MUHIT: A Multilingual Harmonized Dictionary

Sameh Alansary
Bibliotheca Alexandrina, ElShatby, Alexandria, Egypt.
Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University
ElShatby, Alexandria, Egypt.
sameh.alansary@bibalex.org

Abstract
This paper discusses a trial to build a multilingual harmonized dictionary that contains more than 40 languages, with special reference to Arabic which represents about 20% of the whole size of the dictionary. This dictionary is called MUHIT which is an interactive multilingual dictionary application. It is a web application that makes it easily accessible to all users. MUHIT is developed within the Universal Networking Language (UNL) framework by the UNDL Foundation, in cooperation with Bibliotheca Alexandrina (BA). This application targets to serve specialists and non-specialists. It provides users with full linguistic description to each lexical item. This free application is useful to many NLP tasks such as multilingual translation and cross-language synonym search. This dictionary is built depending on WordNet and corpus based approaches, in a specially designed linguistic environment called UNL™ that is developed by the UNLD foundation. This dictionary is the first launched application by the UNLD foundation.

Keywords: MUHIT, multilingual lexical database, English WordNet.

1. Introduction
Much effort has been exerted into lexical databases over the years. The effort invested in the study and representation of lexical items to express the underlying difficulties existing in; first, language vagueness and polysemy. Second, language gaps and mismatches. To develop multilingual lexical resources, there are two approaches. 1) Reusing existing resources. 2) Building machine readable dictionaries from scratch. A multilingual lexical database should ideally have a structured set of language-independent meanings, operating as its interlingua and should meet a number of requirements (Hans C. Boas, 2009e). 1) languages will be connected at the level of meanings, and not at the level of words. 2) The number of language pairs in a database grows exponentially, so the meanings of the different languages should not be linked up in pairs, but should be connected via an intermediate set of meanings. 3) Since not all meanings are lexicalized in every language, there will be lexical gap. This problem cannot be solved by forcing every meaning in the interlingua to be expressed in every language to overcome lexical gaps (Maarten Janssen, 2000; Sammer, Soderland, 2007).

This paper presents an attempt to build a multilingual lexical database that includes more than 40 languages within the universal networking language (UNL) framework focusing on the Arabic part explaining its methodology, size, characteristics and types of information it provides.

Section 2 presents what is MUHIT explaining the size of the Arabic share and of other languages. Section 3 represents how MUHIT was developed, illustrating some of the challenges that were faced during the development. Section 4 explains the Arabic linguistic infrastructure of MUHIT. Finally section 6 discusses how to use MUHIT and its search options accompanied by results.

2. What is MUHIT
MUHIT is an ocean of knowledge. It is an abbreviation for (MULTilingual Harmonized dICtionary), however, it is not just an abbreviation, it constitutes a meaningful word. The name "MUHIT" has been inspired by the Arabic word “المحيط” (al-Muhit), which means "Ocean" and "comprehensive". It also involves the ideas of "environment", “ambience” or “surroundings”. The Arabic name of the application may have been inspired by the fact that MUHIT contains more than 2 million Arabic word forms, most of which we owe to the work done at the Library of Alexandria.

MUHIT is a multilingual lexical database produced by the UNDL Foundation within the UNL framework (Uchida, Zhu, Della Senta,1999) (S. Alansary, M. Nagi and N. Adly, 2010), where entries have been interlinked by sense, and natural language word forms have been associated to a uniform concept identifier (UW) (Martins and Avetisyan, 2009), which makes it possible to search for words with the same sense in the same language or in different languages. MUHIT contains more than 10,000,000 word forms collected from more than 40 languages, the number of entries varies from one language to another and is continuously increasing. The Arabic language ranks first in terms of the number of word forms inside MUHIT as it represents about 20% of the whole size; the Arabic share is 2,332,765 word forms. Languages vary in terms of the number of word forms each language includes, some languages include few word like Abkhazian language which has only one word form, lao language has only (30) word forms and Hindi language has (67) word forms. While some other languages have substantially larger number of word forms like English which includes (398,304) word forms, Russian which includes (1,585,693) word forms and Spanish includes (1,374,495) word forms. For more information about languages of MUHIT (http://www.unlweb.net/muhit/index.php?muhit=report).
3. Methodology

This section will present the sources of the Arabic words of MUHIT, and the environment that MUHIT was developed within, focusing on some challenges that have been faced during the development of MUHIT.

