MUHIT: A Multilingual Lexical Database

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*—There is a clear need for dictionaries translating between a large number of languages. The creation of a dictionary of good quality takes a lot of time, and given the fact that 5000-6000 languages yield 25-30 million pairs of languages, it is important to have a database that provides the possibility to translate directly between pairs of languages. The problem is that words are often hard to match across languages: different words from different languages do not have the same range of meanings, not all words from one languages have an equivalent in the other, etc. This paper sheds light on a multilingual database in which most of these problems are solved.**

1 INTRODUCTION

Over the last few decades, a large amount of new lexical resources have arisen: machine readable dictionaries, lexical databases, full-form lexicons, morphological databases, semantic networks, dictionary databases, etc. It is commonly accepted that there are about five to six thousand languages. Many pairs of languages such as X and Y, do not have a dictionary from X to Y or from Y to X, there are only dictionaries for the pairs from X to English/French/Spanish as the source language or dictionaries for English/French/Spanish to Y as the target language. There is a clear need for dictionaries translating between the large number of different languages without the intervention of a small number of Western European languages with a colonial past. Creating a dictionary of good quality exhausts a lot of time and effort. Given the fact that 5000-6000 languages would yield 25-30 million pairs of languages, this makes it important to have a database that provides the possibility to translate directly between pairs of languages. Since creating a bilingual dictionary for every language pair is not a viable option, we need a way to reach the same result as if we had built bilingual dictionaries, we need to have a system that overcomes all the difficulties that prevent the linking of dictionaries. A well-known problem is that words are often hard to match across languages i.e. different words from different languages do not have the same range of meanings, not all words from one language have an equivalent in the other, etc. Moreover, a multilingual lexical database should meet a number of requirements [1].

- Languages will be connected at the level of meanings, and not at the level of words. Words can display ambiguities, they are arbitrary and by no means identical across languages. Hence, words of different languages are never said to be literally translated, but they share a meaning.
- With an increasing number of languages grows exponentially, the meanings of the different languages should not be linked up in pairs, but are to be connected via an intermediate set of meanings. Since using one of the languages as an intermediate structure would make the system dependent on the particularities of that central language, this intermediate structure should be a language-independent, interlingual set of meanings, to avoid a number of undesirable effects.
- Since not all interlingual meanings are lexicalized in every language, there will be lexical gaps. A mechanism should be present to produce translations to overcome such lexical gaps since it would be unproductive and undesirable to resolve this problem by forcing every meaning in the interlingua to be expressed in every language, in order to avoid the existence of lexical gaps.

In this paper, we will be presenting MUHIT (Multilingual Harmonized database) as a solution. Since creating a bilingual dictionary for every language pair is not a viable option. We need a way to reach the same effect without actually doing so. Moreover, creating a bilingual dictionary for every language pair does not seem as a likely solution because the huge number of the bilingual dictionaries would lead to slow processing and bad interaction. We need a system that overcomes all these difficulties. This system is not supposed to be language depended, but it should be concept depended, which means including all the possible concepts in order to enable each language to describe those concepts freely, resulting in the prevention of the overlapping across languages. The main purpose of this paper is to explore an electronic system (Multilingual Harmonized database, MUHIT) that yields bilingual dictionaries for every pair of languages presented within the system. Section 2 discusses the difference between the lexical database and the dictionary database. Section 3 sheds light on MUHIT in details. Section 4 represents the development of MUHIT explaining its methodology. Section 5 represents the linguistic infrastructure of MUHIT. Section 6 discusses how to use the system and shows the different information that users can get

from using MUHIT. Section 7 represents how to participate in updating and enhancing MUHIT. Section 8 represents different applications that can be built depending on MUHIT system and finally section 9 concludes the paper.

2 MULTILINGUAL DATABASE VERSES MULTILINGUAL DICTIONARY?

The purposes of usage of a lexical database are different from those of a dictionary. MUHIT database differs from the design of dictionary databases in a number of things. Since it does not list only lemmas, but complete inflected forms as wellas the amount and type of information stored for each lemma is different [2].

