J. King Saud Univ., Vol. 14, *Comp. & Info. Sci*., pp. 1-23 (A.H. 1422/2002)

A Relational Approach to the Design of an Arabic Lexical Database

Suleiman Hussein Mustafa
Department of Computer Science
Yarmouk University, Irbid, Jordan

(Received 28 November 1999; accepted for publication 12 May 2001)

Abstract. In this paper, the Arabic lexicon has been investigated in the context of relational database theory. A feature analysis of lexical entities has been carried out which shows that lexical attributes can be classified into five categories comprising nineteen attributes, including form attributes, morphological attributes, functional attributes, meaning attributes, and referential attributes. Based on this analysis, eleven database relations have been identified which form the backbone of an Arabic lexical database, including: words, roots, forms, infinitives, verbs, nouns, plurals, particles, meanings, lexical functions, and cross-references. The design ideas discussed in this paper were tested using a sample of lexical items selected from a modern printed dictionary. The results of developing an experimental lexical database indicate that the relational approach provides an efficient method for storing and retrieving Arabic lexical information. It should be mentioned, however, that several problems were encountered when the printed data was translated into a database form. Some of these problems are inherent in the Arabic lexicon itself, while others are due to the way by which lexical information is presented by paper-based dictionaries.

1. Background

Language is composed of three basic components: phonetic, lexical, and syntactic. But they are not independent; one component is with little use without the other two [1]. Hence, there is an intimate connection between the general rules incorporated into a grammar (or NLP system) and the nature of entries in the lexicon. The lexicon provides the information not predictable from the rules. Such information feeds the rules and ensures they function correctly [2].

The lexical component appears to be the most interesting area of computer applications. That is because the lexicon involves a tertiary structure in which the following types of facts are represented: data (which is expressed by lists of words), information (which is obtained from the network of word relations) and knowledge (which is expressed by the concepts implied in the definitions) [3]. Some of that lexical structure is phonetic, some syntactic, some semantic.

There is a general consensus among scholars involved in natural language processing that the computational lexicon is a fundamental element in building computer systems in various areas of application. Examples of such areas include information retrieval, natural language front ends, text understanding, text generation, machine translation, and speech synthesis.

The set of tokens in a lexicon is not a list of isolated words, each is standing on its own; rather, one can say, it is a complex structure of lexical, semantic, and pragmatic inter-relations. An emerging line of research and development in recent years has focused on this major feature of the lexicon and how the relational database approach can be used in the construction of computational lexicons. Mel'cuk's new ideas of what he called an explanatory combinatorial dictionary (ECD) [4] and lexical functions (LFs) [5] have given a momentum to this research direction. A number of researchers [6-8] have added, or suggested modifications, to the taxonomy proposed by Zholkovsky and Mel'cuk.

ECD is assumed to give detailed information on any lexical unit and its relationships with other words. Mel'cuk maintains that this kind of information is just what a computer needs for NLP systems. On the other hand, LFs are a set of formal tools designed to describe, in a fully systematic and compact way, all types of genuine relations that obtain between lexical units (LUs) of any language. LFs have been grouped under a dichotomy of paradigmatic vs. syntagmatic relations. The first category subsumes all contrast and substitution relations that may hold between LUs in specific contexts, while the second holds between LUs that can appear together (i.e., co-occur) in the same phrase [9].

The last twenty-five years have seen a number of relational models in linguistics, psychology, and anthropology. The well-known psycho-linguist George Miller, for instance, has viewed the lexicon as a large matrix with all the words in a language along the top of the matrix and all the different meanings, those words can express, down the side. The matrix can be accessed from both sides. Relations between forms and meanings are represented as many: many mapping [1]. Computational linguists were quick to apply these theoretical models in various areas of natural language processing, such as question answering systems, machine translation, text generation and automatic paraphrase [10].

The utility, potential and real, of machine-readable dictionary (MRD) sources to most areas of natural language processing (NLP) is beyond question. Generating word lists, deriving semantic taxonomies of various types, providing browsing functionality, parsing of dictionary definitions, semantic analysis and processing, and text generation are examples of the areas wherein lexical databases have been utilized [2]. Applying the relational database approach to the design and construction of a lexicon makes it possible to extract completely different types of information (i.e., structured in several ways and at various levels) from the same basic data. By using a large number of secondary or alternate keys, all the relevant attributes can be directly accessed by many search keys [11, 12].

A great number of experimental lexical databases have been developed and tested. For many, data was extracted from commercial MRDs [13], such as Longman, Webster, and Oxford, while for others data was compiled and entered by manual or text parsing techniques. Their different computational orientations have lead to differences in their range of application, content, and structure. Many of the popular general-reference dictionaries (such as those mentioned above) have been converted into on-line database systems [14] and some have also become available through the Internet.

2. Related Work

While most people agree that the availability of well-formed computational lexicon is a prerequisite for any successful natural language processing (NLP) system, very little research has been carried out in the field of Arabic computational lexicography. Most of the work in this area has been a side effect of developing experimental NLP or developing commercial software for word processing.

On theoretical foundation side, Ali [3] presented a model for building an Arabic lexical information system. His study provides a comprehensive discussion of all the aspects and the problems to be considered in designing such a system. The whole review is based on the idea that the lexicon is viewed as a complex system in which all linguistic knowledge is made available. This idea contrasts with that of Hassan [15] who believes that the lexicon does not contain a network of natural relationships and its contents cannot be represented in a tabular form, therefore we cannot consider it as a system.

Ali views the lexical database as having a set of five basic files to be supported by four lists. The files represent irregular plurals, semantic features, semantic domains, word definitions, and grammatical categories, while the lists consist of root morphological forms, trilateral infinitives, exceptional cases, and semantic relations of multi-word entries.

Another conceptual model of an Arabic lexical knowledge-base was also reported by Hashish [16]. He viewed the database as a set of nodes and links along with a set of selected features working together with the augmented known Arabic rules for the conjugation and derivation of all words in the deep form and the lexical rules for obtaining the surface (written) forms. According to Hashish, the database should be composed of five files, each of which represents one of the Arabic root categories.

On the practical side, a number of researchers have reported various implementations of Arabic computational lexicons. But, the majority were designed as part of lexical or/and syntactic analysis systems and did not incorporate anything more than the minimum of information required to perform the intended NLP task. Very few of the implementations were addressed from a database perspective. In the following paragraphs, we review the major efforts in this regard.

Shalabi [17] reports that, in the course of developing, what he calls, an automatic parser for Al-Alamiah software company (*Sakhr*), a lexical relational database containing 170,000 entries (stems) was built. He claims that the database includes all linguistic information (morphological, lexical, and semantic) for all Arabic words. The approach taken in this system seems to differ from that reported by InfoArab software company which designed a lexical system, called *Abjad-Hawaz*. It has been reported [18] that the lexicon includes about seven million Arabic words.

Al-Hannash [19], has also worked on a lexical database as part of a research program for Arabic language processing at the Center for Informatics and Computer Arabization in Morroco. The database contains about 45,000 simple entries and about 30,000 multi-word entries.

