AN ONTOLOGICAL MODEL FOR REPRESENTING SEMANTIC LEXICONS: AN APPLICATION ON TIME NOUNS IN THE HOLY QURAN

Maha Al-Yahya, Hend Al-Khalifa
Information Technology Department, King Saud University, Riyadh, Saudi Arabia

Alia Bahanshal, Iman Al-Odah
King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia

and Nawal Al-Helwah
Arabic Language Department, Princess Norah University, Riyadh, Saudi Arabia

*Corresponding Author:
E-mail: malyahya@ksu.edu.sa

Paper Received September 30, 2010; Paper Revised December 29, 2010; Paper Accepted December 29, 2010
ABSTRACT

Although Arabic is the language of over two hundred million speakers, little has been achieved in regards to computational Arabic resources, especially lexicons. Most of what has been developed was originally tailored for Roman languages, and is not necessarily satisfactory for the Arabic community. In this research, we propose a computational model for representing Arabic lexicons using ontologies. Ontologies are knowledge representation structures which form the central building block of the Semantic Web. The model is based on the field theory of semantics from the linguistics domain, and the data which drives the design of the model is obtained from the most accurate text that presents superiority and perfection of the Arabic language, the Holy Quran. Creating such lexicons will be invaluable in a number of Arabic applications. This paper presents the design and implementation of the proposed ontological model. Results of its application on “Time nouns” vocabulary of the Holy Quran are presented.

Key words: Holy Quran, Arabic language, ontology, lexicon, semantic web
AN ONTOLOGICAL MODEL FOR REPRESENTING SEMANTIC LEXICONS: AN APPLICATION ON TIME NOUNS IN THE HOLY QURAN

1. INTRODUCTION

Although Arabic is the language of over two hundred million speakers, little has been achieved in regards to computational Arabic resources, especially lexicons. Most of what has been developed was originally tailored for Roman languages, so is not necessarily satisfactory for the Arabic community. A computational lexicon is the central building block from which numerous natural language applications can be developed [1]. We aim to fill this gap by proposing a new ontological model which provides the foundation for building a computational lexicon for the Arabic language based on semantic field theory (see [2]).

A lexicon is the vocabulary of a language, along with a lexical entry which gives more detail about the word [3]. The content of each lexical entry depends on the purpose of the lexicon. A lexical entry may include various word properties such as semantics (meanings and relationships with other words), phonetics (sound/pronunciations), morphology, and syntax (grammatical behavior). The lexical entries which we address in our model are those related to the semantics of the word.

An important element in our model is the ontology. Ontologies are knowledge representation structures capable of capturing domain knowledge in a formal way. They are well known in the Artificial Intelligence community and recently adopted by the Semantic Web community. Ontologies provide structured descriptions of domain facts by representing a consensual agreement on the concepts and relations in that domain. An ontology is similar to a dictionary or glossary, but with greater detail and structure that enables computers to process its content by formalizing concepts and relations. Ontology-based linguistic resources are valuable for any natural language processing application, especially Semantic Web applications. The ontology we present in this paper is applied on nouns from the “Time” semantic field. However, it can be extended to include nouns from other semantic fields.

Our proposed model is classified as a data-driven model, in which all Time nouns from the Holy Quran are used to derive the resulting ontological structure. Although the vocabulary we model is limited to that appearing in the Holy Quran, the meaning of the vocabulary is derived from authoritative Arabic lexicons known in the field (see [4]). To enable interoperability, sharing and reuse of this valuable resource, the developed ontology is implemented using a recent W3C standard for representing ontologies, Web Ontology Language (OWL) [5].

This paper is organized as follows: the next section presents the theoretical background of our approach, the semantic field theory; Section 3 provides a review of relevant literature on computational lexicons; Section 4 presents the design and evaluation of our ontological model; Section 5 presents a discussion of the results and findings; and finally, Section 6 concludes the paper with a summary of the work done in this project and possible directions for future work.

2. SEMANTIC FIELD THEORY

Within the field of linguistics, the definition of meaning is considered one of the most ambiguous and most controversial. In general, however, there are two main schools of thought when defining meaning: the analytical (referential) approach, and the operational approach. The analytical approach defines meaning by analyzing componential features of words, and the operational approach studies the words in usage [6].