On the one hand, lexicographers define the dictionary as a collection of words in one or more specific languages, often listed alphabetically, with usage information, definitions, etymologies, phonetics, pronunciations, and other information[3]. Multilingual dictionary is designed to link between two languages; one is source and the other is target. On the other hand, the database is also an organized collection of data. However, the data are typically organized to model relevant aspects of reality in a way that supports processes requiring this information. Usually databases have a system or schema which is especially designed for applications that interact with the user [4]. Lexical database can be used in a search engine providing human users with lexical information, and also in NLP applications, computer aided language-learning systems, computer aided linguistic research, etc. In addition, the multilingual lexical databases are set up as to have one source and many target languages (all the languages in the system) [2].Multilingual dictionaries are likely to be designed by a pivot language or meta-language (most of them is English) which is used as a bridge to cross over to any other language, That would permit many drawbacks to occurs like lexical gaps [1].

3 WHAT IS MUHIT?

"MUHIT" is an abbreviation for (MUltilingual Harmonized dIcTionary) but 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". Moreover, it is part of one of the most celebrated Arabic dictionaries (al-Qamus al-Muhit), compiled by al-Firuzabadi (1329–1414) that has been widely used for centuries [5].

MUHIT is a multilingual electronic lexical database which has been developed within the universal networking language (UNL) framework [6], [7], [8] and it is one of the UNDL Foundation [9] products in cooperation with Bibliotheca Alexandrina. MUHIT is available on the UNL[lab](http://www.unlweb.net/lab/) where entries have been interlinked by sense, and natural language word forms have been associated to a uniform concept identifier (UWs), the words of UNL [7], [10]. MUHIT contains more than 10,000,000 word forms collected from more than 40 languages, table1 shows these languages and the number of word forms from each language.

TABLE 1: THE EXACT FIGURES OF THE CURRENT VERSION OF THE SYSTEM

Language	Word Forms
Abkhazian	1
Afrikaans	12,008
Arabic	2,333,305
Armenian	1,700,166
Assamese	232
Baatonum	4,116
Bengali	9,183
Bulgarian	58,194
Chinese	66,652
Croatian	87,821
Dutch	3,067
English	398,304
Estonian	148,929
French	938,576
Georgian	6,436
German	78,442
Greek (Modern)	13,950
Gujarati	360
Hindi	1,694
Hungarian	180,159
Indonesian	9,626
Italian	170,890
Japanese	8,748
Kannada	3,337
Khmer	4,481
Lao	30
Latin	1,368,012
Malay	15,275
Nepali	3,360
Panjabi	2,630
Persian	9,098
Polish	1,502
Portuguese	368,776
Romanian	5,063
Russian	1,611,082
Serbian	84,935
Sinhala	2,104
Slovak	3,452
Slovenian	467,601
Spanish	1,372,813
Swahili	2,287
Swedish	5,411
Tamil	10,839
Telugu	39,512
Thai	5,383
Turkish	5,578
Ukrainian	191,704
Vietnamese	9,871
TOTAL	**11,824,995**

As shown from table1, the number of entries varies from one language to another and is continuously increasing. Languages vary in terms of the number of word forms each language includes, some languages include few word forms 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. Arabic language ranks first inside MUHIT as it represents about 20% of the whole size; the Arabic share is 2,332,765 word forms (http://www.unlweb.net/muhit/index.php?muhit=report).MUHIT also contains rare languages not found even in other multilingual lexical databases. For example languages as Baatonum, Panjabi, Nepali and Sinhala are not found in famous multilingual dictionary like Google.

MUHIT was developed mainly for cross-language word search. This means that MUHIT can help users in finding and using information in their native or non–native languages. This is clear from the design of MUHIT.

4 HOW IS MUHIT DEVELOPED?

As mentioned in section 2, MUHIT contains more than 10,000,000 word forms in more than 40 languages. All entries have been introduced through various projects in the UNL[arium] (http://www.unlweb.net/unlarium/), The UNL[arium] is the UNDL Foundation's language resources management system. It is a web-based integrated development environment for creating and editing language resources for natural language engineering especially related (but not limited) to the UNL framework. Some of these projects are WordNet based and some others are corpus based. This section introduces the different projects integrated in building MUHIT lexical database, the aim of each project and the achievements of each project in different languages. In addition, this section will also present the participants of building MUHIT.

