A Pilot PropBank Annotation for Quranic Arabic

Wajdi Zaghouani
University of Pennsylvania
Philadelphia, PA USA
wajdiz@ldc.upenn.edu

Abdelati Hawwari and **Mona Diab**
Center for Computational Learning Systems
Columbia University, NYC, USA
{ah3019,mdiab}@ccls.columbia.edu

Abstract

The Quran is a significant religious text written in a unique literary style, close to very poetic language in nature. Accordingly it is significantly richer and more complex than the newswire style used in the previously released Arabic PropBank (Zaghouani et al., 2010; Diab et al., 2008). We present preliminary work on the creation of a unique Arabic proposition repository for Quranic Arabic. We annotate the semantic roles for the 50 most frequent verbs in the Quranic Arabic Dependency Treebank (QATB) (Dukes and Buckwalter 2010). The Quranic Arabic PropBank (QAPB) will be a unique new resource of its kind for the Arabic NLP research community as it will allow for interesting insights into the semantic use of classical Arabic, poetic literary Arabic, as well as significant religious texts. Moreover, on a pragmatic level QAPB will add approximately 810 new verbs to the existing Arabic PropBank (APB). In this pilot experiment, we leverage our knowledge and experience from our involvement in the APB project. All the QAPB annotations will be made freely available for research purposes.

1 Introduction

Explicit characterization of the relation between verbs and their arguments has become an important issue in sentence processing and natural language understanding. Automatic Semantic role labeling [SRL] has become the correlate of this characterization in natural language processing literature (Gildea and Jurafsky 2002). In SRL, the system automatically identifies predicates and their arguments and tags the identified arguments with meaningful semantic information. SRL has been successfully used in machine translation, summarization and information extraction. In order to build robust SRL systems there is a need for significant resources the most important of which are semantically annotated resources such as proposition banks. Several such resources exist now for different languages including FrameNet (Baker et al., 1998), VerbNet (Kipper et al. 2000) and PropBank (Palmer et al., 2005). These resources have marked a surge in efficient approaches to automatic SRL of the English language. Apart from English, there exist various PropBank projects in Chinese (Xue et al., 2009), Korean (Palmer et al. 2006) and Hindi (Ashwini et al., 2011). These resources exist on a large scale spearheading the SRL research in the associated languages (Carreras and Marquez, 2005), Surdeanu et al. (2008). However, resources created for Arabic are significantly more modest. The only Arabic Propank [APB] project (Zaghouani et al., 2010; Diab et al., 2008) based on the phrase structure syntactic Arabic Treebank (Maamouri et al. 2010) comprises a little over 4.5K verbs of newswire modern standard Arabic. Apart from the modesty in size, the Arabic language genre used in the APB does not represent the full scope of the Arabic language. The Arabic culture has a long history of literary writing and a rich linguistic heritage in classical Arabic. In fact all historical religious non-religious texts are written in Classical Arabic. The ultimate source on classical Arabic language is the Quran. It is considered the Arabic language reference point for all learners of Arabic in the Arab and Muslim world. Hence understanding the semantic nuances of Quranic Arabic is of significant impact and value to a large population. This is apart from its significant difference from the newswire genre, being closer to poetic language and more creative linguistic ex-

Workshop on Computational Linguistics for Literature, pages 78–83,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

pression. Accordingly, in this paper, we present a pilot annotation project on the creation a Quranic Arabic PropBank (QAPB) on layered above the Quranic Arabic Dependency Treebank (QATB) (Dukes and Buckwalter 2010).

2 The PropBank model

The PropBank model is a collection of annotated propositions where each verb predicate is annotated with its semantic roles. An existing syntactic treebank is typically a prerequisite for this shallow semantic layer. For example consider the following English sentence: 'John likes apples', the predicate is 'likes' and the first argument, the subject, is 'John', and the second argument, the object, is 'apples'. 'John' would be semantically annotated as the *agent* and 'apples' would be the *theme*. According to PropBank, 'John' is labeled ARG0 and 'apples' is labeled ARG1. Crucially, regardless of the adopted semantic annotation formalism (PropBank, FrameNet, etc), the labels do not vary in different syntactic constructions, which is why proposition annotation is different from Treebank annotation. For instance, if the example above was in the passive voice, 'Apples are liked by John', John is still the agent ARG0, and Apples are still the theme ARG1.