The semantic field theory (conceptual spheres), which forms the theoretical foundation of our computational model, follows an analytical approach. It was first introduced by Professor Jost Trier (see [2]). According to this theory, the meaning of a word is considered within a given view of the world. It is dependent on its relation to other words in the same semantic field (conceptual area) [7]. It assumes that the lexicon is structured into semantic fields according to a set of primitive features. Word meaning is established by the position within the field, and the relationship it has with other words in its field. Consequently, words do not exist in isolation, and there is always a kind of relation, a “sense relation”, as Buren [8] states. Among these relations are synonymy, defined by Lehrer [2] as bilateral implication: two words are synonymous if they can be mutually defined. Another relation is antonymy, and there are different types of this relation between words. For example, the type of relationship between “man” and “woman” is antonymous but different from that of “hot” and “cold”. The reason is that in the former, the denial of one means the other is true, whereas in the latter if something is not hot it does not mean it is cold. There are other levels or grades of meaning between the two extremes. Buren [8] calls the first “contradictories” and the second “contraries”. According to Lehrer [2] contraries are called gradable antonyms, which can be measured on a scale, and occur at the end of the scale. In addition, modifiers may be used with these scales such as “very, so, quite” to locate a precise point on the scale. Since contradictory antonyms are not gradable, modifiers cannot be used. Other semantic relations exist, such as incompatibility and converseness [2].
Central to semantic field theory is componential analysis (semantic analysis) [9]. Componential analysis is an approach which was originally used in the 1950s and 1960s [10,11] describing kinship terms in anthropology. Later, the approach was used in the field of semantics to study the distinction of meaning. Componential Analysis is considered a device for semantic classification [12], which is the “cornerstone of any study in semantics” [6].

Using componential analysis, a word meaning can be defined in terms of a number of specific atomic components and decompositions, sometimes called primitives or universals, which represent the distinctive features of a given word [13,14]. Words grouped into a semantic field share a common aspect of meaning. However, no two words share every part of meaning. Each word has a specific part (component) of meaning which is exclusive for that word and distinguishes it from the rest of the words in the field. This is called a distinctive feature. Individual components form the basis for structuring a specific semantic field. The inverse is also true, as [2] suggests that the organization of words in a semantic field can serve as a basis for determining components.

Using this approach to meaning enables us to differentiate between different words within the same semantic field (closely related words), especially for abstract terms which are difficult to distinguish semantically [15]. The following example illustrates the componential analysis of the words “man”, “woman”, “boy”, and “girl”. These words all belong to the same semantic field of “human”, and they are defined in terms of two semantic dimensions: “adulthood” and “gender”. The meaning of the individual words can be expressed as a combination of these features [16].

- Man= Human + Adult + Male
- Woman= Human + Adult + Female
- Boy= Human + Child + Male
- Girl= Human + Child + Female

These formulae are called componential definitions of the semantic units, and can be regarded as formalized dictionary definitions [16]. Componential analysis provides an economic way of representing word meaning compared to traditional dictionary or lexicon entries. It provides a more precise description through components. Therefore, semantic field theory and componential analysis offer a systematic framework for analyzing the semantics of language vocabulary. Componential analysis is invaluable for language learning and vocabulary acquisition, literacy writing, lexical change, and classification, among other useful applications.

3. RELATED WORK

Relevant work in the area of computational lexicons can be classified into three major types based on the structural model adopted for representing word semantics. These models include semantic network models, frame-based models, and ontology-based models.

3.1. Semantic Network Models

In semantic network models, lexical units are represented in a network structure and a word's meaning is defined by the number and type of connections it has with other words in the network. These connections represent different lexical relations such as hyponymy, synonymy, antonymy, etc. WordNet [17] is probably one of the most common and widely used lexicons based on a semantic network. It is one of the earliest, developed since the early 80’s. WordNet was built by lexicographers on the basis of analysis of language. The design is based on psycholinguistic and computational theories of human lexical memory. WordNet uses a semantic network structure representing words and concepts as an interconnected system consistent with the way humans organize their mental lexicon. The main taxonomic structure consists of a hierarchy of hyponyms. The synset is the set of synonyms, and plays a central role in presenting a lexical concept in WordNet; it acts as the root for other semantic relations. WordNet synsets are divided into nouns, verbs, adjectives, and adverbs. There are other semantic relations represented in WordNet. These include antonymy, hyperonymy/hyponymy, meronymy/holonymy (member, substance, part), entailment, cause, attribute, and similarity.

Following on the success of WordNet, word nets for other languages have been developed, such as EuroWordNet [18], which contains networks for Dutch, Italian, Spanish, English, Czech, Estonian, French, and German.

3.2. Frame-Based Models

Frame-based models are based on Frame Semantics formalisms [19]. An application of this model is the FrameNet project [20]. FrameNet provides a computational representation of semantic and syntactic combinatory possibilities of each sense of a word. A similar project for verbs is VerbNet [21,22]. VerbNet is an online verb lexicon for the English language. It is based on a hierarchical structure of verb classes. Verbs in the lexicon are linked to other lexical resources such as WordNet and FrameNet. Each class is described using thematic roles, restrictions on the arguments, and frames consisting of a syntactic description and semantic predicates with a
temporal function. Each frame is associated with explicit semantic information, expressed as a conjunction of Boolean semantic predicates such as motion, contact, or cause. Since its release, it has been used in a number of NLP applications for characterizing verbs and verb classes.