Another lexical database subsystem has been reported by Al-Hafez and his colleagues [20]. They have described a knowledge-based lexicon as part of an undergoing project for a comprehensive Arabic NLP system. The lexicon is divided into two aspects: morphological and semantic. Detailed information is incorporated about each word, such as its grammatical categories, gender, number, case, and affixes. The word meaning is represented by a combination of semantic aspects, including primitives, features, fields, domains, and rules of disambiguation.

Ditters [21] has been working on a project for automatic syntactic analysis of modern standard Arabic (called ASCAMSA). He presents a set of basic formal rules for describing entries in an Arabic lexicon. The lexicon consists of 7,000 entries of consonantal roots in the first stem active voice. The general form of a lexical verb entry is as follows: VP 3LEX (r1, r2, r3, complementation, vowelpast, vowelpresent, infinitive).

Hamrouri [22] presented a case for compressing lexical data using the primitive morphological properties of Arabic words. His results indicate that the current methods for storing the lexicon are not optimal and can be compressed by applying compression techniques to word affixes.

Ben Ahmed and Zriqui [23] have examined the possibility of generating a theoretical trilateral lexicon using the morphemic structure combination (MSC) method. Their results show that the five rules used in the generation process did not prove to be adequate for acceptable, error-free and usable trilateral roots. Their approach seems to be similar to that adopted by Al-Fedaghi and Yaseen [24] who report that their experiments resulted in 96.9% success in generating correct Arabic words. A rather different approach has been reported by Al-Jabri and Mellish [25], who have used semantic descriptions on the basis of standard Arabic morphological forms to generate Arabic words. No results have been reported so far.

This paper describes a methodology for constructing and structuring a relational Arabic lexical database. It also presents the problems encountered in developing an experimental system which has been implemented using Microsoft Access database management system. The system incorporated a sample of about one thousand lexical items selected from one of the most commonly used modern printed Arabic dictionaries (i.e., Intermediary Dictionary)[1]

3. Lexical Entities and Attributes

There is no consensus among scholars working in the area of computational linguistics regarding the nature of lexical entities that should be included in a lexicon. The simplest notion of a lexicon holds that it is a collection of words, with associated information about them [1, p.32]. But, what is it we mean by "word" ? Is it a concept, a single orthographic representation, or a morpheme (bound or free)?

In addition, lexical units are often not just single word items. Many words, in some languages, are morphologically complex forms and some phrases of the type we encounter in printed dictionaries (like idiomatic expressions) are treated as multi-word lexical units. Complex forms are rarely used and represented in Arabic dictionaries. For instance, a dictionary of modern Arabic, containing about 54,000 entries, has been examined to count the number of multi-word entries. The result showed that the number of such lexical units did not exceed 168 entries.

---

[1] المعجم الوسيط (مجمع اللغة العربية في القاهرة)

This is a very small number when compared with number of idiomatic expressions reported by Al-Hannash [26]. He claims that the number goes beyond 30,000 entries (out of the 75,000 entries in his database) which would represent about 40% of the Arabic lexical database. He based his work with these expressions on the assumption that Arabic words assume different meanings in different contexts.

The other issue raised, in this regard, is how comprehensive the lexicon should be in terms of derivatives? Some researchers maintain that it is a matter of personal taste how far the lexicon should include regular derivations (e.g., weakness from weak), morphological forms (such as plurals), and other computable word forms as separate entries rather than leaving them to be computed [27]. Others believe that the lexicon may not include entries for some derived words whose behavior is predictable on the basis of morphological rules [2].

Another question that should be addressed in this regard concerns the details that should be included under each lexical entry. A criticism that is always raised against classical Arabic dictionaries is the lack of consistency and acceptable level of details.[28] As Hassan indicates, the lexicon must include the following types of data: pronunciation, spelling, morphology, orthographic variations, derivation, meanings, examples, and usage [15].

One of the problems to be considered is that of orthographic resemblance (like lead and lead in English) - what looks identical, in the absence of diacritics, may represent different types of information. A given nondiacritized Arabic word (such as a word composed of “meem + noon"[2]) could be treated sometimes as a verb, a noun, a relative pronoun, or a particle. But, there are not many words in Arabic of this category. Contrary to English, which depends on the context to determine a given word’s classes, words in Arabic assume independent meaningful entities both inside and outside context. The vocalized word “Rajol”[3] for instance, is always classified as a noun regardless of any context in which it could appear [29].

For the purpose of the present research study, we viewed a lexical unit as having five major aspects: an orthographic form (which specifies the presence or absence of diacritical marks), a morphological base (which specifies the root, the derivation form which is known in Arabic as “wazn”, the infinitive, and the tense form), a function (which specifies its role in terms of part of speech and class), a meaning or set of meanings (which specifies its lexical and functional content along with fields of

---

[2] Word with diacritics:"مَنَ،مَن،المن،مِن".

[3] Word: "رجل" means “a man”.

application and examples of usage), and a cross reference aspect (which specifies its relationships with other lexical units and its different orthographic variations, if any). Fig.1 presents these aspects and their various components.

Traditionally, Arabic lexical units are classified into three categories which are the basic building blocks of a lexical database: verbs, nouns, and particles. Based on the general lexical structure presented in Fig. 1 a feature analysis was carried out to identify the attributes that should be considered for inclusion in the proposed lexical database. Figs (2 – 4) and appendices (1 - 3) present the results of this analysis.[4]

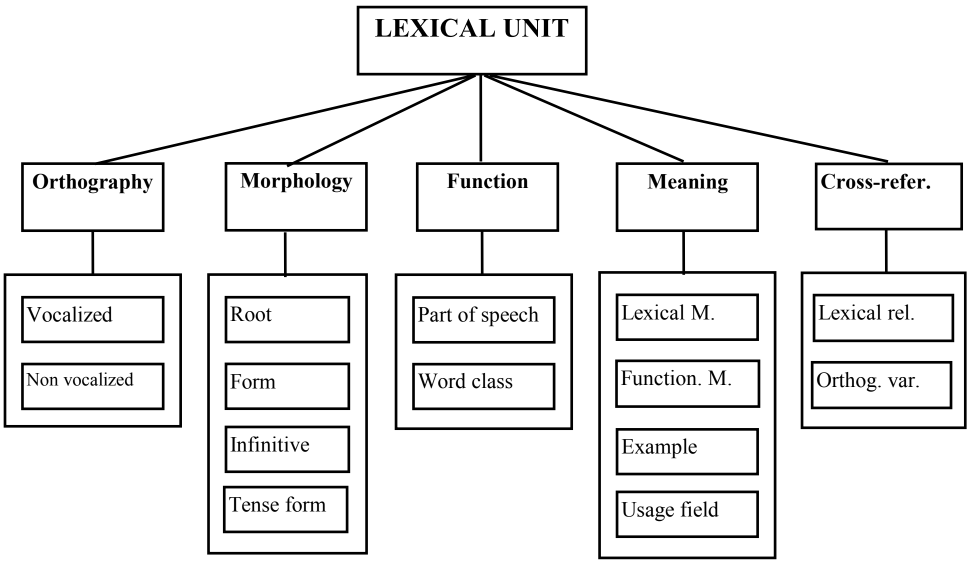


Fig. 1. Arabic lexical unit structure.