3.1. How MUHIT was Developed

This sub-subsection describes the approach adopted in building MUHIT. There are several approaches to build multilingual lexical databases such as Parallel Wordlists approach, Hub-and-Spoke Mode approach, WordNet approach, Acquilex et al. approach and Corpus Based approach (Maarten, 2002). The methodology adopted in developing the lexical database “MUHIT” depends on the combination of WordNet approach and corpus based approach which have been developed through the UNLarium environment. MUHIT was built by integrating two language resources The International Corpus of Arabic (ICA) and the English WordNet 3.0. All languages were required to link their dictionaries with the English WordNet as the first phase and collect 50,000 most frequent natural language words as the second phase. Arabic collected the most frequent 50,000 lexemes from ICA. There were two starting points. The first was from the WordNet and the second was from the natural language in order to increase the richness of MUHIT and grantee that most of the words were covered.

3.1.1. The English WordNet

The English WordNet 3.0 has been used in building MUHIT. The English WordNet is a large lexical database of English developed at the Cognitive Science Laboratory of Princeton University, it is widely used in the fields of computational linguistics and natural language processing (Fellbaum, 1998; Vossen, 1998). In the WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets) each expressing a distinct concept and each assigned a distinct ID number. Synsets and their IDs (accompanied by a gloss, examples and a frequency number) can differentiate between even the slightest nuances of meaning. For example, “bank” have six meanings in the WordNet, they would be represented by six different IDs and six different Arabic words that represent their meanings in Arabic as shown in figure 2. English WordNet contains 117,659 senses. To integrate the English WordNet 3.0 in building MUHIT, the team tried to link every concept “sense” in the English WordNet with a suitable Arabic head word. Consequently, about 117,000 new Arabic words have been introduced to MUHIT.

3.1.2. The International Corpus of Arabic (ICA)

The International Corpus of Arabic (ICA) has been used in building MUHIT, because it contains 100 million words. The collection of samples is of written Modern Standard Arabic selected from a wide range of sources. It is designed to represent a wide cross-section of regional variety of Arabic; it is stimulating the first systematic investigation of the national variety as being used all over the Arab world (Alansary, Nagi and Adly 2006). Corpora have proven to be very useful resources for linguists who believe that their theories and descriptions of Arabic should be based on real, rather than contrived data. A list of most common 50,000 Arabic lexemes have been extracted from the ICA, these lexemes were mapped with our dictionary, about 30,000 new words were added to the MUHIT.

3.2. The UNLarium Environment

MUHIT was developed in The UNLarium environment. The UNLarium is an integrated development environment for producing language resources for natural language processing (NLP). It is mainly a web-based database management system where registered users are able to create, edit and export dictionary entries and grammar

1 http://www.unlweb.net/unlarium/index.php?action=export
2 http://www.unlweb.net
rules according to the UNDL Foundation’s standards for language engineering. Although originally conceived inside the UNL framework, the UNLarium intends not to require any deep knowledge on UNL, and its data may be used in several NLP systems, in addition to UNL-based applications. Furthermore, the system is supposed to be used as a research workplace for exchanging information and testing several linguistic constants that have been proposed for describing and predicting natural language phenomena. One of our main goals is to create a language-independent meta-language that would be as comprehensive, as harmonized and as confluent as required by multilingual processing.

Arabic entries in MUHIT were added through the UNLarium environment by adding the Arabic lemma then choosing its appropriate lexical category, lexical structure and part of speech. By choosing certain lexical category, a number of other linguistic features appear in the entry creation form. If the chosen lexical category is noun the gender, number, inflectional paradigms and Subcategorization frames of nouns will appear. If the chosen lexical category is verb the aspect, transitivity, inflectional paradigms and Subcategorization frames of verbs will appear. If the chosen lexical category is adjective the degree, distribution, inflectional paradigms and Subcategorization frames of adjectives will appear. After adding all linguistic features of the Arabic lemma the entry is added to UNLarium and to MUHIT through Submitting the entry in UNLarium which would be directly reflected in MUHIT. Figure 3 shows the form in the UNLarium environment for creating entries in MUHIT. All information appearing in this form will be discussed in details in section 4.