3.3. Ontology-Based Models

Ontology-based models use an ontology formalism for representing lexical entries. With recent advancements in Semantic Web technologies, ontology-based lexicons are becoming widespread. A rising trend is linking computational lexicons to upper (general) ontologies such as SUMO (Suggested Upper Merged Ontology) [23] and CYC [24], and linguistic foundational ontologies such as DOLCE [25].

There are a number of projects where an ontological structure has been adopted in lexicon design. For example, WordNet has been restructured according to the principles of formal ontology in the OntoWordNet project [26] and has been represented using the W3C standard Web Ontology language (OWL). WordNet has also been linked to SUMO, the suggested upper merged ontology1. In addition, FrameNet has been linked to the Suggested Upper Merged Ontology (SUMO) [27], and has been represented using OWL [28].

Another computational lexicon development project strongly influenced by Semantic Web technologies is the multilingual MILE project [29]. The design of the lexicon is based on defining lexical classes for creating objects to be used in building MILE conformant lexical entries. Lexical objects include semantic and syntactic features, semantic relations, syntactic constructions, predicate and arguments, etc. Lexical classes are organized in a hierarchy and defined using RDF schema.

The Lexical Markup Framework (LMF)2 is a standardized metamodel for representing computational lexicons. It provides the basic vocabulary required for generating a computational lexicon. It consists of three basic packages: the core package which describes the basic entries in a lexicon, the syntactic package which describe the morphology model and syntactic constructions, and the semantic package which provides vocabulary to describe one sense and its relation to other senses in the same language in a way similar to the synset notion of WordNet.

LexOnto [30] is a lexicon ontology developed for the lexicon engineer to map information in domain ontologies to natural language lexical frames (sub-categorization frames) which can aid in NLP applications. It aims to associate lexical structures with ontological classes and properties found in domain ontologies. In other words, it translates an ontological structure into a lexical frame for NLP applications. A crucial difference between our model and LexOnto is that LexOnto is not exactly a lexicon; it maps domain knowledge represented in ontologies into linguistic structures which can be processed by NLP applications. In addition, LexOnto focuses on subcategorization frames which are considered as a syntactic element of a lexical entry. However, our model focuses more on the semantics.

Another related model built upon LexOnto is LexInfo [31]. LexInfo is an ongoing project which utilizes LingInfo [32] and LMF. LexInfo enables the representation of morphological relations between terms, and the syntactic decomposition of composite terms. It also maps subcategorization frames into ontological structures. The motivation for such a project is to associate linguistic information (part of speech, morphological decomposition, etc.) with ontology elements in an attempt to facilitate tasks such as ontology learning and natural language generation from ontologies.

Another related system is the SIMPLE lexicon [33], which describes lexical entries in terms of their semantic properties. The model is designed to facilitate cross-language linking. The basic unit of the SIMPLE lexicon model is called Semantic Units (SemUs). Each SemanticUnit has an associated semantic type from the SIMPLE ontology. The main entities in the SIMPLE ontology are SemanticUnits, Semantic Type, and Templates. SemanticUnits represent the primary means for encoding word senses. Each SemanticUnit is assigned a semantic type from the ontology, plus other information specified in the associated template, which contribute to the characterization of the word-sense (similar to features). Semantic types involve structured information represented as template. The semantic types themselves are organized into an ontology [34]. Templates facilitate the lexicographer’s task by providing a structured form to encode information about a certain lexical unit. The SIMPLE lexicon contains a set of 150 common templates.

GOLD (General Ontology for Linguistic Description) [35] is an ontology for descriptive linguistics. It describes the basic concepts and relations in the field of describing natural language. It was designed to solve the problem of linguistic data mark-up. The focus of GOLD is on form units, and not meaning or semantics. GOLD is an effort to connect this linguistic domain ontology to an upper ontology such as SUMO.

1 http://sigma.ontologyportal.org:4010/sigma/KBs.jsp
2 http://www.lexicalmarkupframework.org/
With regards to ontological work for the Arabic language, there have been very few efforts. However, there are some ongoing projects, including Arabic WordNet [36], work by [37], the Al-Khalil project [38], and the Quranic Ontology Corpus3.

Arabic WordNet (AWN) [36]. AWN is a free lexical resource for Modern Standard Arabic based on the widely used Princeton WordNet for English. AWN is based on the design and contents of WordNet to enable machine translation. Similar to WordNet, AWN has also been linked to the Suggested Upper Merged Ontology (SUMO) [39]. This is an ongoing project since 2005. It follows a development process similar to WordNet, and has a comparable semantic structure. It utilizes the SUMO ontology to link to Wordnets in other languages. The AWN browser4 provides a tree-based representation of the terms. It presents top level concepts of nouns, verbs, adjectives, and adverbs. Synsets map to a general SUMO term or a term that is directly equivalent to the given synset. AWN has been used in a number of applications such as passage retrieval for question answering [40].

