the ontology learning

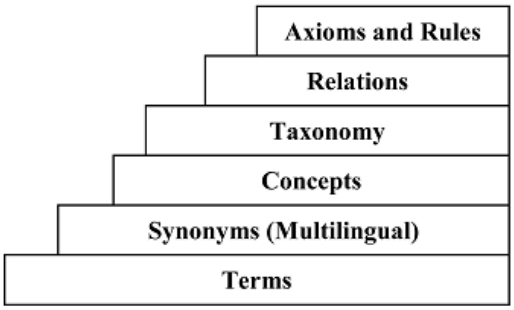

Figure1.Conventional Ontology Learning Cake

The ontology learning cake of **Error! Reference source not found.** delineates the transformation of key terms of the text corpora into higher level abstract entities like synonyms, concepts, Taxonomy and relations. The approach in figure 1 does not take into consideration the learning of various literary forms. Hence the conventional learning cake can be elaborated to accommodate the learning of advanced literary forms, themes or topics. This is shown at the highest abstraction level in Fig. 2, thus the cardinality of the candidate entity set at a lower level is higher than the next layer.

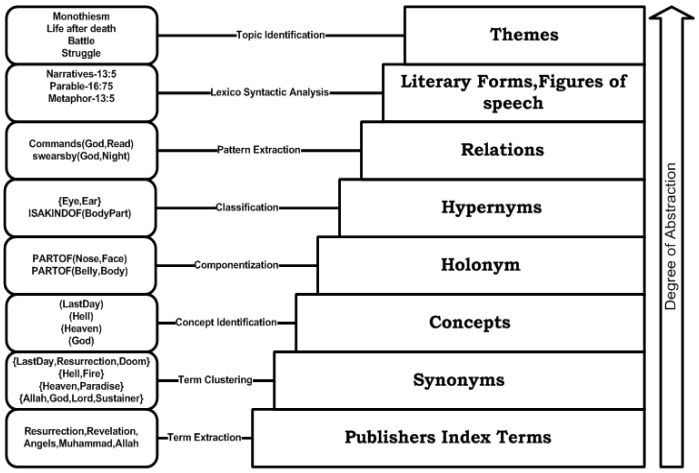

Figure 2.Ontology Learning Cake for QLO.

A. Brief over view of the Phases involved in learning the QLO

- **Term Extraction**: Important terms as listed in the publisher supplied index file are extracted to get a consistent list of terms which represent the document.
- **Term Clustering**: WordNet is used to cluster the terms by grouping the terms having common synonyms. The output of this phase is called Syncluster.
- **Concept Identification**: Concept Identification is performed in a semiautomatic fashion by selecting one of the terms in the set SynCluster as the concept.
- **Componentization**: Holonym relationships provided by WordNet are utilized to get the "part of" relationship thus aggregating the concepts.
- **Classification**: Hypernym relationships provided by WordNet is utilized to get the "is a" relationship, thus classifying the concepts.
- **Pattern Extraction**: Relationships between concepts are identified based on syntactic patterns that exist between two concepts.
- **Lexico Syntactic Analysis**: Natural language processing techniques are used to extract the various literary forms by using a custom built database of Lexico syntactic patterns.
- **Topic Identification**: Themes or recurring main ideas of Quran are recognized as topics, topics give a brief view of multiple concepts.

V. METHODOLOGY

The methodology and architecture used to generate literature ontology from an annotated Quran corpus is depicted in figure 2.

It is required that the ontology cover all the key terms present in the extracted Key Phrases list.To evaluate the term coverage the Ontology Term Coverage Metric (OTCM) is used.

This metric act as a feedback for the ontology Conversion

© 2010, IJARCS All Rights Reserved

algorithm, hence it is instantaneously calculated each time the ontology evolves from a prior state.

Lexico literary Syntactic rules have to be developed manually for each literary form to identify the literary features from the text As shown in figure 2, the annotated Quranic corpus is first tokenized and sentences are separated by the sentence splitter. The Stanford POS tagger is used to generate the POS tags and dependency between the words in the sentence. The JAPE transducer is used to extract the patterns based on the literary pattern database. This database of lexico-syntactic patterns is formulated by Quranic domain expert.