![Arabic Dictionary](Image)

Figure 3: The entry creation form in The UNLarium environment

In order to ensure the quality of the created entries, users are assigned a profile, which is defined according to several characteristics. They can be promoted or demoted at any time depending on their participation in the project. The initial (default) level is Observer. Permissions are related to the scope of actions, as follows:

- Observers are allowed to browse dictionaries and grammars, navigate the system, but cannot add entries or grammar rules;
- Trainees are allowed to add entries, but only under supervision;
- Authors are allowed to add entries, but may edit only their own data;
- Editors may also edit authors' data, but cannot edit other editors' data;
- Revisers may edit editors' data, but cannot edit other revisers' data;
- Managers may edit any data, create projects and delete entries; and
- Super managers may edit the source code of the system.

In order to avoid problems, every entry or rule is double-checked inside the UNLarium: first by the editor, and then by the reviser. Permissions may be canceled depending on the users’ track history. Authors can be demoted to Observers, if their entries achieve more than 10% of errors.

3.3. The Unified Tagset

The set of features in any dictionary depends on the structure of the natural language and may vary a lot. In order to better standardize lexical resources inside the UNL framework, the UNDL Foundation recommends the adoption of a set of tags for some specific and pervasive grammatical phenomena. This section presents the role of the unified tagset in building MUHIT. This tagset is a set of features and several of the linguistic constants have been already proposed in the Data Category Registry (ISO 12620)^4, they also represent widely accepted linguistic concepts. The purpose of this Tagset is providing the technical means for describing any linguistic behavior which should be done in a highly standardized manner, so that others could easily understand and exploit the data for their own benefit. The main intention is to create a harmonized system in order to make language resources as easily understandable and exchangeable as possible. The linguistic information inside MUHIT is a list of features extracted from the tagset.

The tagset is capable of defining the pervasive morphological, semantic, syntactic and even pragmatic phenomena. Each language can, then, choose the set of values applicable to the phenomena it reflects. For instance, the UNL tagset covers all possible values of countability; singular, dual, trial and quadral, paucal, multal, plural, singular tantum, plural tantum and invariant although the value of “dual”, for example, would never be used in languages such as English or French; however, it will definitely be used in Arabic. The tagset is arranged in a taxonomic hierarchy depicted in a tree that contains list of attributes with their values. For example, tagset includes the attribute “Gender” which has the values; “feminine”, “masculine”, “neuter”, “common” and “variable”. Also, part of speech, number, valency, voice, distribution, etc have lists of values.

---

3 www.undl.org

4 [http://www.isocat.org/](http://www.isocat.org/)
The Tagset was designed in order to be as comprehensive, few, short and mnemonic as possible in order to ensure comprehensibility and consistency. It standardizes both: the tags of the linguistic attributes, and the tags of the values they may assume. Such standardization is crucial in ensuring translatability across the languages participant in the MUHIT. In addition, standardizing the set of tags across the UNL community facilitates understanding and exchanging the available language resources of different languages among the various UNL language centers. The tagset can be reached at http://www.unlweb.net/wiki/Tagset.

3.4. Challenges of Building MUHIT

This section presents some of the linguistic challenges that have been faced during building MUHIT. Some of these challenges are related to the selection of Arabic words and others are related to linking the Arabic words with the English WordNet.

3.4.1. Linking Arabic Words with English WordNet

In the attempt of linking the Arabic words with the English WordNet many challenges have appeared, such as lexical gaps, culture specific words and technical and scientific terms.

- **Lexical gaps**
  Words in the WordNet are classified into different parts of speech according to the English language, however, their Arabic counterparts does not always belong to the same part of speech. For example, the word “summa cum laude” which means (with highest honor; with the highest academic distinction) is classified as an adjective in the WordNet, but the correct Arabic translation is “بامثلة” which is classified as an adverb.

- **Missing appropriate UW**
  Some Arabic words do not have an appropriate UW (ID) to represent its exact meaning; for example the Arabic verb “اسم” which means (saying “In the name of God the Merciful”) has no ID to represent its meaning; also the Arabic noun “الاسم” which means (saying “Crescent in the end of the month”) has no ID to represent its meaning. Such Arabic words were assigned appropriate UW according to the ontology. For example the UW of the Arabic noun "اسم" is crescent(ic>natural object).