Belkridem and El Sebai [37] describe an ontological representation for the Arabic Language. The design of the ontology is based on a taxonomic relation between the main morphological classes of the Arabic language. Verbs are classified according to derivation rules of the Arabic language. Although the research is relevant, only a theoretical model is described and no implementation is provided.

Another recently published work is Al-Khalil linguistic ontology [38]. Al-Khalil is an OWL ontology which is based on the GOLD linguistic ontology [35]. Its objective is to provide a reference for the description of Arabic linguistics focusing on Arabic traditional grammar. The ontology design follows the phonetic order as in the book Kitab AlAyn by Al-Khalil Ibn Ahmad AlFarahidi. Although the project is promising, there is no implementation available.

The Quranic Arabic Corpus ontology provides an ontological classification of the concepts found in the Holy Quran. Searching for the word “فجر” (dawn) in the ontology returns the subtree (concept/event/physical event/dawn), the verse in which it is cited, along with its translation, and a visual concept map. Although it provides higher level conceptualizations, it does not address the meaning of specific words in the Holy Quran.

From our review of related work, three models are comparable to our work: WordNet, LexInfo, and the SIMPLE lexicon. Although WordNet organizes word senses (synsets) into a hierarchical structure and expresses various relationships between concepts (e.g., synonymy, antonymy, etc.), it does not explicitly describe what a concept means at the level of features and primitive word components as in componential analysis. Our model shares with the SIMPLE lexicon the idea of defining semantic features for a word sense. However, the conception of SIMPLE features is different. A word sense has a certain type defined in a taxonomy. For each type, there is an associated template which contains attribute-value pairs for the word sense. The LexInfo model captures the meaning of linguistic constructions, whereas our model captures the meaning of individual words. The work we present here is not concerned with linguistic terminology; instead, it looks into deeper meaning based on the field theory of semantics. Compared to AWN, our work differs in the underlying theoretical framework. AWN is based on the notion of synsets, while our model is based on componential analysis of word senses and semantic field theory. Since AWN is based on a model used by WordNet, AWN entries can be mapped to entries in WordNets for other languages. Our model, however, only supports part of the Arabic language at the current stage.

4. AN ONTOLOGICAL MODEL FOR ARABIC LEXICONS

The approach we followed in the ontology development is based on the UPON (Unified Process for ONtology) ontological engineering approach [41]. We had an expert in Arabic linguistics involved throughout the development process. In the following sub-sections, we first describe the goal and scope of our ontology, then we describe the design of our ontology, and how word meaning and lexical relations can be realized from this design. Finally, we present a preliminary evaluation of the ontology. Ontology design was an iterative process, and before we reached the final design we experimented with different models [42,43].

4.1. Goal and Scope

The goal of the ontology is to provide a computational model capable of representing word meaning using semantic field theory and componential analysis. Representing meaning in such a way will enable the creation of useful NLP applications for Arabic, such as semantic analysis. Semantic analysis is concerned with identifying the componential formulae, and semantic relations which exist between words. Users of the ontology include linguists, language learners, and researchers in NLP and Semantic Web applications.

The ontology is limited to Time nouns which appear in the Holy Quran. There are two reasons for choosing this sample of Arabic words. The first is that the Holy Quran represents the purest and most authentic form of the

3 http://corpus.quran.com/ontology.jsp
4 http://www.globalwordnet.org/AWN/AWNBrowser.html
classical Arabic language. The second is that we already have readily available a detailed componential analysis of this vocabulary [4]. Moreover, we limited componential analysis to individual words, and did not cover the meaning of larger linguistic units such as sentences, text, or context.

The vocabulary contains a total of 59 words (see Table 1). The meaning of these words was derived from authoritative Arabic dictionaries and lexicons (details in [4]). We used 28 words as a basis for the model design; the remaining 31 will be used for validation of the resulting model. The specific aspects of meaning we focused on in this model are componential analysis of word features (atomic components of word meanings). Other word properties such as phonetics (sound/pronunciations), morphology, and (syntax grammatical behavior) are not included.

Table 1. Time Nouns from the Holy Quran

<table>
<thead>
<tr>
<th>Time Nouns</th>
<th>Arabic Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>أَنْ</td>
<td>الأَنْ</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
<tr>
<td>يَدْرَكُ</td>
<td>يَدْرَكُ</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
<tr>
<td>يَدْرَكُ</td>
<td>يَدْرَكُ</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
<tr>
<td>يَدْرَكُ</td>
<td>يَدْرَكُ</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
<tr>
<td>يَدْرَكُ</td>
<td>يَدْرَكُ</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
<tr>
<td>يَدْرَكُ</td>
<td>يَدْرَكُ</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
<tr>
<td>يَدْرَكُ</td>
<td>يَدْرَكُ</td>
</tr>
<tr>
<td>يَلْتَوُى</td>
<td>يَلْتَوُى</td>
</tr>
</tbody>
</table>

4.2. Conceptual Analysis

Although the componential formulae are available, the conceptual classification of the words is not. The first stage in building the model was to analyze the components in the formulae and determine the semantic dimensions. Once determined, we organized the words into a hierarchical classification with general concepts at the top, and specific at the bottom. Words in the hierarchy are associated with components via ontological relations. This classification structure of the ontology implies that the deeper you go into the hierarchy, the more arguments the componential formula will have, and therefore the meaning narrows. In contrast, words at higher levels have fewer arguments in their componential formula, and therefore the meaning broadens.