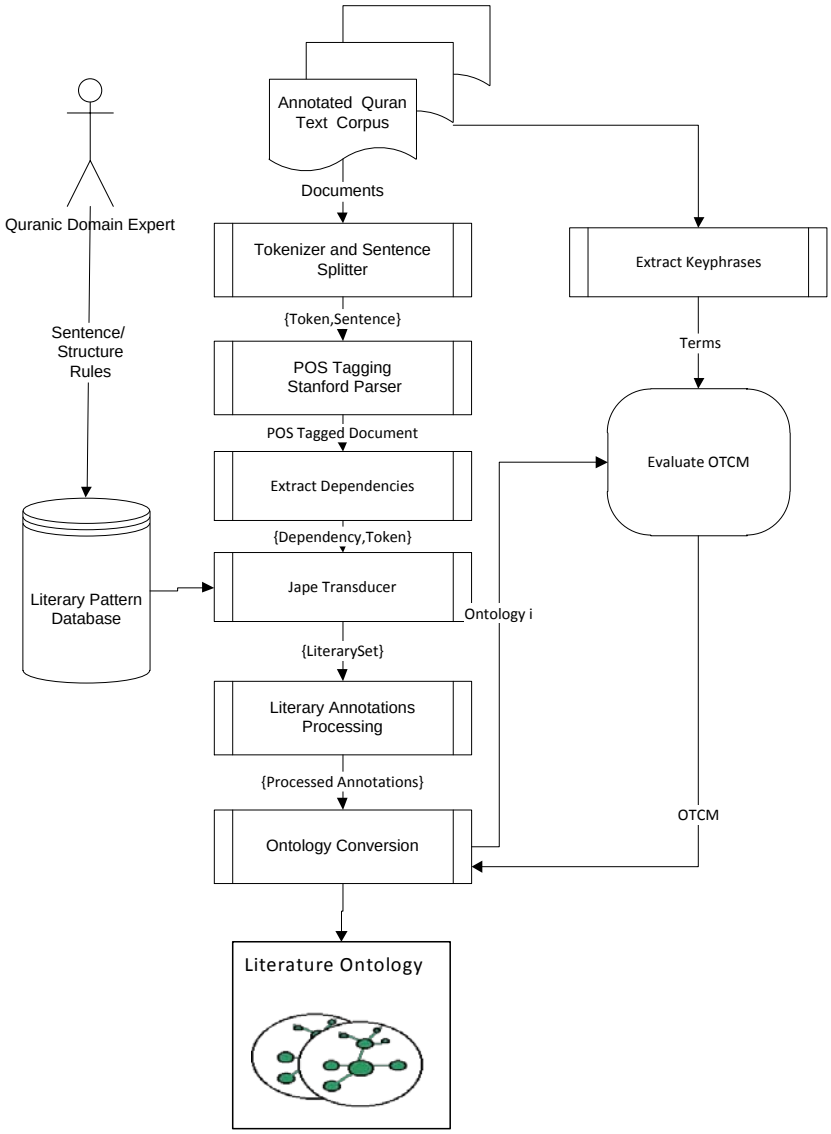

Figure2.Pipeline for Extracting Literary forms.

B. Pipeline for Extracting Literary forms

Literary forms are recognized by the pipeline shown in Fig. 3 It has the following phases:

1. *Tokenizer and Sentence splitter*: It separates individual words of the Quranic text and then constructs sentences based on the punctuations. Sentence and token annotations are obtained as output from this phase.
2. *POS tagging:* This phase assigns parts of speech to the individual token strings. Stanford parser was used to POS tag the token annotations
3. *Extract Dependencies:* This phase assigns dependency relationship between the tokens, Stanford parser was used for this purposeas shown in Fig. 4

   Stanford dependencies like pobj,det,appos were extracted from a sample sentence from the Quran as follows:

   "In the name of God, the beneficent, the merciful".
4. *JAPE Transducer:* JAPE is a Java Annotation Patterns Engine. JAPE provides finite state transduction over annotations based on regular expressions.Lexico syntactic patterns are formulated as JAPE. The formulated patterns for the literary forms as discussed in section III.A is used to recognize the literary forms. The grammar consists of a set of phases, each of which consists of a set of pattern/actionrules.
5. *Literary Pattern database:* A collection of files is used to store the JAPE grammar corresponding to the lexico syntactic patterns for the various literary forms. The lexico syntactic patterns were formulated by domain expert after careful analysis of the grammatical structure of the English translations of the Quran.