3 Motivation and Background

The main goal behind this project is to extend coverage of the existing Arabic PropBank (APB) to more verbs and genres (Zaghouani et al. 2010; Diab et al. 2008). APB is limited to the newswire domain in modern standard Arabic (MSA). It significantly lags behind the English PropBank (EPB) in size. EPB consists of 5413 verbs corresponding to 7268 different verb senses, the APB only covers 2127 verb types corresponding to 2657 different verb senses. According to El-Dahdah (2008) Arabic Dictionary, there are more than 16,000 verbs in the Arabic language. The Quran corpus comprises a total of 1466 verb types including 810 not present in APB. Adding the 810 verbs to the APB is clearly a significant boost to the size of the APB (38% amounting to 2937 verb types).

In the current paper however we address the annotation of the Quran as a stand alone resource while leveraging our experience in the APB annotation process. The Quran consists of 1466 verb types corresponding to 19,356 verb token instances. The language of the Quran is Classical Arabic (CA) of 77,430 words, sequenced in chapters and verses, dating back to over 1431 years. It is considered a reference text on both religious as well as linguistic matters. The language is fully specified with vocalic and pronunciation markers to ensure faithful oration. The language is poetic and literary in many instances with subtle allusions (Zahri 1990). It is the source of many other religious and heritage writings and a book of great importance to muslims worldwide, including non speakers of Arabic.

Dukes and Buckwalter (2010) started the Quranic Arabic Corpus, an annotated linguistic resource which marks the Arabic grammar, syntax and morphology for each word. The QATB provides two levels of analysis: morphological annotation and syntactic representation. The syntax of traditional Arabic grammar is represented in the Quranic Treebank using hybrid dependency graphs as shown in Figure 1.[1] To the best of our knowledge, this is the first PropBank annotation of a religious and literary style text.

The new verbs added from the Quran are also common verbs widely used today in MSA but the Quranic context adds more possible senses to these verbs. Having a QAPB allows for a more semantic level of analysis to the Quran. Currently the Quranic Corpus Portal[2] comprises morphological annotations, syntactic treebanks, and a semantic ontology. Adding the QAPB will render it a unique source for Arabic language scholars worldwide (more than 50,000 unique visitors per day).

Linguistic studies of the Quranic verbs such as verbal alternations, verb valency, polysemy and verbal ambiguity are one of the possible research directions that could be studied with this new resource. On the other hand, the Arabic NLP research community will benefit from the increased coverage of the APB verbs, and the new domain covered (religious) and the new writing style (Quranic Arabic). Furthermore, Quranic citations are commonly used today in MSA written texts (books, newspapers, etc.), as well as Arabic social media intertwined with dialectal writings. This

[1] This display is different from the other existing Arabic Treebank, the Prague Arabic Dependency Treebank (PADT) (Smrž et al., 2008).

[2] http://corpus.quran.com/

makes the annotation of a Quranic style a rare and relevant resource for the building of Arabic NLP applications.

4 Methodology

We leverage the approach used with the previous APB (Zaghouani et al. 2010; Diab et al. 2008). We pay special attention to the polysemic nature of predicates used in Quranic Arabic. An Arabic root meaning tool is used as a reference to help in identifying different senses of the verb. More effort is dedicated to revision of the final product since unlike the APB, the QAPB is based on a dependency Treebank (QATB) not a phrase structure Treebank.[3]

For this pilot annotation experiment, we only annotate the 50 most frequent verbs in the corpus corresponding to 7227 verbal occurrences in the corpus out of 19,356 total verbal instances. In the future plans, the corpus will cover eventually all the 1466 verbs in the whole Quranic corpus. Ultimately, it is our plan to perform a merging between the new frame files of the QAPB and the existing 1955 Frame files of the Arabic PropBank