A. Methodology

This 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 [1].The methodology adopted in developing the computational lexicon "MUHIT" depends on the combination of WordNet approach and corpus based approach through many projects which have been developed through the UNL[arium] environment. To develop multilingual lexical resources there are two methods. First method depends on reusing existing resources. Second method depends on building machine readable dictionaries from scratch [10]. MUHIT is a combination of both methods.

B. Projects

This subsection presents the different projects that MUHIT is composed of, illustrating the aim and the progress of each project. BRUNO, MIR, LPP, LIS, LEWIS & SHORT will be discussed. Each project was developed for a certain aim but all their lexicon were integrated to form MUHIT lexicon. For more details about each project and other projects see (http://www.unlweb.net/wiki/Projects).

1) BRUNO

The project BRUNO (Basic Resources for UNLizatiOn) is devoted to the creation of NL-UNL (analysis) dictionaries. BRUNO project has two goals. First, is to provide several word-to-concept monolingual databases (i.e., encoding or reader's dictionaries). Second, is to find concepts that are not enclosed in the WordNet3.0 and should be incorporated to the UNL Dictionary. BRUNO is language dependent project. Every language has its own set of entries to be addressed. It includes more than **50,000** lemmas to be addressed in universal concepts. Table2 shows the progress of BRUNO project among different languages illustrating that Arabic ranks as first in terms participation.

TABLE 2: BRUNO PROGRESS REPORT

Language	Target	Current	Accomplished
Arabic	51,937	51,937	100.00%
Armenian	601	535	89.02%
Bulgarian	2,242	2,032	90.63%
Chinese	4,746	4,185	88.18%
Hungarian	1,925	1,816	94.34%
Kannada	700	192	27.43%
Khmer	303	198	65.35%
Malay	4,175	4,175	100.00%
Panjabi	998	250	25.05%
Slovenian	2,050	2,045	99.76%
Telugu	2,057	42	2.04%
Thai	1,975	1,369	69.32%
Ukrainian	2,258	1,964	86.98%

2) MIR

MIR is a centralized repository of lexical data extracted from the WordNet3.0. It contains 117,659 UWs representing different sets of synonyms (or synsets) of English to be associated to the corresponding lexical items of any language. MIR project is divided into two phases, MIR1 consists of 27,241 concepts and MIR2 consists of 90,373 concepts. MIR project has

two main goals. First, is to provide a concept-to-word multilingual database. Second, is to assign a degree of universality to each of the senses registered in the WordNet3.0. Tables 3shows the progress of MIR1 and table 4 shows the progress of MIR2among different languages.

TABLE 3: MIR1 PROGRESS REPORT

Language	Target	Current	Accomplished
Afrikaans	27,241	2,864	10.51%
Arabic	27,241	27,087	99.43%
Armenian	27,241	27,158	99.70%
Assamese	27,241	139	0.51%
Azerbaijani	27,241	120	0.44%
Baatonum	27,241	3,078	11.30%
Bengali	27,241	1,625	5.97%
Bulgarian	27,241	3,979	14.61%
Chinese	27,241	14,532	53.35%
Croatian	27,241	3,278	12.03%
Dutch	27,241	1,628	5.98%
English	27,241	27,241	100.00%
Estonian	27,241	3,116	11.44%
French	27,241	27,201	99.85%
Georgian	27,241	1,490	5.47%
German	27,241	26,645	97.81%
Greek (Modern)	27,241	3,168	11.63%
Gujarati	27,241	120	0.44%
Hindi	27,241	868	3.19%
Hungarian	27,241	4,020	14.76%
Indonesian	27,241	995	3.65%
Italian	27,241	3,414	12.53%
Japanese	27,241	1,669	6.13%
Kannada	27,241	1,078	3.96%
Khmer	27,241	1,334	4.90%
Lao	27,241	30	0.11%
Latin	27,241	7,646	28.07%
Malay	27,241	5,282	19.39%
Nepali	27,241	1,699	6.24%
Panjabi	27,241	725	2.66%
Persian	27,241	3,049	11.19%
Polish	27,241	1,148	4.21%
Portuguese	27,241	27,004	99.13%
Romanian	27,241	1,287	4.72%
Russian	27,241	27,191	99.82%
Serbian	27,241	2,849	10.46%
Sinhala	27,241	1,325	4.86%
Slovak	27,241	1,282	4.71%
Slovenian	27,241	8,468	31.09%
Spanish	27,241	27,229	99.96%
Swahili	27,241	1,659	6.09%
Tamil	27,241	3,062	11.24%
Telugu	27,241	2,115	7.76%
Thai	27,241	2,843	10.44%
Turkish	27,241	1,891	6.94%
Ukrainian	27,241	6,228	22.86%
Vietnamese	27,241	1,699	6.24%