- **The UW is a culture-specific words**
  Some UWs represent culture-specific words which do not have Arabic equivalent, because these concepts are not found in the Arabic culture. In this case, we transliterated the name into Arabic. For example, “مودرا” which means “ritual hand movement in Hindu religious dancing”. There were two options to translate this word into Arabic; first, to transliterate it to be “مودرا”. Second to describe its meaning by Arabic words to be "الرقص حركة آيدي في “المودرا”". The decision was to translate it into "مودرا”.

- **Technical term**
  The headword of the UW is a name referring to a specific entity, but translatable (can be translated with Arabic lexical items). In this case the UW is represented by Arabic words, as in “pac-man strategy” with the gloss “the target company defends itself by threatening to take over its acquirer”, it will be represented in Arabic as “استراتيجية باك مان”.

  - **Biological Taxonomy**
    Biological classification, or scientific classification in biology, is a method to group and categorize organisms into groups such as genus or species. Each organism is given a scientific (or Latin) name and occasionally a common name. These types of words consists of two parts; the classifications (Kingdom, Phylum, Class, order, family, Genus, etc) and the scientific name. we translated the classification and added the scientific nameas it is. For example, the word “family Jungermanniaceae” was translated as “فصيلة Jungermanniaceae”.

3.4.2. Selecting Arabic Words

The task of finding free electronic sources in order to obtain lists of Arabic words was very difficult; it took long time and a lot of searching. Even if found, these lists were unusable, because they were always lists of word forms not lemmas which means that morphological analysis was needed.

3.4.3. Named Entities Insertion

It is planned to enrich MUHIT with the most common named entities found in Wikipedia. The insertion of this type of words will make MUHIT more useful for different uses and research studies that require information about named entities such as question answering. They will be fully described linguistically.

4. Arabic Linguistic Infrastructure of MUHIT

The linguistic infrastructure of MUHIT is a set of linguistic information developed to describe every natural language word. In order to better standardize lexical resources inside the UNL framework, the UNDL Foundation recommends the adoption of what so called “Tagset” which is a standard and universal list of features that is required for providing lexical resources. These tags have been proposed to the Data Category Registry (ISO 12620). This list of features is suitable for describing all linguistic phenomena of all languages. The focus here will be on the Arabic language.

4.1. Morphological Information

Morphological information is the information that can describe the morphological behavior of the words such as part of speech, inflections of words, etc. This section presents some morphological information that are assigned to the Arabic words of MUHIT.

4.1.1. Part of speech

The Arabic entries are classified into different classes and each class may include subclasses. These classes are noun, verb, adjective, adposition, adverb, etc. For example, nouns are classified into common nouns and proper nouns, so words like “لاعيف” “disk” is common noun where words like “محمد” “Mohamed” is proper noun. Verbs are also classified into subclasses, full verb, copula verb, and modal verb. For example, the verb “أكل” ‘eat’ is a full verb, “يدا” ‘seem’ is a copula verb, and “كان” ‘was’ is
an auxiliary verb. All the other parts of speech are also classified into subclasses (Alansary, 2012). For more information see (http://www.unlweb.net/wiki/Part_of_speech).

4.1.2. The lexical structure feature
Arabic words are classified into sub word (bound morphemes) such as "س" /sin/ the future prefix, simple words as "ق" /qar/ "read" and multword expressions which are lexical structures made up of a sequence of two or more lexemes such as "جمهورية مصر العربية" "Arab Republic of Egypt". For more information see (http://www.unlweb.net/wiki/Lexical_structure).

4.1.3. The inflectional paradigms
Inflection is the modification of words to express different grammatical categories such as tense, mood, voice, aspect, person, number, gender and case. Conjugation is the inflection of verbs; declension is the inflection of nouns, adjectives and pronouns. In the UNL framework, inflection is indicated by a set of transformations carried over the base form for generating the different word forms. Arabic is one of the highly inflected languages. For example, by assigning a feature such as M121 to the verb "لَعْبَ "he uses", 146 different verb forms will be generated including the forms "لَعِبَ "he uses", "لَعِبَ "both are using", "لَعِبَ "they are using", "لَعِبَ "she used", "لَعِبَ "he uses". Each paradigm contains a list of inflectional rules that are responsible for generating different word forms. Figure 4 shows a sample of the different inflectional forms of the Arabic verb "لَعْبَ "he uses". The Arabic lexical database contains 343 inflectional paradigms, representing 10,566 morphological rule (Alansary, 2012).