© 2010, IJARCS All Rights Reserved

6. *Literary Annotations processing:* The annotations denoting the various literary forms need to be converted to an intermediate form before it can be converted into ontological classes or relationships.The intermediate forms are than converted into ontology triples in OWL format.
7. *Ontology Conversion:* The literary forms in the intermediate form need to be converted into a literature ontology.The OTCM (Ontology Term Coverage Metric) quantifies the degree up to which all key terms extracted by the keyphrase extractor are represented by the ontology, thus at each iteration the OTCM metric is recomputed. Further iterations are stopped when OTCM value converges to a fixed value.Ontology conversion seeks to maximize this metric.
8. *Extract Keyphrase:* Since all unique terms in the Quran need to be represented as concepts, all unique terms from the Quran verses is to be extracted in this phase.

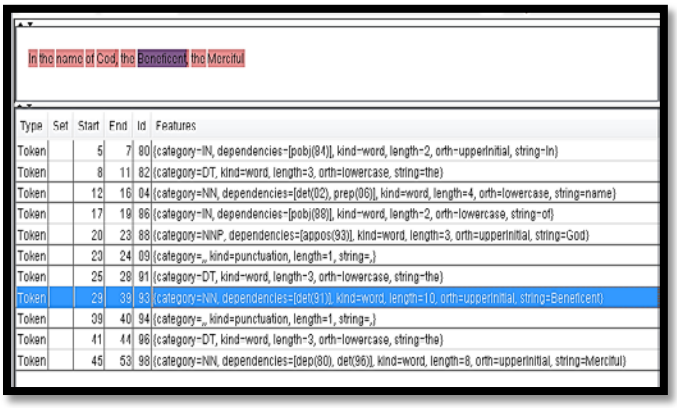

Figure 4.Extracted dependencies.

C. Evaluating the Ontology Term Coverage Metric (OTCM)

Once an ontology is generated there needs to be a metric to evaluate the completeness of the ontology. As no standard metrics are currently availaible, a new metric namely OTCM has been formulated to evaluate the completeness of the generated ontology. OTCM metric matches each term in the term list with an existing concept and relation in the generated ontology.The resultant matching score called as OTCM is defined as follows:

$$\frac{\sum_{t\in CS_g}\sum_{k\in RS_g}\sum_{t\in T_l} 0.8 * match(C_i, T_j) + 0.2 * match(R_k, T_j)}{|C+T+R|} \quad (1)$$

Where CS_g&$RS_g$are the concept set and relation set for ontology g having concept C_i , Relation R_k and T_j denotes the term found in the index term list.and the $|C+T+R|$ represents the total number of ontology concepts, Key Terms and Object Properties respectively.

Match function is defined as follows:

IF similarity(X_1,X_2)>ε
Match(X_1,X_2)=1
ELSE
Match(X_1,X_2)=0

Where ε is a small empirically derived threshold that distinguishes a match based on a Wordnet based similarity module that combines path similarity metrics and Wu-Palmer Similarity [18].The OTCM metric is calculated at the end of eachiteration to increase the completeness of the generated ontology.

D. Examples of Literary Lexico-Syntactic Literary Concept Extraction

Stanford parser was used to generate the POS tag and dependencies of each sentence in the Quran. JAPE grammars were used to formulate the lexico syntactic patterns for ontology concept, Metaphor, etc. The open source GATE tool was used to execute the JAPE code on the English Quranic corpus as shown in the following examples.

Ontology Concept Detection Pattern	
Formalization	$\forall_{i\in(NP\cap KP)}$i → Concept($i$)
Example Q 37.83-113	NP= {face, reward, peace,...} ⬇ Concept(face),Concept(reward), Concept(peace) …

Figure3. Concept Detection

As shown in Fig.5 the ontology concept corresponds to a key noun phrase occurring throughout the Quran where NP is noun phrase and KP is key phrase.

Similarly in Fig 6 instances of a concept can be recognized as Proper nouns (PN) like name of prophets are the candidate instances.