[4] The abbreviations in Figures 1-4 are read as follows:M=Meaning, Rel=Relation, Orthog=Orthograhic, Var=Variation, Refer=Reference, Voc=Vocalized, Lex=Lexical, Func=Functional, Int=Intransitive, Neut=Neutral

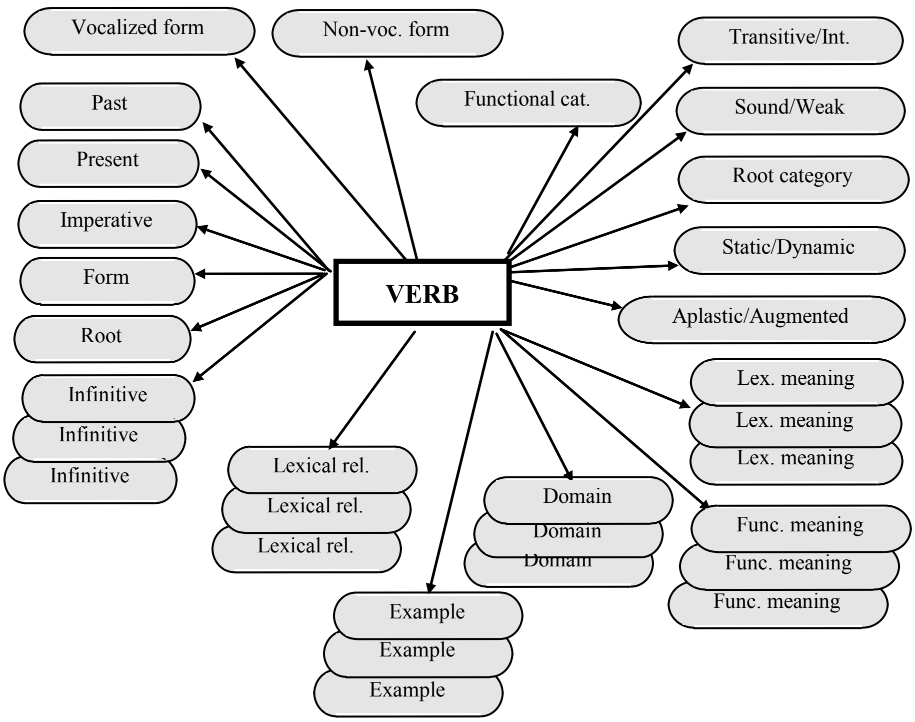


Fig. 2. Arabic verbal attributes structure[5].

As Fig. 2 shows, a lexical database should include about nineteen attributes, some of which are aggregates (such as meanings, examples, ... etc.). Using the lexical structure of Fig.1, these attributes are grouped in five categories as follows:

a. Form attributes: vocalized form, and non-vocalized form.
b. Functional attributes: functional category and lexical class, in terms of transitivity (transitive or intransitive), transformation (sound or week), root category, static or dynamic, aplastic or augmented. A taxonomy of verbal classes is given in *Appendix* (1).

---

5 Vocalized / Nonvocalized (مشكول أو غير مشكول), transitive or intransitive (متعدي أو غير متعدي), sound or week (صحيح أو معتل), static or dynamic (جامد أو مشتق), aplastic or augmented (مجرد أو مزيد).

c. Morphological attributes: tense modes (past, present, and imperative), form, root, and infinitive.
d. Meaning attributes: lexical meanings, functional meanings, domains of application, and examples of usage.
e. Lexical relations.

As Fig. 3 indicates, the noun lexical unit involves nineteen major attributes. Many of them are similar to those found in the analysis of verbal features such as the orthographic forms and variations, meanings, and domains of usage. The other attributes are unique to the nominal lexical unit, including gender, number, and the various lexical classes. As in the case of verbal lexical units, an attempt was made to devise a nominal taxonomy in order to direct the process of assigning values to nominal lexical classes (see *Appendix* 2).

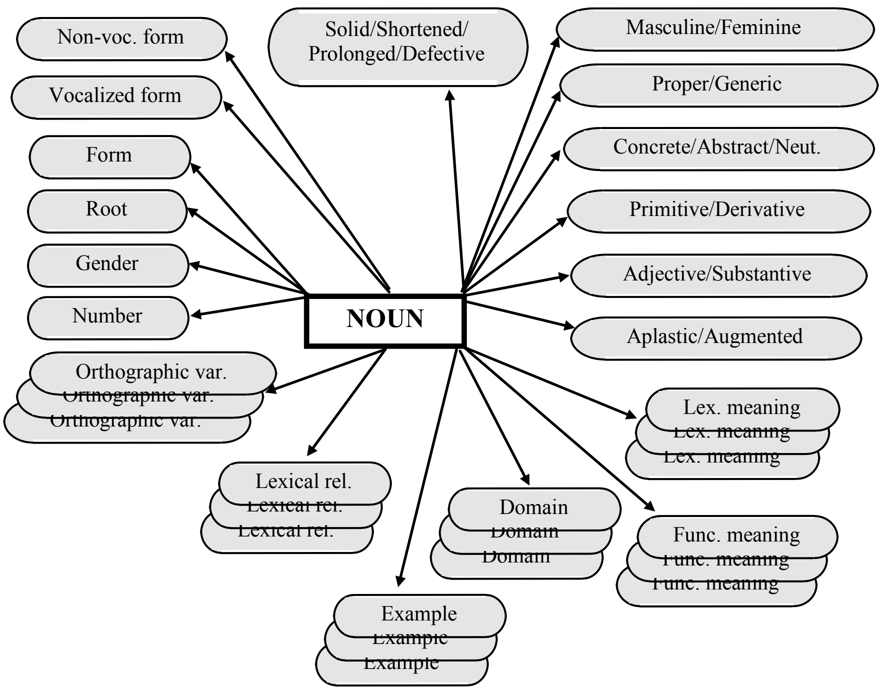


Fig.3. Arabic nominal attributes structure[6].

6 proper / generic (علم أو جنس), concrete / abstract/ neutral (اسم ذات، أو معنى، أو عام), solid / shortened / prolonged / defective (صحيح الاخر، أو مقصور، أو ممدود، أو منقوص), primitive / derivative (جامد أو مشتق), adjective / substantive (صفة أو موصوف).

It is important to note here that there is no consensus among Arab grammarians as to what comes under a lexical taxonomy. Because the same lexical unit assumes different lexical and functional roles, it is classified under different categories. In some cases, the same category is treated differently by different grammarians. This represents a problem that has to be addressed in the design of a relational lexical database. The taxonomies provided at the end of this paper were intended to address such a problem.