4.1 The annotation process

The PropBank annotation process is divided into two steps: a. creation of the frame files for verbs occurring in the data, and b. annotation of the verbal instances with the frame file ids. During the creation of the Frame Files, the usages of the verbs in the data are examined by linguists (henceforth, “framers”). During the frameset creation process, verbs that share similar semantic and syntactic characteristics are usually framed similarly). Once a predicate (in this case a verb) is chosen, framerlook at an average sample size of 60-70 instances per predicate found in the Quranic corpus in order to get an idea of its syntactic behavior. Based on these observations and their linguistic knowledge and native-speaker intuition, the framers create a Frame File for each verb containing one or more framesets, which correspond to coarse-grained senses of the predicate lemma. Each frameset specifies the PropBank core labels (i.e., ARG0, ARG1,…ARG4) corresponding to the argument structure of the verb. Additionally, illustrative examples are included for each frameset, which will later be referenced by the annotators. Note that in addition to these core, numbered roles, PropBank also includes annotations of a variety of modifier roles, prefixed by ARGM labels from a list of 15 arguments (ARGM-ADV, ARGM-BNF, ARGM-CAU,ARGM-CND, ARGM-DIR, ARGM-DIS, ARGM-EXT, ARGM-LOC, ARGM-MNR, ARGM-NEG, ARGM-PRD, ARGM-PRP, ARGM-REC, ARGM-TMP, ARGM-PRD). Unlike the APB frame files creation, where no specific Arabic reference is used, for this project, an Arabic root meaning reference tool developed by Swalha (2011) is used by the framers to ensure that all possible meanings of the verbs in the corpus are covered and all various senses are taken into account. The Arabic root-meaning search tool is freely available online.[4] The search is done by root, the tool displays all possible meanings separated by a comma with citation examples from many sources including the Quran. Once the Frame files are created, the data that have the identified predicate occurrences are passed on to the annotators for a double-blind annotation process using the previously created framesets. Each PropBank entry represents a particular instance of a verb in a particular sentence in the Treebank and the mapping of numbered roles to precise meanings is given on a verb-by-verb basis in a set of frames files during the annotation procedure. To ensure consistency, the data is double annotated and finally adjudicated by a third annotator. The adjudicator resolves differences between the two annotations if present to produce the gold annotation. A sample Frameset and a related annotation example from the QAPB are shown in Table 1. During the annotation process, the data is organized by verb such that each verb with all its instances is annotated at once. In doing so, we firstly ensure that the framesets of similar verbs, and in turn, the annotation of the verbs, will both be consistent across the data. Secondly, by tackling annotation on a verb-by-verb basis, the annotators are able to concentrate on a single verb at a time, making the process easier and faster for the annotators.

[3] The Propbank style of annotation are already used with other languages on top of dependency Treebank structures such as the Hindi Treebank project (Ashwini et al., 2011).

[4] Available at :<http://www.comp.leeds.ac.uk/cgi-bin/scmss/arabic_roots.py>

FrameSet Example	Annotation Example
Predicate: wajadaوَجِدَ Roleset id: f1, to find **Arg0: the finder** **Arg1: thing found**	**Rel:** wajada, **وَجِدَ** **Arg0: -NONE- *** **Gloss: You** **Arg1: هُ** **Gloss: it** **ArgM-LOC: عِنْدَ اللَّهِ** **Gloss: with Allah** **Example in Arabic:** **وَمَا تُقَدِّمُوا لِأَنْفُسِكُمْ مِنْ خَيْرٍ تَجِدُوهُ عِنْدَ اللَّهِ** **Gloss: and whatever good you put forward for yourselves - you will find it with Allah**

Table 1. The frameset / Annotation of wajada

4.2 Tools

Frameset files are created in an XML format. We use tools used in the APB project. The Frame File editing is performed by the Cornerstone tool (Choi et al., 2010a), which is a PropBank frameset editor that allows creation and editing of PropBank framesets without requiring any prior knowledge of XML. Moreover, we use Jubilee[5] as the annotation tool (Choi et al., 20010b). Jubilee is a recent annotation tool which improves the annotation process of the APB by displaying several types of relevant syntactic and semantic information simultaneously. Having everything displayed helps the annotator quickly absorb and apply the necessary syntactic and semantic information pertinent to each predicate for consistent and efficient annotation. Both tools are currently being modified in order to handle the Dependency TreeBank structure, originally the tool was designed specifically to handle phrase structure Tree format. Moreover, since the file formats and the tree formats in the dependency Treebank are different from the previous APB effort, a revision in the Quranic Treebank output had to be done. This involves mainly a change in the annotated data format in order to add the role labels in the annotation file. For the moment, all of the 50 XML Frame files have been created and some manual annotation is performed to illustrate the feasibility of the experiment.