| Arabic Word | M121 | M122 | M123 | M124 |
|-------------|------|------|------|------|
| لَعْبَ | 2142 | 2142 | 2142 | 2142 |
| لَعِبَ | 2142 | 2142 | 2142 | 2142 |
| لَعِبَ | 2142 | 2142 | 2142 | 2142 |

Figure 4: All the rules of the paradigm M121 and the generated forms of the verb "لَعْبَ "he uses"

4.2. Morpho-syntactic Information
Morpho-syntactic feature is a feature which is relevant to syntax, means that it is involved in either syntactic agreement or government. Gender, number, and person are involved in agreement in a large number of languages. This section presents some of this information that are assigned to the Arabic words of MUHIT.

4.2.1. Transitivity
It is used to describe the syntactic behavior of the verbs and the type of their arguments. The Arabic lexicon classifies verbs according to transitivity into two main classes, intransitive verbs and transitive verbs. The intransitive verbs are in turn classified into accusative verb whose subject is not the agent, as in the sentence "أَنْفَقْتُ النَّارَ "(Water flow) and unergative verb whose subject is the agent, as in the sentence "عَمِلَ الْوَلَدُ "(the boy walk). Transitive verbs are further classified into four types, direct transitive; a verb which takes a subject and a single direct object, such as the verb "شَربَ الرَّجُلَ "(the man drink water), indirect transitive; a verb which takes a subject and a single indirect object, such as the verb "رَجَبَ الرَّجُلَ "(the man drink the water), ditransitive; a verb which takes a subject and two objects, such as the verb "عَلَّمَ · "(Affirmative). For more information see (http://www.unlweb.net/wiki/Transitivity).

4.2.2. Tense
It is used in the grammatical description of verbs, referring primarily to the way the grammar marks the time at which the action denoted by the verb took place. It can be broadly classified as: past tense as in "كتبت "wrote", present tense as in "كتبت "writes" and future tense as in "كتبت "will write". For more information see (http://www.unlweb.net/wiki/Tense).

4.2.3. Gender
Linguistically, some languages like Arabic has two genders; masculine and feminine, however, Gender of natural language words within the UNL framework is classified into four genders; masculine such as "كرسي" "chair", "رجل" "man", "جدار" "wall", feminine such as "ثَاءُرَ "woman", "نساء" "women", "كَُبَّرِ "girl", "جَرِيْدَ" "newspaper", common such as "ضار" "thing", "حَيْثَ "woman", "دَرْءَ "girl", "مَعْدَة" "model", "لاَيْمَا "nobody" or " بغب "glass". In the case of common and variable, the words may be classified as either masculine or feminine. The difference is that, in common gender, a change of the gender changes the meaning of the natural gender of the reference. For example, "رَجَل "man" and "وَالِدَ "father" or "كَانَ "woman" both mean 'this glass'. Gender attribute is important in generating the different word forms of both adjectives and verbs as in the adjective "مشترطة" "active", the word form "مشط" "active", describes masculine noun and the word form "مشطا" "active", describes feminine noun. Similarly with the verb "كتب "write" and the word form "كتب "he is writing", indicates that the verb agent is a masculine noun and the word form "كتب "she is writing" indicates that verb agent is feminine noun. For more information about gender see (http://www.unlweb.net/wiki/Gender).
4.2.4. Number
The number feature is mainly for describing nouns. Some languages like Arabic classify nouns according to numbers into three classes; singular, plural and dual, other languages like English classify nouns into two classes; singular and plural. The UNL Tagset classifies nouns into three main classes; singular, plural and invariant and each class includes subclasses. For example, the first main class singular includes the subclass “singular tantum” which describes words that are singular but do not have plural form such as “كتاب” “writing”. The second main class plural includes; dual nouns such as “أيام”, “two books”, pronominal nouns which are nouns that refer to few of a class such as “كثير من” “few years”, mutal nouns which are nouns that refer to many of a class such as “أطفال” “children”, and plural tantum which refers to plural nouns that do not have a singular form such as “ذوي” “spouses”. The number attribute is also assigned to verbs to specify the number of their subject and to adjectives to specify the number of the Substantive. For example, the verb “كتب” “he wrote” is assigned as singular to indicate that its subject is singular, and the verb “كتبوا” “they wrote” is assigned plural feature to indicate that its subject is plural and so on forth.