TABLE 4: MIR2PROGRESS REPORT

Language	Target	Current	Accomplished
Afrikaans	90,373	1,897	2.10%
Arabic	90,373	90,164	99.77%
Armenian	90,373	466	0.52%
Baatonum	90,373	170	0.19%
Bengali	90,373	157	0.17%
Bulgarian	90,373	2,064	2.28%
Chinese	90,373	3,998	4.42%
Croatian	90,373	1,986	2.20%
Dutch	90,373	152	0.17%
English	90,373	90,356	99.98%
Estonian	90,373	1,587	1.76%
French	90,373	4,898	5.42%
German	90,373	1,799	1.99%
Greek (Modern)	90,373	1,999	2.21%
Hungarian	90,373	2,199	2.43%
Indonesian	90,373	90	0.10%
Italian	90,373	1,687	1.87%
Japanese	90,373	263	0.29%
Khmer	90,373	634	0.70%
Latin	90,373	5,301	5.87%
Malay	90,373	3,093	3.42%
Nepali	90,373	1,657	1.83%
Persian	90,373	1,238	1.37%
Portuguese	90,373	3,819	4.23%
Romanian	90,373	395	0.44%
Russian	90,373	2,310	2.56%
Serbian	90,373	1,288	1.43%
Slovak	90,373	127	0.14%
Slovenian	90,373	2,528	2.80%
Spanish	90,373	1,786	1.98%
Tamil	90,373	1,935	2.14%
Telugu	90,373	1,771	1.96%
Thai	90,373	1,919	2.12%
Turkish	90,373	151	0.17%
Ukrainian	90,373	2,380	2.63%
Vietnamese	90,373	924	1.02%

3) LPP Dictionary

The Project Le Petit Prince (LPP) aims at building a dictionary for translating Le Petit Prince novel; published by Antoine de Saint-Exupéry in 1943,to the Universal Networking Language (UNL). The project LPP has two main goals: to set standards and guidelines for human UNLization and to test several tools that have been developed at the UNDL Foundation. The total number of word of LPP is 1,832.Table5 shows the progress of building LPP dictionary among different languages.