4.3 Impact of the dependency structure Treebank

Having The Quran corpus annotated using a dependency structure Treebank has some advantages. First, semantic arguments can be marked explicitly on the syntactic trees (such as the Arg0 Pron. In Figure 1), so annotations of the predicate argument structure can be more consistent with the dependency structure as shown in Figure 1.

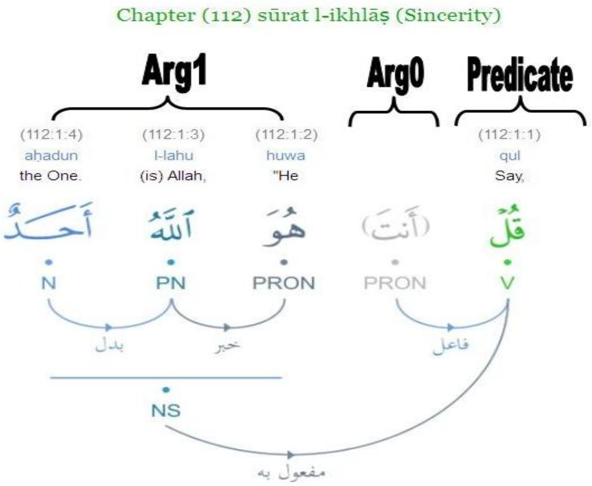

Figure 1. Semantic role labels to the QATB

Secondly, the Quranic Arabic Dependency Treebank (QATB) provides a rich set of dependency relations that capture the syntactic-semantic information. This facilitates possible mappings between syntactic dependents and semantic arguments. A successful mapping would reduce the annotation effort.

It is worth noting the APB comprises 1955 verbal predicates corresponding to 2446 framesets with an ambiguity ratio of 1.25. This is in contrast to the QAPB where we found that the 50 verbal predicate types we annotated corresponded to 71 framesets thereby an ambiguity ratio of 1.42. Hence these results suggest that the QAPB is more ambiguous than the newswire genre annotated in the APB. By way of contrast, the EPB comprises 6089 verbal predicates corresponding to 7268 framesets with an ambiguity ratio of 1.19.

21 verb types of the 50 verbs we annotated are present in both corpora corresponding to 31 framesets in QAPB (a 1.47 ambiguity ratio) and 25 framesets in APB (1.19 ambiguity ratio). The total verbal instances in the QAPB is 2974. 29 verb

[5] Cornerstone and Jubilee are available as Open Source tools on Google code.

types with their corresponding 40 framesets occur only in the QAPB (58% of the list of 50 verbs). This translated to a 1.38 ambiguity ratio.

In the common 21 verb types shared between APB and QAPB corpora we note that 12 predicates share the same exact frame sets indicating no change in meaning between the use of the predicates in the Quran and MSA. However, 9 of the verbal predicates have more framesets in QAPB than APB. None of the verbal predicates have more framesets in APB than QAPB. Below is an example of a verbal predicate with two different framesets.

FrameSet Example	**Annotation Example**
Predicate: >anozal أَنْزَلَ Roleset id: f1, to reveal Arg0: revealer Arg1: thing revealed Arg2: start point Arg3: end point, recipient	Rel: >anozal Arg0: نَا Gloss: we Arg1: آيَاتٍ بَيِّنَاتٍ Gloss: *clear verses* Arg3: إِلَيْكَ Gloss: to you Example in Arabic: وَلَقَدْ أَنْزَلْنَا إِلَيْكَ آيَاتٍ بَيِّنَاتٍ *We have certainly revealed to you verses [which are] clear proofs*

Table 2. The frameset / Annotation of >anozal (QAPB)

FrameSet Example	**Annotation Example**
Predicate: >anozal أَنْزَلَ **Roleset id*: f1, to release*** Arg0: *agent releasing* Arg1:*thing released*	Rel: >anozal Arg0: زياد Gloss: Zyad **Arg1:NONE-*** Gloss: He ARGM-TMP: منتصف الثمانينات Gloss: the mid-eighties Example in Arabic: عودة الطلب على أغاني شريط أنا مش كافر الذي أنزله زياد منتصف الثمانينات The songs of the Album I am not a disbeliever released by Ziad during the eighties are popular again.