However, such taxonomies are hierarchical in nature. This might lead to functional transitive dependencies between different classes. This represents another problem that had to be considered in the design of the lexical database. In the relational database theory, non-key attributes are assumed to be mutually independent and fully dependent on the primary key. The existence of such transitive dependencies between attributes implies that an attribute cannot be updated independently of all the rest [30]. This special nature of lexical data has led some researchers in the field to advocate the idea of having a dedicated database management system for lexical databases [31].

The simplest of the three lexical units is the particle. As we consider the analysis of its attributes as shown in Fig. 4 and *Appendix* (3), we find that, one of the five major features represented in Fig.1 (i.e., the morphological base) does not exist in the case of particles and only the functional meanings are present in the list of features. As in the case of verbal and nominal lexical units, the taxonomy presented in *Appendix* (3) was intended to identify the values that could be assigned to the lexical and functional particle classes and roles.

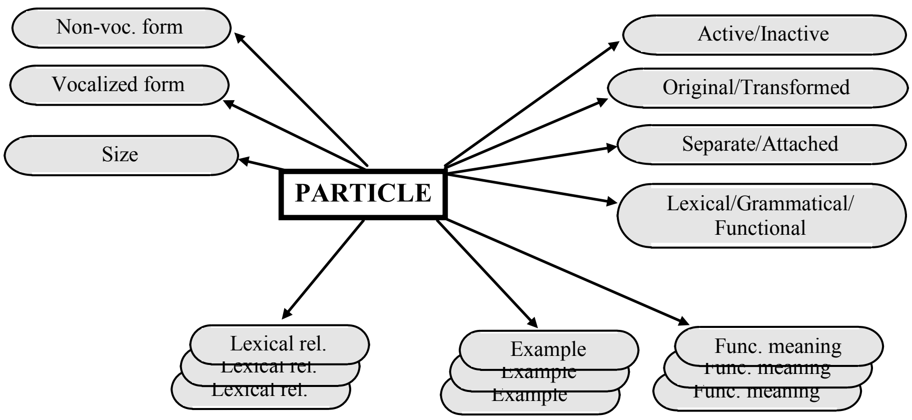


Fig. 4. Arabic particle attributes structure[7].

[7] active / inactive (عاملة أو غير عاملة), original / transformed (أصلية أو محوّلة), separate / attached (منفصلة أو متصلة).

4. Lexical Relational Structure

Having identified the lexical entities and their attributes, as described in the previous section, the next step was to identify the lexical database relations and the relationships between various relational entities. What follows is a list of the relational tables that were included in the design of the lexical database, along with their formal definitions. (Note that the abbreviation *Voc* = Vocalized, *Cat* = Category, *Intran* = Intransitive).

- *Words [Non-Voc-Word, Part-of-Speech, Reference-Exists]*
- *Roots [Voc-Root, Non-Voc-Root, Category]*
- *Forms [Voc-Form, Non-Voc-Form]*
- *Infinitives [Non-Voc-Infinitive, Non-Voc-Root, Non-Voc-Form]*
- *Cross-References [Voc-Word, Non-Voc-Word, Voc-Cross-Reference]*
- *Verbs [Voc-Verb, Non-Voc-Verb, Past, Present, Imperative, Voc-Form, Voc-Root, Transitive-Intran-Cat, Transitive-Cat, Solid-Weak-Cat, Solid-Cat, Hamzated-Cat, Weak-Cat, Root-Cat, Aplastic-Augmented-Cat, Augmented-Cat, Static-Dynamic-Cat, Conceptual-Cat]*
- *Nouns [Voc-Noun, Non-Voc-Noun, Voc-Root, Voc-Form, Gender, Number, Proper-Generic-Cat, Generic-Cat, Primitive-Derivative-Cat, Derivative-Cat, Adjective-Cat, Substantive-Cat, Semi-Noun-Cat, Aplastic-Augmented-Cat, Solid-Nonsolid-Cat]*
- *Plurals [Voc-Plural, Voc-Noun, Non-Voc-Noun]*
- *Particles [Voc-Particle, Non-Voc-Particle, Size, Active-Inactive-Cat, Original-Transformed-Cat, Nominal-Verbal-Cat, Grammatical-Functional-Cat]*
- *Meanings [Voc-Word, Non-Voc-Word, Lexical Meaning, Functional Meaning, Example, Domain of Usage]*
- *Lexical-Functions [Voc-Word, Non-Voc-Word,Word-Related-To, Type-of-Lexical-Relation]*

The first relation was intended to act as a comprehensive index for all Arabic lexical items included in the database, through which a user can access any given word in its nonvocalized form. Knowing, through this table, that a lexical item is a verb, a noun, or a particle a link can be established with the appropriate relation (i.e., the relation *Verbs*, *Nouns*, or *Particles*) where lexical information is located. On the other hand, if the user is interested in knowing the meanings of a given word, a link is established with the table *Meanings* through the attribute *Non-Voc-Word*.

Other relations that are considered as support tables include: *Roots*, *Forms*, *Infinitives*, and *Cross-References*. Knowing the root of a given word, the user can tell, through *Roots*, to what category it belongs. The table *Forms* was intended to list all the

Arabic morphological forms, with and without diacritical marks, so that any form can be traced in the other relations. Roots and morphological forms are also used to access infinitives through the table *Infinitives*. As of the table *Cross-References*, its purpose is to refer to other orthographic or lexical variations.

The purpose of the other relational tables is to provide complete lexical and semantic information about words. Such information is found in the following relations: *Verbs*, *Nouns*, *Particles*, *Plurals*, *Meanings*, and *Lexical-Functions*. The last is supposed to maintain links between related lexical items.

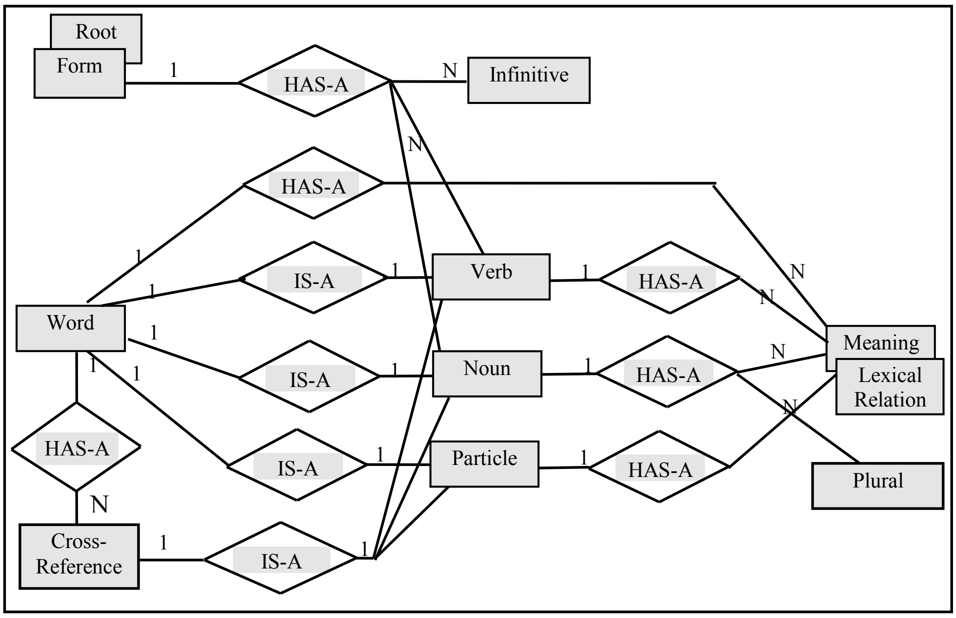


Fig. 5. Relationships between lexical entities in the database.