4.3. Classes

Once the conceptual classification was done, a generalized model based on this classification was developed and mapped to ontological structures. The ontology consists of 18 classes, seven of which are general for any semantic field, and eleven of which are specific for the “Time” semantic field. Figure 1 shows the specification of the model represented using UML (Unified Modeling Language) notations. Lines ending with a triangle indicate subsumption relationship, while lines ending with an arrow indicate an association between the two classes. Class names are described in English and Arabic in the figure; however, in the actual implementation, they are in Arabic. An annotation is provided with an English description. Below we provide a textual description of each of the classes in our ontology.

LinguisticConcept: a class which represents all terminology used in our ontology.

SemanticField: a class representing all existing semantic fields within a language. A SemanticField contains a collection of one or more SemanticDomain(s) and includes one or more SemanticDimension(s). An example member of this class is the “Time/Zaman” SemanticField.

SemanticDomain: a division within a semantic field which divides the semantic field into conceptual spheres. A SemanticField contains a collection of one or more SemanticUnit(s). An example member of this class is the “Day” SemanticDomain.

SemanticUnit: a class representing a single word. A SemanticUnit has a minimum of one Feature, and one or two DistinctiveFeature(s) which differentiates the meaning of the word from others within the semantic domain or language. A SemanticUnit can be part of another semantic unit. For example, “Ghadat” is part of “Bukrah”. Since OWL does not provide any primitive for part-whole relationship, we have defined the (isPartOf) relationship between semantic units as a transitive property. An example SemanticUnit is the word “Evening”.

Feature: a feature of a certain SemanticUnit, which can be shared with other SemanticUnit(s). A single Feature may be shared with more than one SemanticUnit. An example member of this class is the Feature “Dark” for the SemanticUnit “Evening”.

DistinctiveFeature: a feature of a certain semantic unit, which is unique, and can only be shared where word synonymy exists. A single SemanticUnit can have at most two DistinctiveFeature(s).
**SemanticDimension**: a class representing meaning dimensions which exist for **SemanticField(s)** of a language. In our model, we explicitly describe the **TimeDimension**. Other dimensions can be added, such as “Human” **SemanticDimension**. Semantic dimensions are conceptualizations of features which are logically grouped together.

**TimeDimension**: a class representing the **SemanticDimension(s)** of the Time **SemanticField**. The Time semantic field has 10 dimensions. These dimensions were defined based on the componential formulae obtained from the componential analysis of each Time noun which appears in the Holy Quran [4]. The analysis reveals that numerous features can be combined into one conceptual dimension, thus forming a semantic dimension. For example, the noun “winter” has a feature of “cold”, and the noun “summer” has the feature “hot”; “hot” and “cold” are associated with one semantic dimension “temperature”, in which each feature occupies a specific point along that dimension. A single semantic unit does not necessarily need to be associated with all semantic dimensions. **TimeDimension** includes the following subclasses:

- **Order**: a **TimeDimension** class representing the position of the Time noun within a specific semantic domain. This class has three individuals (beginning, middle, end).
- **Temperature**: a **TimeDimension** class representing the temperature associated with a Time noun. There are two individuals in this class (hot, cold).
- **Dynamism**: a **TimeDimension** class representing the dynamistic nature of the Time noun. This class has two individuals (static, dynamic).
- **Explicitness**: a **TimeDimension** class representing the specificity of the Time noun. It involves two individuals (vague, exact).
- **Distribution**: a **TimeDimension** class representing the nature of the Time noun distribution. This class has three individuals (discrete-discontinuous, periodic, continuous-sequential).
- **Embodiment**: a **TimeDimension** class representing type of embodiment of the Time noun. It involves two individuals (concrete, abstract).
- **Period**: a **TimeDimension** class representing the time period of the Time noun. It involves three individuals (past, present, future).
- **Limit**: a **TimeDimension** class representing the limit of the Time noun. It involves two individuals (finite, infinite).
- **Range**: a **TimeDimension** class representing the range of the Time noun. It involves four individuals (short, medium, extending, long).
- **Illumination**: a **TimeDimension** class representing illumination of the Time noun. It involves two individuals (light, dark).
4.4. Individuals and Properties