4.2.5. Person
It is a category that defines the deictic reference to a participant in an event, such as the speaker, the addressee or others. It is classified into first person, second person and third person. First person have two subclasses: first person singular as the Arabic word “أنا” ‘I’ (1PS) and first person plural as “نحن” ‘we’ (1PP). Second person have two subclasses: second person singular as in “أنت” ‘you’ (2PS) and second person plural as in “أنتم” ‘you’ (2PP). Third person have two subclasses: third person singular as in "هم" ‘he’ (3PS) and third person plural as in "هم" ‘they’ (3PP). Person attribute is also assigned to verbs to specify the person of the verb subject. For example, the Arabic verb "سمع" ‘hear’ is assigned the person feature (3PS) to specify that the verb subject is third person singular (3PS) and the verb form "سمع" ‘I hear’ is assigned the person feature (1PS) to describes that the verb subject is first person singular (1PS). For more information about person see (http://www.unlweb.net/wiki/Number).

4.3. Syntactic Information
Syntactic information describes the principles and processes by which sentences are constructed. It deals with phrase and sentence formation out of words. This subsection will discuss some syntactic information that are assigned to words in MUHIT such as valency, aspect and subcategorization information.

4.3.1. Valency
Valency or valence is a category that indicates the number of syntactic arguments required by any predicate. Verb valency can be: monovalent such as in the sentence "شوف" ‘show’; "أعطي الرجل جاره هدية" ‘the man gave his neighbor a gift’ and trivalent such as in the sentence " اللغة العربية أظرف الله أعمالهم حسنات عليهم" ‘Thus will Allah show them their deeds as regrets’. In some cases the predicate is considered valent which means that this predicate does not have arguments. For example “beautiful” (adjective) is valent. For more information about valency see (http://www.unlweb.net/wiki/Valency).

4.3.2. Aspect
Grammatical aspect is a feature for verbs which is used to indicate the temporal internal structure of an action, event, or state, from the point of view of the speaker. There are two types of grammatical aspect inceptive(ICP) as in the Arabic sentence "بدأ يشرح الدروس" ‘start to explain the lesson’ and causative (CAU) as in the Arabic sentence "جعله يقبل الأمر" ‘made him accept it’. For more information about aspect see (http://www.unlweb.net/wiki/Aspect).

4.3.3. Subcategorization Frames
They are sets of rules used to generate syntactic structures out of the base form. Subcategorization frames that determine the number and types of the necessary syntactic arguments (specifiers, complements and adjuncts) of the verb. Subcategorization frames that are used in case of valent words whose syntax needs to follow a general rule, i.e., whenever there can be stated a regular pattern for generating constituents linked to the base form, such as specifiers, complements and adjuncts. For example the Arabic sentence "ذهب الرجل إلى الجار" ‘the man gave money to his neighbor’. The Subcategorization frame is:

\[
VS(+NP,+NOM,+APER,+ANUM,+AGEN)VC(+NP,+ACC)VC(PH([لـ]+),+DAT);
\]

![Figure 5: The syntactic tree of the sentence](image)

This subcategorization frame indicates that the verb “جهب” ‘give’ has three arguments; verb specifier (noun phrase), verb complement (noun phrase), another verb complement (adverbial phrase) and the head of this phrase is the preposition "لـ" ‘to’ as shown in figure 5.

4.4. Semantic Information
It focuses on the relation between on the one hand signifiers, like words, phrases, signs, symbols, and on the other hand what they stand for, their denotation.
4.4.1. Semantic classification of the words

The semantic classification adopted in MUHIT is the English WordNet 3.0. ontology. In WordNet, English nouns, verbs, adjectives and adverbs are organized into sets of synonymous words (called synsets), each synset representing one distinct concept. Each entry in this classification carries a set of features and attributes, all subclasses of this concept inherit the properties of that class (Fellbaum, 1998).

4.4.2. Animacy

It is a semantic category assigned to nominal concepts. It indicates human or animal referents. Animacy may assume two possible values: animate (ANM), if the referent is a living object; human or animal, as "مدرس" ‘teacher’ - "رجل" ‘man’ - "قطة" ‘cat’ or inanimate (NANM) which refers to any other non-living referent as "سيارة" ‘car’ - "حرية" ‘boat’ - "حرية" ‘freedom’.