Relationships between the various entities are presented in Fig. 5. As this figure shows, the entity "*word*" is related to the entities "*Verb, Noun,* and *Particle*" with a one-to-one *IS-A* relationship, while the rest of connections between entities are represented by one-to-many and *HAS-A* relationships. These relationships are maintained in the tables through the non-vocalized form of the given words. For instance, a word from the table "*Words*" could be accessed in the table "*Verbs*" or "*Nouns*" or "*Particles*" using the attribute "*Non-Voc-Word*". The same also applies to other tables, where the primary key could be the vocalized form.

The fact that the non-vocalized form of words was used in the table "*Words*" as a primary key whereas the vocalized form was used as primary key in other relations is to allow the normal user to access any word in the database through the table "*Words*" without the need to worry about diacritical marks. Even people with solid background in vocalization might have access problems when different levels of vocalization is applied for different words. But, while the non-vocalized form is more efficient in terms of access, it fails to provide a unique primary key for storing Arabic words. It was, therefore, necessary in the design of the lexical database to use the vocalized form as a primary key in the relations: "*Verbs, Nouns, Particles,* and *Plurals*", with the non-vocalized form being used as a secondary key.

As of the last two relations (i.e., *Meanings* and *Lexical-Functions*), no primary key was used due to the fact that access to these functions is dependent on the other relations (i.e., *Words*, *Verbs*, *Nouns*, and *Particles*) which are related to these two entities by a one-to-many relationship. The user can access the two relations in one of two ways:

(a) either access the table "Words" then use the key "*Non-Voc-Word*" to the required meanings/lexical functions, or
(b) initially access the given word in "*verbs/Nouns/Particles*" then use the key represented by the vocalized form to access its meanings/lexical functions.

As an example of how these relational tables are used, let us assume that a user submits a query relating to the word "منزل" ("*menzil*" = "a house"). At first, by accessing the table "*Words*" we find that this word belongs to the nouns category (i.e., Part-Of-Speech = "N"). An access link is then established with the table "*Noun*" through which we can retrieve most of the features of the word, including its vocalization. If the list of meanings is required by the user, the access mechanism will lead us to the table "*Meanings*" using the vocalized form of the given word. The retrieval navigation process continues as long as there is more information requested by the user of the database.

5. Database Implementation

The ideas presented in this paper were tested using a sample of about one thousand lexical items (representing 300 root entries) which were chosen from the Arabic "Intermediary Dictionary". The dictionary follows an old Arabic tradition of grouping words under their roots. Each entry was analyzed in terms of the lexical items included along with their associated attributes and inter-relationships. The analysis also involved the checking of orthographic variations, homonyms, synonyms, irregular forms, and various inflections.

Although, the listing of entries and their derivatives in the dictionary follow a predetermined order, the information given under each head-word (be it a direct entry or a sub-entry) presented a real challenge in the development of the prototype lexical database. Differences in the design philosophy between the printed dictionary form and the lexical database form posed several problems which can be summarized in the following paragraphs. One could assume that the design orientations of both forms share the lemma through which all (or most) of the inflected word forms are grouped together or related in some way to maintain the morphological relations. But, in a lexical database, more emphasis is directed towards the macrostructure of the lexicon along with the access mechanism by which every lexical entity can be equally accessible.

5.1 Qualifying phrases

At first sight, one might be likely to think of words in the dictionary of a given language as being independent distinct lexical entities. Once you consider how these words are represented in the dictionary, it turns out that this is not true. A great number of the lexical entries in the Arabic dictionary, especially in the case of verbs, are qualified by context information such as agents, subjects, or other qualifying phrases as in the following examples (quoted from the Intermediary Dictionary).

(أمطرت) السماء: .... (مضغ) الطعام: ....

(امتعض) من الأمر: ....   (المتعمدة) من الرطب: ....

(معط) في القوس: ....   (ماطله) بحقه: ....

(مضمض) النعاس في عينه: .....   (مقت) إلى الناس: ....

In developing the database, the lexical tokens (designated in the dictionary as head-words) were considered the building blocks of all tables, regardless of the qualifying phrases associated with them. Such context information was treated as part of the list of meanings rather than the primary or secondary key information. It follows that head-words such as "أمطرت" ("*amtaret*" = "it rained") or "ماطله" ("*matalahu*" = "he procrastinated"), which are given in a form that conforms with the qualifying phrases, would be included in the database access keys as: "أمطر" or "ماطل" instead.

5.2 The definite article "*al*" (أل)

To be consistent with their definitions, almost all nominal entries in the Arabic dictionary are listed in the definite case. Consider, for instance, the following two items ("*al-maheed*" and "*al-mehr*"):

(المهيد): الزبد الخالص.   (المهر): صداق المرأة.

Since words are ordered according to their roots, this practice does not affect the efficiency of access to any given word. Even in the case of non-derived words which are listed in their proper alphabetical positions (or in the case where the alphabetical order is strictly applied regardless of root affiliation), the definite article is ignored when entries are sorted.

This is not the true in the case of lexical databases where a strict alphabetical order is maintained using one of the well-known sorting algorithms. As a result, all nouns would be expected to cluster in one area, which would lead, in turn, to significant deterioration in efficiency of both sorting and access.

Therefore, it was decided in implementing the database to remove "al" from all lexical entries. Meanwhile, consistency between these entries and their definitions was maintained by repeating the full word (i.e., with "*al*") in the definitions as follows ("*mehr*" and "*maheed*"):

(مهيد): المهيد الزبد الخالص.   (مهر): المهر صداق المرأة.

5.3 Completeness of information

There is a great difference between a manual dictionary and the lexical database, as described here, in terms of the kind of information that ought to be included. Some information that might be viewed by dictionary compilers as irrelevant or easy to deduce by users (on the assumption that they already know the morphological rules) must not be viewed so by designers of lexical database. For instance, a dictionary might not include the plural of a word like "مدرسة" ("*medrasah*" = "a school"), or "جامعة" ("*jami'ah*" = "a university") on the assumption that all speakers of the language would know their plurals.

As one exposes this assumption to testing, it turns out that the normal user of manual dictionaries might not be able to tell a lot of what would be considered default information. For this reason and for computational linguistic reasons, the proposed lexical database model, described in this paper, is based on the assumption that the information given about any lexical item must possess a high level of completeness.