Individuals in the ontology are of three types: SemanticDimensions, SemanticUnits (Quran Time vocabulary), and Features. Words are represented as individuals since they must have features. Features are represented as individuals, since they describe the lexical unit and are not subject to further classification. Properties are relations which exist between instances in the ontology. Details of the properties are shown in Table 2. There are two types of properties in our ontology: object properties and datatype properties. Object properties relate individuals with each other, and datatype properties relate individuals to literals (string/numeric values). Properties related to SemanticDimension(s) have certain predefined values defined using a hasPolarityValue property. These values are defined in the ontology using the owl:oneof construct for defining an enumeration. For example, values for the hasOrder property are defined as follows:

```
<owl:Class rdf:ID="Order">
  <owl:oneOf rdf:parseType= "Collection">
    <owl:Thing rdf:about="#أول"/>
    <owl:Thing rdf:about="#وسط"/>
    <owl:Thing rdf:about="#آخر"/>
  </owl:oneOf>
</owl:Class>
```

To represent feature polarity among a specified SemanticDimension, we use the hasPolarity property on features. This property is linked to a numeric data value (integer) and may hold the integer values (+1) for positive poles, (-1) for negative poles, and (0) for neutral poles – neither positive nor negative. Neutral poles are applicable where the dimension is not binary and represents a scale.
Table 2. Properties in the Ontology

<table>
<thead>
<tr>
<th>Property</th>
<th>Domain</th>
<th>Range</th>
<th>Restrictions</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>hasDistinctiveFeature</td>
<td>SemanticUnit</td>
<td>DistinctiveFeature</td>
<td>minCardinality=1, maxCardinality=2</td>
<td></td>
</tr>
<tr>
<td>hasFeature</td>
<td>SemanticUnit</td>
<td>Feature</td>
<td>minCardinality=1</td>
<td></td>
</tr>
<tr>
<td>hasSemanticUnit</td>
<td>SemanticUnit</td>
<td>SemanticUnit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hasPart</td>
<td>SemanticUnit</td>
<td>SemanticUnit</td>
<td></td>
<td>Transitive</td>
</tr>
<tr>
<td>hasDomain</td>
<td>SemanticUnit</td>
<td>SemanticDomain</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasOrder</td>
<td>SemanticUnit</td>
<td>Order</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasIllumination</td>
<td>SemanticUnit</td>
<td>Illumination</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasExplicitness</td>
<td>SemanticUnit</td>
<td>Explicitness</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasPeriod</td>
<td>SemanticUnit</td>
<td>Period</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasDuration</td>
<td>SemanticUnit</td>
<td>Duration</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasRange</td>
<td>SemanticUnit</td>
<td>Range</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasEmbodiment</td>
<td>SemanticUnit</td>
<td>Embodiment</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasDistribution</td>
<td>SemanticUnit</td>
<td>Distribution</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasLimit</td>
<td>SemanticUnit</td>
<td>Limit</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasDynamism</td>
<td>SemanticUnit</td>
<td>Dynamism</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasStem</td>
<td>SemanticUnit</td>
<td>String</td>
<td></td>
<td>Functional</td>
</tr>
<tr>
<td>hasPolarity</td>
<td>Feature</td>
<td>Integer</td>
<td></td>
<td>Functional</td>
</tr>
</tbody>
</table>

Figure 2 shows a sample set of individuals in the ontology where ellipses represent individuals and rectangles represent data values. The diagram is read in the direction of the arrows. To simplify the diagram, not all properties are shown.
4.5. Ontology Testing and Evaluation

The ontology was implemented using OWL, and can be obtained from the project homepage\(^5\). Using Semantic Web standards such as OWL to implement the ontology enables the shared and open access to such a valuable resource. In addition, it paves the way to the development of powerful Semantic Web applications.

According to [44], there are four important characteristics when evaluating an ontology:

- **Syntactic quality**: that measures the quality of the ontology according to its formal style, the way it is written;
- **Semantic quality**: where the primary concern is the absence of contradictory concepts;
- **Pragmatic quality**: that refers to the ontology content and usefulness for users, irrespective of its syntax and semantics;
- **Social quality**: that reflects more general criteria, for instance the numbers of other ontologies that link to it, e.g., by defining their terms using its definitions, and the number of times it is accessed (when public) from within and outside of the community that manages it.

Syntactic quality and semantic quality are both verified using the development tool we used, the Protégé ontology development editor. Since our ontology is an experimental prototype, we are not concerned with the social quality, which is only relevant when the ontology is published in the open domain. The focus of our evaluation in this research is on pragmatic quality, and therefore the evaluation will include two stages:

1. Evaluate how well the ontology can present word meaning (formulae and relations) via semantic analysis.
2. Test our ontology on the remaining Time nouns, not used during the design, and test it on new words from a new semantic field, also from the Holy Quran, and observe to what extent the model can accommodate them.