Table 3. The frameset / Annotation of >anozal (APB)

The two frames of verb '' >anozal " can clarify the meaning differences between MSA and QA as used in the Quran. Although both APB and QAPB have this verb, they have different senses leading to different semantic frames. In the QAPB the sense of revealed is only associated with religious texts, while in MSA it has the senses of released or dropped.

5 Conclusion

We have presented a pilot Quranic Arabic PropBank experiment with the creation of frame files for 50 verb types. At this point, our initial study confirms that building a lexicon and tagging the Arabic Quranic Corpus with verbal sense and semantic information following the PropBank model is feasible. In general, the peculiarities of the Quranic Arabic language did not seem to cause problems for the PropBank annotation model. We plan to start the effective annotation of the resource in order to finalize the creation of a QAPB that covers all 1466 verbal predicates. Once released, the data will be freely available for research purpose.

References

Vaidya Ashwini, Jinho Choi, Martha Palmer, and Bhuvana Narasimhan. 2011. Analysis of the Hindi Proposition Bank using Dependency Structure. In *Proceedings* of the fifth Linguistic Annotation Workshop. ACL 2011, pages 21-29.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings* of COLING-ACL '98, the University of Montreal, pages 86–90.

Xavier Carreras and Lluıs Marquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings* of the Ninth (CoNLL-2005), pages 152–164.

Jinho Choi, Claire Bonial, and Martha Palmer.2010a. PropBank Instance Annotation Guidelines Using a Dedicated Editor, Cornerstone.In *Proceedings* of the (LREC'10), pages 3650-3653.

Jinho Choi, Claire Bonial, and Martha Palmer.2010b. PropBank Instance Annotation Guidelines Using a Dedicated Editor, Jubilee.In *Proceedings* of the (LREC'10), pages 1871-1875.

Mona Diab, Aous Mansouri, Martha Palmer, Olga Babko-Malaya,Wajdi Zaghouani, Ann Bies, and Mohammed Maamouri. 2008. A Pilot Arabic PropBank. In *Proceedings* of the (LREC'08), pages 3467-3472.

Kais Dukes and Tim Buckwalter. 2010. A Dependency Treebank of the Quran using Traditional Arabic Grammar. In *Proceedings* of the 7th International Conference on Informatics and Systems (INFOS).

Antoine El-Dahdah. 2008. *A Dictionary of Arabic Verb Conjugation*. Librairie du Liban, Beirut, Lebanon.

Daniel Gildea, and Daniel Jurafsky. 2002. Automatic-Labeling of Semantic Roles. *Computational Linguistics* 28:3, 245-288

Karin Kipper, HoaTrang Dang, and Martha Palmer. 2000. Class-Based Construction of a Verb Lexicon. In *Proceedings* of the AAAI-2000 Seventeenth National Conference on Artificial Intelligence, pages 691-696.

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma-Gaddeche, Wigdan Mekki, Sondos Krouna, Basma-Bouziri, and Wajdi Zaghouani. 2011. Arabic Treebank: Part 2 v 3.1. LDC Catalog No.:LDC2011T09

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31:1

Martha Palmer, Shijong Ryu, Jinyoung Choi, Sinwon Yoon, and Yeongmi Jeon. 2006. LDC Catalog LDC2006T03.

Otakar Smrž, Viktor Bielický, IvetaKouřilová, Jakub Kráčmar, Jan Hajič and Petr Zemánek. 2008. Prague Arabic Dependency Treebank: A Word on the Million Words.In *Proceedings* of the Workshop on Arabic and Local Languages (LREC 2008).

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluıs Marquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing on syntactic and semantic dependencies. In *Proceedings* of CoNLL'08, pages 159–177.

Majdi Swalha. 2011. *Open-source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora*. PhD thesis, Leeds University.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In Dekang Lin and Dekai Wu, editors, *Proceedings* of EMNLP 2004, pages 88–94.

Wajdi Zaghouani , Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. 2010. The revised Arabic PropBank. In *Proceedings* of the Fourth Linguistic Annotation Workshop (LAW IV '10), ACL, pages 222-226.

Maysoon Zahri. 1990. *Metaphor and translation*. PhD thesis, University of Salford.