As we started implementing the lexical database using the printed dictionary, described above, a great amount of the information required by the design was not available. Therefore, it was important to check other dictionaries (which helped in some cases) and to apply the rules of morphology (wherever necessary) to complete the missing information. Based on the experience gained from this project, one can assume that almost about one third of the amount of information stored in the database came from supplementary sources.

5.4 Semantic knowledge

Although the prototype system described in this paper was not intended to be a lexical knowledge base, including semantic knowledge was assumed to be an inevitable part of any lexical database. Such knowledge was incorporated in the system under a number of entities (particularly the attributes relating to lexical classes, meanings, and lexical functions). The task of including semantic information was one of the most problematic issues in developing the lexical database. Not only because such information is severely scarce in dictionaries, but also because this information has been rarely a major concern in the study of the Arabic lexicon.

There are, of course, some sources [32-34] that can be used for this purpose besides what is included in dictionaries, but the job of locating semantic information and developing semantic relations is a very tedious one. In many cases, this information has to be deduced from whatever data available in dictionaries or other sources. This means that language expertise would be important to tell what semantic content a given lexical token might convey.

The word taxonomies presented at the end of the paper were intended to provide a framework for assigning semantic values and maintaining semantic and lexical relationships between related lexical items.

6. Conclusion

This paper presents a model for storing and retrieving Arabic lexical information based on the database relational approach. While this idea is not new in the field of natural language processing and information retrieval, most of the researchers who have investigated the issue of designing an Arabic lexical database directed most of their focus on the theoretical aspects of the subject. Instead, the research reported in this paper, paid more attention to the real practical problems of designing and implementing a relational lexical database.

The analysis of entities was based on the traditional paradigm of categorizing Arabic words into verbs, nouns and particles. An attempt was made to present each category in the form of a taxonomy. Attributes, on the other hand, were also categorized into five groups: form attributes (vocalized form, and non-vocalized form), morphological attributes (tense modes: form, root, and infinitive), functional attributes (in terms of transitivity, transformation, root category, static or dynamic, aplastic or augmented), meaning attributes (lexical meanings, functional meanings, domains of application, and examples of usage), and lexical relations.

Based on this analysis, eleven database relations were identified which formed the backbone of an Arabic lexical database, including: words, roots, forms, infinitives, verbs, nouns, plurals, particles, meanings, lexical functions, and cross-references.

The design ideas discussed in this paper were tested using a sample of lexical items selected from a modern printed dictionary. The results indicate that the relational approach provides an efficient method for storing and retrieving Arabic lexical information. It should be mentioned, however, that several problems were encountered when the printed data was translated into a database form. Some of these problems are inherent in the Arabic lexicon itself, while others are due to the way by which lexical information is presented by paper-based dictionaries.

Some of the major problems encountered evolved around the following issues: making distinction between a word with diacritical marks and without for the purpose of facilitating access to data, dealing with transitive dependencies in presenting functional classes, handling qualifying phrases and the definite article used traditionally in the printed dictionary, and maintaining the completeness of lexical and semantic data as required by the database model being suggested and tested.

Acknowledgment. The work presented in this paper has been partially supported by a grant from Yarmouk University, Irbid, Jordan.

References

[1] Miller, George. *The Science of Words*. New York: Scientific American Library, 1991.
[2] Boguraev, Bran and Briscoe, Ted (Eds.*) Computational Lexicography for Natural Language Processing*. London: Longman, 1989.
[3] Ali, Nabil. “Automating the Arabic Lexicon”. In: Ali, Nabil. *Computers and the Arabic Language*[8]. Kuwait: Tareep, (1988), 457-529.
[4] Mel’cuk, Igor and Zholkvosky, Alexander. “The Explanatory Combinatorial Dictionary”. In: Evens, Martha Walton (Ed.) *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, 41-74.
[5] Mel’cuk, Igor. “Lexical Relations: A Tool for the Description of Lexical Relations in a Lexicon”. In: Wanner, Leo. *Lexical Functions in Lexicography and Natural Language Processing*, 37-102.
[6] Steele, J. and Meyer, I. “Lexical Functions in an Explanatory Combinatorial Dictionary: Kinds, Descriptions, and Examples in English”. In: Steele, J. (Ed.*) Meaning-text Theory: Linguistics, lexicography, and Implications*, 41-61.
[7] Grimes, Joseph E. “Inverse Lexical Functions.” In: Steele, J. (Ed.*) Meaning-text Theory: Linguistics, lexicography, and Implications*, 350-364.
[8] Ramos, Mararita and Tutin, Agne’s. “A Classification and Description of Lexical Functions for the Analysis of their Combinations”. In: Wanner, Leo. *Lexical Functions in Lexicography and Natural Language Processing*, 147-179.

---

[8] Arabic title: “اللغة العربية والحاسوب”.

[9] Wanner, Leo. *Lexical Functions in Lexicography and Natural Language Processing*. Amesterdam/Philadelphia: John Benjamines, 1996.

[10] Evens, Martha Walton (Ed.) *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. Cambridge: Cambridge University Press, 1988.

[11] Calzolari, Nicoletta. "The Dictionary and the Thesaurus can be Combined". In: Evens, Martha Walton (Ed.) *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, 75-96.

[12] Tompa, F. "Database Design for a Dictionary of the Future"`. *Preliminary Report Center for the New Oxford English Dictionary*, University of Waterloo, Waterloo, Ontario, 1986.

[13] Heid, U.; Heyn, M. and Christ, O. "Extracting Linguistic Information from Machine-readable Versions of Traditional Dictionaries". In: Kiefer, F. *et.al*., (Ed.) *COMPLEX '92: Papers in Computational Linguistics*. Budapest: Hungarian Academy for Sciences, (1994), 137-157.

[14] Al-Shawi, Hiyan *et.al.* "Placing the Dictionary Online". In: Boguraev, Bran and Briscoe, Ted. (Eds.) *Computational Lexicography for Natural Language Processing*, (1989), 41-63.

[15] Hassan, Tamam. *Arabic Language: Semantics and Etymology*[9]. 2nd ed. Cairo: Al-Haiah Almisriah Al-Aammah Lil-Kitab, 1979.

[16] Hashish, Muhammed A. "Arabic Language Processing." *Proceedings of the Symposium on Using Arabic in Information Technology*, Riyadh (S. Arabia), 10-14, (May 1992), 73-82.

[17] Shalabi, Ashraf. "Automatic Parsing of Arabic Sentences". *Proceedings of the 4th International Conference and Exhibition on Multi-lingual Computing: ICEMCO-94*, 7-9, (April 1994), 3.2.1-10.

[18] PC Magazine "Abjad Hawaz Lexicon"[10]. *PC Magazine* (Arabic Ed.), 1, No.1, (Nov.1994), 16-19.

[19] Al-Hannash, Muhammed. "A Database for Arabic Idiomatic Expressions"[11]. *1st Conference on Arabic Computational Linguistics*, Cairo, 20-22 June 1992.