4.5.1. Stage1: Semantic analysis

Semantic analysis is concerned with identifying the componential formulae, and semantic relations which exist between words. Based on the field theory of semantics, the meaning of a word is defined as a set of features that distinguish it from other words in the language. These features are represented as a formula, and can be extracted from our ontological model by following an algorithmic procedure. For example, determining the meaning of the word “Bukrah بَكْرَة” can be determined by following the steps below:

1. We start with an empty formula, \( (=\text{Bukrah}) \).
2. Then, we extract all individuals in which the word (بَكْرَة) (Bukrah) is related to via the property \(\text{isPartOf} \), including the word itself, and add them to a set \(S \). \(S = \{\text{اليوم}, \text{بَكْرَة}\}\)
3. From the set \(S \), we select the element which is an instance of \(\text{SemanticDomain} \). We retrieve the \(\text{SemanticField} \) to which the \(\text{SemanticDomain} \) is part of, and add it to the formula, and therefore get \(\text{Bukrah}=\text{Time} \). We also add the domain name to the formula.

\[
\begin{align*}
\text{بَكْرَة}=\text{زَمَن}+\text{نَهَار} \\
\text{Bukrah}=\text{Time}+\text{day}
\end{align*}
\]

4. For all \(\text{SemanticDimension(s)} \) within the \(\text{SemanticField} \), check the properties associated with the \(\text{SemanticDomain} \) and add the instances of these dimensions to the formula:

\[
\text{بَكْرَة}=\text{زَمَن}+\text{نَهَار}+\text{مَحْدُود}+\text{نَور} \\
\text{Bukrah}=\text{Time}+\text{day}+\text{exact}+\text{light}
\]

5. For other words in the set, we check remaining semantic dimensions, and we add the value of these dimensions to the list. In our case, the only word left in the set is (بَكْرَة), so we add the instances of its dimensions to the formula:

\[
\text{بَكْرَة}=\text{زَمَن}+\text{يَوْم}+\text{مَحْدُود}+\text{أَوْلَ} \\
\text{Bukrah}=\text{Time}+\text{day}+\text{exact}+\text{light}+\text{beginning}
\]

6. If the words in the list include words which are associated with the \(\text{hasFeature} \) property of a feature, we add the features to the formula. In this example, this does not apply.

7. Finally, we add \(\text{DistinctiveFeatures} \) of the target word to the formula. The final formula is:

\[
\text{بَكْرَة}=\text{زَمَن}+\text{يَوْم}+\text{مَحْدُود}+\text{أَوْل}+\text{الضَّحِي}
\]

---

\(^5\) https://sites.google.com/site/semquran/
The above algorithmic steps can easily be implemented in a computer application using the Jena\footnote{http://jena.sourceforge.net/} API, a Semantic Web framework for Java.

As an example of analyzing semantic relations, we consider the antonymy semantic relation and see how it can be derived from our ontological model. Consider that we would like to know if two words are antonyms. First, we need to find the meaning formulae for both words, then we compare the features associated with each formula. If there exists a feature along a common dimension which has a negative polarity (-1) in one formula and positive polarity (+1) in the other, then we can conclude that these two words are antonyms. For example, the words (غدات) and (روح)، have the following componential formulae:

\[
\text{غدات} = \text{زمن} + \text{نهار} + \text{محدد} + \text{أول} + \text{من الفجر إلى الضحاي}
\]

\[
\text{رواح} = \text{زمن} + \text{نهار} + \text{محدد} + \text{آخر} + \text{من زوال الشمس إلى الليل}
\]

Checking the elements in the formula, we can see that (أول) has the value (+1), and (آخر) has the value (-1) along the Order semantic dimension. We can therefore say that these two words are antonyms.

In addition, given a specific word (which exists in the ontology), the model can give suggestions for possible antonyms. First we need to find the meaning formula for the word (as described above), then we examine each feature in the formula:

- If it has a polarity value, then we check it:
  - If it is equal to (+1) or (-1), then the word resides on a pole of the semantic dimension. We retrieve all words which have features associated with an opposite polarity. These words are possible antonyms for the target word.
  - If it does not have a polarity value, then antonyms do not exist.

Therefore, examining the meaning formula for (غدات), we get

\[
\text{غدات} = \text{زمن} + \text{نهار} + \text{محدد} + \text{أول} + \text{من الفجر إلى ضحاي}
\]

Also, checking the features, we can see that the feature (أول) has a value of (+1). Knowing this, we extract all words which have features with the value (-1) as a polarity. We find a number of words and not one. These are (رواح-دالوك-مساء-غروب-أصيل-عصر) (Rawah, Duluk, Evening, Sunset, Alassel, Late afternoon), and they all represent possible antonyms. With regards to synonymy, since this relationship between semantic units (words) can be identified by examining the semantic analysis (components) for each word, if the analysis is identical, then the two words are synonymous.