Ontology Instance Detection Pattern	
Formalization	$\forall_i \in (PN) \rightarrow$ Individual(i)
Example Q 38.45-47	PN={Abraham,Isaac,Jacob} Individual(Abraham),Individual(Isaac) …

Figure4. Instance detection

Metaphor Detection Pattern	
Formalization	similitude * as CONCEPT similitude of CONCEPT similitude of the CONCEPT ⬇ Metaphor(CONCEPT)
Example Q13.35	A similitude of the Garden ⬇ Metaphor(Garden) -

Figure5.Metaphor Detection

In Figure5"Similitude" keyword is used to detect the metaphors and parables in the Quran. The pre recognized concept becomes the subclass of the metaphor superclass.

VI. CONCLUSION AND FUTURE WORK

This paper delineates a semi-automatic methodology to generate a literature ontology recognizing the literary forms present in the EnglishTranslation of grouped Quranicverses.
An ontology comprising the various literary forms was generated based on the proposed methodology. A new metric namely OTCM was formulated to evaluate the completeness

© 2010, IJARCS All Rights Reserved

of the generated ontology at each phase. The final ontology when evaluated with OTCM resulted in a high OTCM score reflecting high concept coverage of the final ontology.

Thereexist more than a hundred literary forms for the English language alone out of which we have proposed methods to identify the most important ones like concepts ,metaphorsetc. Work is still in progress to formulate the lexico syntactic patterns for other important literary forms.

Higher precision and recall can be obtained if we can employ different translations of Quran which can be aligned leading to a low false positive rate for literary feature detection.

V. REFERENCES

[1] A. Zouaq and R. Nkambou, "Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project," IEEE Transactions on Knowledge and Data Engineering (10.1109/TKDE. 2009.25), 2009.

[2] P. Buitelaar, P. Cimiano, and B. Magnini, "Ontology learning from text: methods, evaluation and applications," Computational Linguistics, vol. 32, no. 4.

[3] A. Gómez-Pérez and D. Manzano-Macho, "A survey of ontology learning methods and techniques OntoWeb Consortium," OntoWeb Consortium, 2003.

[4] R. Navigli, P. Velardi, and A. Gangemi, "Ontology learning and its application to automated terminology translation," IEEE Intelligent Systems, vol. 18, no. 1, p. 22–31, 2003.

[5] P. Velardi, R. Navigli, A. Cucchiarelli, and F. Neri, "Evaluation of ontolearn, a methodology for automatic learning of ontologies," in Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, 2005.

[6] R. Engels and T. Lech, "Generating ontologies for the semantic web: Ontobuilder," in Towards the Semantic Web: Ontology-driven Knowledge Management, England: John Wiley & Sons, 2003.

[7] M. Shamsfard and A. Barforoush, "The state of the art in ontology learning: A framework for comparison," Knowledge Engineering Review, pp. 18-4, 2003.

[8] J. Dror, D. Shaharabani, R. Talmon, and S. Wintner, "Morphological Analysis of the Qur'an," Literary and linguistic computing, vol. 19, no. 4, p. 431, 2004.

[9] K. Dukes and N. Habash, "Morphological Annotation of Quranic Arabic," in Language Resources and Evaluation Conference (LREC). Valletta, Malta, 2010.

[10] K. Dukes and T. Buckwalter, "A Dependency Treebank of the Quran using Traditional Arabic Grammar," in 7th international conference on Informatics and Systems. Cairo, Egypt, 2010.

[11] K. Dukes, E. Atwell, and A. B. M. Sharaf, "Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank."

[12] M. Sanderson and B. Croft, "Deriving concept hierarchies from text," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, p. 206–213.

[13] B. J. Wielinga, A. T. Schreiber, J. Wielemaker, and J. A. C. Sandberg, "From thesaurus to ontology," in Proceedings of the 1st international conference on Knowledge capture, 2001, p. 201.

[14] H. A. Tzortzis, "An Introduction to the Literary & Linguistic Excellence of the Qur'an."

[15] M. A. A. Haleem, Understanding the Qur'an: themes and style. I.B. Tauris, 1999.

[16] "The Qur'an Oaths : Farahi's Inte rpretation," in Islamic Studies, 1990, Spring Issue., .

[17] P. Cimiano, "Ontology Learning and Population from Text," PhD thesis at the Universität Karlsruhe (TH), FakultätfürWirtschaftswissenschaften, 2006.

[18] E. Blanchard, M. Harzallah, H. Briand, and P. Kuntz, "A typology of ontology-based semantic measures," EMOI-INTEROP'05, Proc. Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability.

© 2010, IJARCS All Rights Reserved