5. Search options and results of MUHIT

The usage of MUHIT could not be easier. The interface is very simple and intuitive. It only consists of a search bar and a help button. The system searches for the string in all existing dictionaries. This search is performed not only for the citation forms of the words but for all existing inflections as well. For more information see (http://www.unlweb.net/muhit/index.php?muhit=help#C).

All this linguistic information appears with the results of using MUHIT multilingual lexical database. MUHIT allows searching by lexeme, word form or wildcard search options, it also allows searching by the concept. The results provide the user with a number of information such as the number of results and number of languages that include the search word. In addition, the results inform about the different parts of speech of the search word if found. Figure 6 shows the different related information of the search word.

(1) The features of the search word such as part of speech, the lexical structure, number, inflectional paradigm and subcategorization frame of the word.
(2) Different inflections of the word such as plural, dual and feminine of nouns. All these inflectional forms are not stored in MUHIT, but generated by the stored inflectional paradigms, see section a.2, c.
(3) All the different possible translations from other languages.
(4) The available synonyms of the word in the same language.

Figures 7, 8 and 9 shows these types of information for the word "تعلم".

Figure 6: Results of Searching by Inflection

Figure 7: Linguistic Features Appears in MUHIT

Figure 8: Different Translations and Synonyms
MUHIT has the possibility to search with the concept and the results provide the number of languages that have translated this concept. For example, searching for the concept “the present time or age” shows that 42 languages have translated this concept and shows the different synonyms in each language. The total number of results that were found for this search is 70 results as shown in figure 10.

6. Conclusion

The paper presents a multilingual lexical database “MUHIT” that has been built within the UNL framework that includes about 40 languages. The linguistic infrastructure of the system has been introduced with special reference to Arabic. MUHIT can be useful for specialists and non-specialists and many applications can be built depending on MUHIT such as multilingual machine translation systems and cross language search systems. The paper has presented the search options of MUHIT and the types of results illustrated with screen shots.

7. References

Hans C. Boas, (2009e). “Recent trends in multilingual computational lexicography,” in: Boas, Hans C. (ed.). Multilingual FrameNets in Computational Lexicography: Methods and Applications. 1–26.

M. Janssen. (2002); “SIMuLLDA : a Multilingual Lexical Database Application using a Structured Interlingua”, Doctoral dissertation, Utrecht University, June, 2002, chapter 1.

Hiroshi Uchida, Meiying Zhu, Tarcisio G. Della Senta. (1999), ”A Gift for a Millennium”, November 1999.

Sameh Alansary. (2012), A UNL based approach for building an Arabic computational lexicon, the 8th international conference on informatics and systems (INFOS 2012) may 14-16, 2012.

Ronaldo Martins, Vahan Avetisyan. (2009), Generative and Enumerative Lexicons in the UNL Framework, Seventh International Conference on Computer Science and Information Technologies (CSIT 2009), 28 September - 2 October, 2009, Yerevan, Armenia Proceedings of CSIT 2009.

Alansary, Sameh, MagdyNagi, NohaAdly. (2010). UNL+3: The Gateway to a Fully Operational UNL System. In Proceedings of Egypt 10th International Conference on Language Engineering, Cairo, Egypt.

C. Fellbaum. (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Sameh Alansary, Magdy Nagi, and Noha Adly. (2007), Building an International Corpus of Arabic (ICA): Progress of Compilation Stage, 7th International Conference on Language Engineering, Cairo, Egypt, December 5 - 6 2007.

M. Agnes and D. B. Guralnik. Webster's New World College Dictionary, IDG Books Worldwide, 4thed, Houghton Mifflin Harcourt, 2001.

J. D. Ullman, J. Widom; 1997. First Course in Database Systems, A, 1/e, Prentice Hall Engineering/Science/Mathematics.

Vossen, P. (ed.). (1998). EuroWordNet: A multilingual database with lexical semantic networks. Kluwer Academic Publishers.

Marcus Sammer and Stephen Soderland. 2007. Building a sense-distinguished multilingual lexicon from monolingual corpora and bilingual lexicons. In Proceedings of Machine Translation Summit XI.