4.5.2. Stage2: Semantic application

The total number of words used during this stage was 51. We tested our ontology on the remaining Time nouns (31 nouns) and on new words from a new semantic field in the Holy Quran, the Human field. We used a selection of 20 words from the Holy Quran along with their features. The general concepts of our model have accommodated the new words seamlessly; however, we added a new semantic dimension specific to the human field.

5. RESULTS AND DISCUSSION

Results show that the model was able to cope with new nouns; however, some semantic dimensions were added to the model to accommodate new features. Lexical relations were also checked to verify that the model captures them sufficiently. Although there exists a dimension of “dynamism” which has two values, \textit{static} and \textit{dynamic}, the model does not capture the element of temporal sequencing in time. For example, \textit{summer} follows \textit{spring}, and \textit{today} comes before \textit{tomorrow}. Our model assumes that if X \textit{isPartOf} Y, then this implies that features associated with word X are also features for word Y, and are added to the componential formula. However, during the evaluation it appears that there exists words for which this statement is not true. For example, \textit{summer} \textit{isPartOf} \textit{year}; however, \textit{year} \textit{hasFeature} \textit{abstract}, while \textit{summer} \textit{has Feature} \textit{concrete}. Therefore, this \textit{Embodiment} feature cannot be inherited. A proposed resolution to this issue is to attach certain properties to \textit{features} which describe the nature of these \textit{features}, whether they are inheritable (shared) or not.
Another interesting finding from our evaluation is what we refer to as the \textit{dispersion effect}. This effect occurs when there is minimal or no inclusion relationship within nouns of a specific semantic field, which results in a shallow and wide structure, instead of a deep and narrow one. This means that componential formula will be extremely short; therefore, meaning representation is not sufficient. With regards to “Vague” nouns, this effect is apparent. The componential formulae were very short, thus not giving depth of meaning as is the case with “Day” nouns. When we applied the model on nouns from a different semantic field “Human”, we also observed the \textit{dispersion effect}. Our findings of a limited number of features (\textit{semantic richness}) for concrete nouns vs abstract nouns, support those reported in the literature. Studies show that words referring to concrete semantic units have richer semantics than abstract ones, and within concrete semantic units, living things have more features than non-living (artifact) things [45].

Another important finding from our evaluation is that within the “Human” semantic field, it was difficult to identify semantic dimensions. This may be due to two reasons: the fact that the “Human” semantic field is a very large field, and the sample chosen is not focused on a specific domain within the “Human” semantic field.

The ontology proposed in this research is unique in representing componential analysis of Arabic vocabulary. Traditional approaches to Arabic language computational models were based on models of Roman languages. However, our proposed model has originated from an authoritative and rich source of Arabic language, \textit{i.e.}, the Holy Quran.

We do not claim that our model is comprehensive. However, we focused on the area where others have not tapped into, that of componential analysis. Additionally, since our model is implemented in OWL, it can easily be extended and linked to other ontologies such as SUMO [23], LMF, and LexInfo [31].

Furthermore, we believe that such a model for representing Arabic lexicons will enable the creation of a plethora of useful applications for processing Arabic natural language. Such applications include simplifying Arabic language teaching for non-Arabic speakers and building intelligent Arabic dictionaries.

Finally, the results of our work can be summarized as follows:

1. Finding appropriate semantic primitives (dimensions) was simpler in concrete concepts and nouns. However, this was not the case with abstract concepts.
2. The evaluation also highlights some difficulties associated with this approach to semantics; for example, identifying semantic dimensions, and those which have polarities was difficult.
3. Although the lexicon is built based on Time nouns in the Holy Quran, the model is capable of accommodating any Time noun in the Arabic language.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented an ontological model for a computational lexicon capable of representing Arabic language lexicons in a way which provides a foundation for building useful Arabic language applications using Semantic Web technologies. The model has been implemented on the Arabic language vocabulary associated with “Time” vocabulary in the Holy Quran. Results of the evaluation indicate that the model is capable of representing word semantics in a way that can facilitate semantic analysis of Arabic words and various useful applications.

The next natural step is to extend the model into other semantic fields and see how it can accommodate them. Since componential analysis and ontology population are human intensive processes, a major direction in future work is looking into strategies for automated ontology population using technologies such as Latent Semantic Analysis [46] and Formal Concept Analysis [47]. In addition, we plan to develop semantic web applications capable of exploiting the rich structure of the ontological model. We intend to develop an application which automatically performs semantic analysis of words. Another useful application on the horizon is word positioning within the semantic field (classification) based on known features of the word. Classifying a new word in the lexicon is not a simple task. However, using the proposed model, the linguist needs only to select certain features and the application can automatically detect the appropriate classification, and suggest it to the user. Moreover, visualizations of the ontology are useful for linguists in observing semantic field characteristics and language behavior in a certain field, such as word density, word movement, and other attributes of a semantic field.

ACKNOWLEDGMENT

This work is supported by King Abdulaziz City for Science and Technology (KACST), Riyadh, Saudi Arabia.

REFERENCES