[20] Al-Hafez, M.Y., Maryati, M., Vella, A. and Clarke, J.D. "Design of an Arabic Language Knowledge-base ass a Lexicon for NLP". *Proceedings of the 3rd International Conference and Exhibition on Multi-Lingual Computing: ICEMCO-92*, 10-12, (Dec. 1992), 2.4.1-17.

[21] Ditters, Everhard "The basic Structure of a Formal Arabic-English Verbal Lexicon". *Proceedings of the 4th International Conference and Exhibition on Multi-Lingual Computing: ICEMCO-94*, 7-9 April 1994), 3.4.1-18.

[22] Hamrouri, Boubaker Meddeb "Logic Compression of Multilingual Dictionaries*". Proceedings of the 4th International Conference and Exhibition on Multi-Lingual Computing: ICEMCO-94*, 7-9 April 1994, pp 3.7.1 - 8.

[23] Ben Ahmed, M. and Zriqui, M. "Automatic Generation of the Arabic Theoretical Lexicon Using Morphematic Structure Combination". *Proceedings of the 4th International Conference and Exhibition on Multi-Lingual Computing (Roman and Arabic Scripts): ICEMCO-94*, 7-9 April 1994, pp 3.1. 1 - 8. {Written in Arabic}

[24] Al-Fedaghi, Sabah and Yaseen, Mustafa. "Applying the Automatic Combinatorial Theory to Non-Diacritized Arabic Text"[1]. *Arabuter*, 3, (18-19 August 1991), 19-29, 31. {Written in Arabic}

[25] Al-Jabri, Saad and Mellish, Chris. "Generating Arabic Words from Semantic Descriptions." Proceedings of the 4th International Conference and Exhibition on Multi-Lingual Computing: ICEMCO-94, (7-9 April 1994), 9.6.1-11.

[26] Al-Hannash, Muhammed. "Databases for a Computational Arabic Lexicon"[12]. *Proceedings of the 3rd International Conference and Exhibition on Multi-Lingual Computing: ICEMCO-92*, (10-12 Dec. 1992), 12.4.1-10.

---

[9] Arabic title: "اللغة العربية: معناها ومبناها".

[10] Arabic Title: "قاموس أبجد هوز"

[11] Arabic title: "قاعدة بيانات التعابير المسكوكة في اللغة العربية".

[12] Arabic title: "قواعد بيانات من أجل بناء معجم آلي للغة العربية".

[27] Evens, Martha "A Lexicon for a Medical Expert System". In: Evens, Martha W. *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*, 97-111.

[28] Nassar, Hussein. *The Arabic Lexicon: Evolution and Development*.[13] Cairo: Daar Misr Lil-Tiba'ah, 1988.

[29] Halwani, Muhammed K. *A New Handbook of the Science of Morphology*[14]. Beirut: Dar Al-Sharq Al-Arabi, 1998.

[30] Date, C.J. *Introduction to Database Systems* (Vol.1). Reading(MA): Addison-Wesley, 1990.

[31] Domenig, M. & Shan, P. "Towards a Dedicated Database Management System for Dictionaries". *Proceedings of the 11th International Congress on Computational Linguistics (COLING 86),* Bonn, Germany, pp 91-96.

[32] Al-Sameraei, Fadhel Saleh. *Samantics of the Arabic Standard Forms*.[15] Kuwait: Kuwait University, 1981.

[33] Al-Shamsaan, Abu Aws Ibraheem. *Verbal Standard Forms: Semantics and Relationships.*[16] Jeddeh: Dar AlMadani, 1987.

[34] Al-Zajjaj, Abu AlQasem Abdelrahmaan. *Particles Conveying Semantic*[17]. Ali AlHamed (Ed.), Beirut: Muasasat AlRisalah and Daar AlAmal, 1986.

[13] Arabic title: "المعجم العربي: نشأته وتطوره".

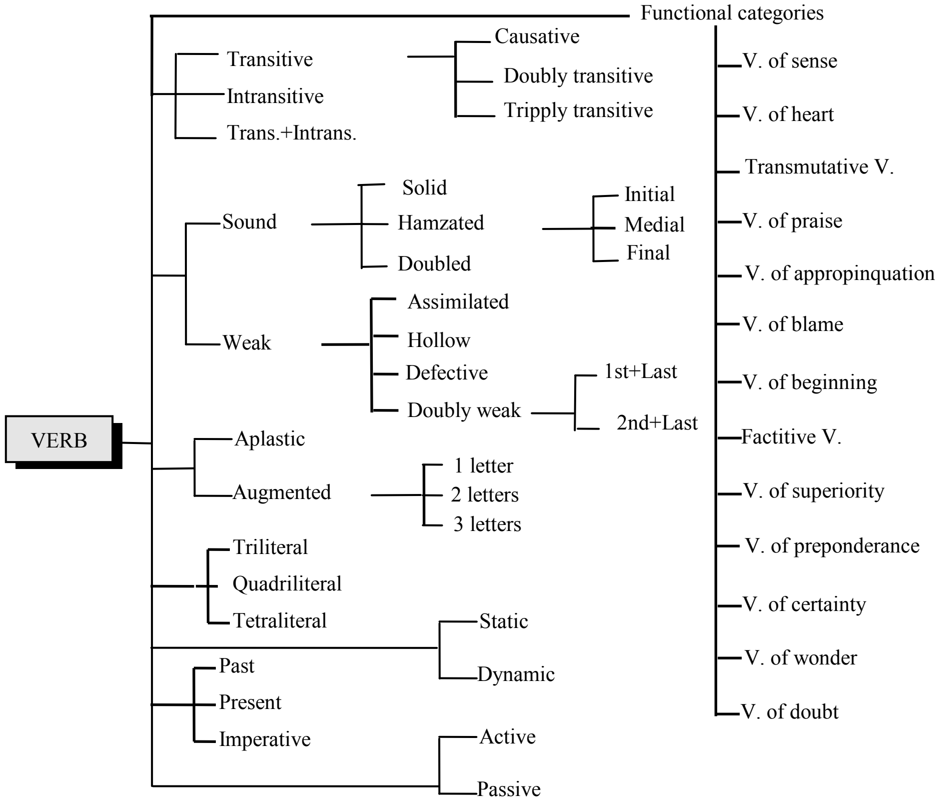
[14] Arabic title: "المغني الجديد في علم الصرف".

[15] Arabic title: "معاني الأبنية في العربية".

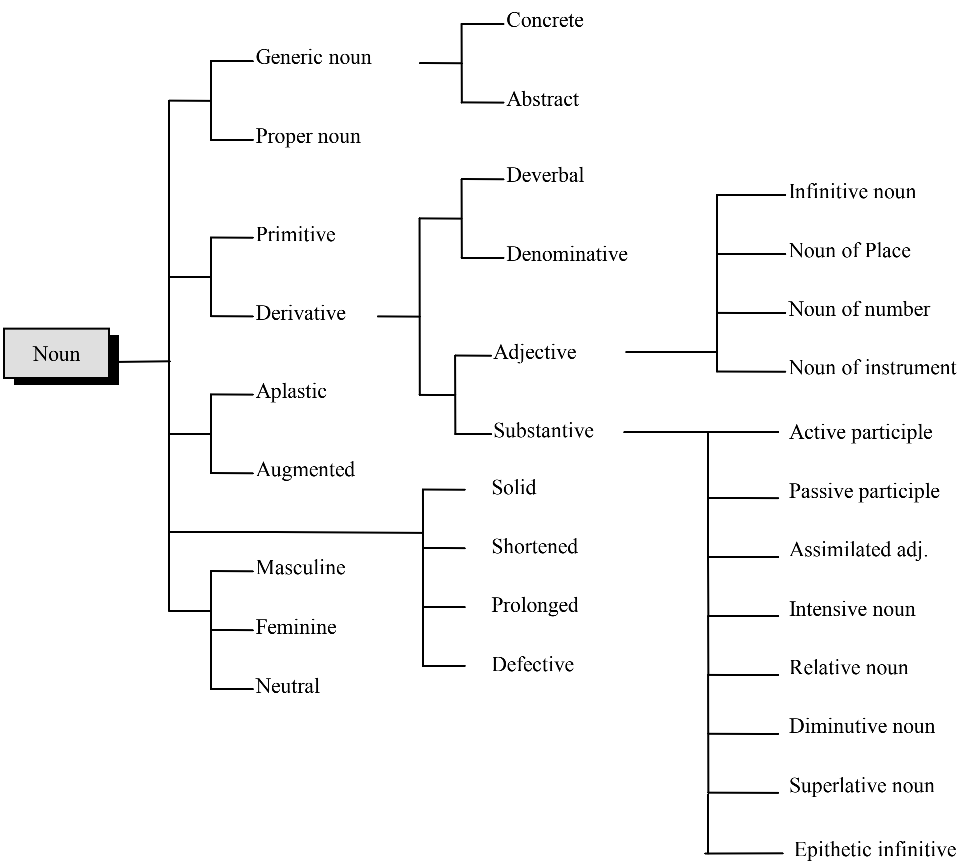
[16] Arabic title: "أبنية الفعل: دلالاتها وعلاقاتها".

[17] Arabic title: "حروف المعاني".

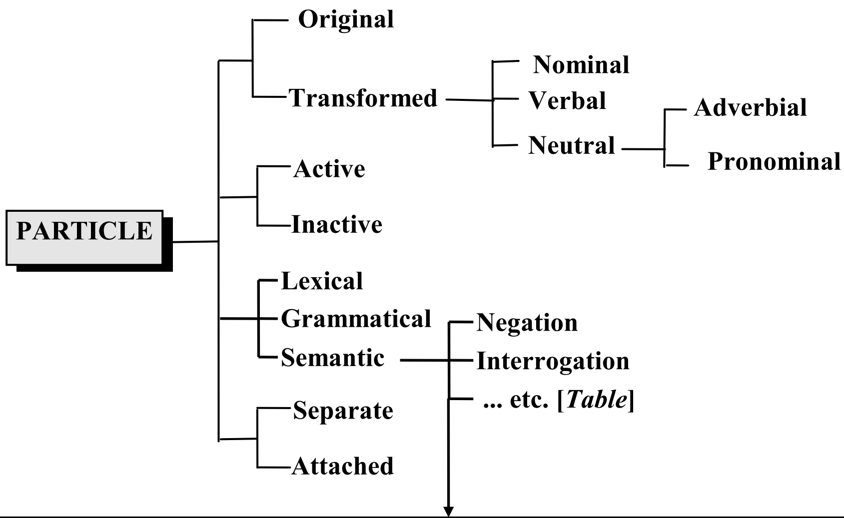
APPENDIX 1: *A TAXONOMY OF ARABIC VERBAL CLASSES*



APPENDIX 2: *A TAXONOMY OF ARABIC NOMINAL CLASSES*



APPENDIX 3 : *A TAXONOMY OF PARTICLE CLASSES*



Function	Function	Function	Function	Function
Introduction ابتداء	Stimulation تحضيض	Interpretation تفسير	Preposition جر	Oath قسم
Exception استثناء	Attainment تحقيق	Seperation تفصيل	Response جواب	Source مصدر
Restriction استدراك	Selection تخيير	Abundance تكثير	Disapproval ردع	Simultaneity معية
Inception استفتاح	Appeal ترجي	Paucity تقليل	Increase زيادة	Lamentation ندبة
Interrogation استفهام	Comparison تشبيه	Optative تمني	Condition شرط	Call نداء
Futurity استقبال	Conjugation تصريف	Premonition تنبيه	Adverb ظرف	Negation نفي
Digression إضراب	Wonder تعجب	Regret تنديم	Exposition عرض	Prohibition نهي
Prevention امتناع	Definition تعريف	Confirmation توكيد	Conjungtion عطف	
Imperative أمر	Causation تعليل	State حال	Purpose غاية	

استخدام المنهج العلاقي في تصميم قاعدة بيانات معجمية عربية

سليمان حسين مصطفى
قسم علوم الحاسب ، جامعة اليرموك، اربد، الأردن

(قدّم للنشر في ١٩٩٩/١١/٢٨م؛ وقبل للنشر في ٢٠٠١/٠٥/١٢م)

ملخص البحث. تمت في هذا البحث دراسة القاموس العربي في سياق نظرية قواعد البيانات العلاقية. حيث جرى تحليل الوحدات اللفظية باستخدام أسلوب "تحليل الخصائص" وفقاً لتصنيف ضم خمس فئات تشمل ١٩ خاصية وهي: الخصائص الشكلية، والخصائص الاشتقاقية، والخصائص الدلالية، والخصائص الإحالية. وبناءاً على هذا التحليل، تم تحديد أحد عشر علاقة (من علاقات قواعد البيانات) تشكل العمود الفقري لقاعدة بيانات معجمية عربية، تضم: الألفاظ، الجذور، الصيغ، المصادر، الأفعال، الأسماء، الجموع، الأدوات، المعاني، الوظائف اللفظية، والإحالات. وقد تمت دراسة الأفكار الواردة في هذا البحث من خلال عينة من الألفاظ تم اختيارها من أحد المعاجم الحديثة. وقد أظهرت التجارب التي أجريت على قاعدة البيانات التجريبية أن المنهج العلاقي يوفر طريقة فعالة لتخزين المعلومات اللفظية واسترجاعها. ولكن ينبغي التنويه إلى أن الدراسة أظهرت وجود العديد من الصعوبات عند تحويل البيانات إلى صيغة قواعد البيانات. بعضها يمكن في طبيعة المعجم العربي نفسه وبعضها الآخر ناتج عن أسلوب عرض المعلومات في المعاجم المطبوعة التي تشكل أساساً معجمياً لإيجاد قواعد بيانات محوسبة.