TABLE 5: LPP PROJECT PROGRESS REPORT

Language	Target	Current	Accomplished
Afrikaans	1,832	1,023	55.84%
Arabic	1,832	1,828	99.78%
Armenian	1,832	1,831	99.95%
Assamese	1,832	97	5.29%
Baatonum	1,832	1,230	67.14%
Bengali	1,832	1,782	97.27%
Bulgarian	1,832	1,764	96.29%
Chinese	1,832	1,279	69.81%
Croatian	1,832	1,828	99.78%
Dutch	1,832	1,780	97.16%
English	1,832	1,832	100.00%
Estonian	1,832	1,831	99.95%
French	1,832	1,828	99.78%
German	1,832	1,747	95.36%
Greek (Modern)	1,832	1,832	100.00%
Gujarati	1,832	120	6.55%
Hindi	1,832	40	2.18%
Hungarian	1,832	826	45.09%
Indonesian	1,832	1,085	59.22%
Italian	1,832	1,823	99.51%
Japanese	1,832	340	18.56%
Kannada	1,832	63	3.44%
Khmer	1,832	188	10.26%
Lao	1,832	21	1.15%
Latin	1,832	1,104	60.26%
Malay	1,832	798	43.56%
Nepali	1,832	248	13.54%
Panjabi	1,832	126	6.88%
Persian	1,832	1,676	91.48%
Polish	1,832	146	7.97%
Portuguese	1,832	1,739	94.92%
Romanian	1,832	187	10.21%
Russian	1,832	1,828	99.78%
Serbian	1,832	1,639	89.47%
Sinhala	1,832	83	4.53%
Slovak	1,832	1,409	76.91%
Slovenian	1,832	1,830	99.89%
Spanish	1,832	1,678	91.59%
Swahili	1,832	191	10.43%
Tamil	1,832	1,673	91.32%
Telugu	1,832	300	16.38%
Thai	1,832	544	29.69%
Turkish	1,832	1,788	97.60%
Ukrainian	1,832	1,832	100.00%
Vietnamese	1,832	206	11.24%

4) LIS Dictionary

This project aims at building a dictionary for the Library Information System (LIS) which is an information retrieval system that aims at performing multilingual search over bibliographical metadata. The main goal of the library information system is to UNLize a small set of MARC21[1]records and to provide the resources necessary to generate it into at least five different languages other than Arabic. The project has been developed by the UNL Center at the Library of Alexandria. Table 6 shows the progress of LIS dictionary among different languages.

TABLE 6: LIS DICTIONARY PROGRESS REPORT

Language	Target	Current	Accomplished
Arabic	2,861	2,689	93.99%
Armenian	2,861	2,844	99.41%
Assamese	2,861	2	0.07%
English	2,861	1	0.03%
French	2,861	2,708	94.65%
Greek (Modern)	2,861	69	2.41%
Hungarian	2,861	25	0.87%
Italian	2,861	2,776	97.03%
Latin	2,861	25	0.87%
Portuguese	2,861	50	1.75%
Russian	2,861	2,736	95.63%
Spanish	2,861	2,734	95.56%
Tamil	2,861	5	0.17%
Ukrainian	2,861	110	3.84%

[1]http://www.loc.gov/marc/specifications/

5) LEWIS & SHORT

The project Lewis & Short Latin Dictionary aims at mapping Latin lemmas extracted from the Lewis & Short Latin Dictionary (1879) into UNL. The lemmas were extracted from the online version of "A Latin Dictionary" available at the Perseus Digital Library. This project is coordinated by the UNL Language Center at the University of Patras, in Greece.

C. Participants

After reviewing the projects that constitute MUHIT, this subsection will present how these projects were developed and the participants; who participate in developing these projects through UNL[arium] environment. Participants are: 1) Language-Centers (LCs), which are responsible for managing the linguistic resources (dictionaries, grammars, and corpora) created for a given language. There are many language centers such as the Arabic language center in Bibliotheca Alexandrina, in Alexandria, Egypt (http://www.bibalex.org/unl/Frontend/home.aspx), The Russian Language Center (http://www.unl.ru/), and the Spanish Language Center(http://www.vai.dia.fi.upm.es/ing/projects/unl/index.htm).2) Volunteers and freelancers participate in any project as fully independent and self-determining contributors and are not committed to any goal, timetable, schedule, deadline or obligation other than complying with the system specifications.

5 MUHIT LINGUISTIC INFRASTRUCTURE

As section 3 reviewed the projects and the participants of MUHIT, it is important to explain the linguistic infrastructure of MUHIT. Linguistic knowledge that appears in MUHIT has been assigned to all words through UNL[arium] encompassing different linguistic levels: morphological information, morpho-syntactic information, syntactic information and semantic information. UNL uses a standard and universal list of features (Tagset) to describe all types of the linguistic information concerning every natural language word in MUHIT. This section will present the linguistic information that constitutes the linguistic infrastructure of MUHIT provided with different examples.

A. Tagset[2]

This subsection presents how natural language words are described in MUHIT using a list of features extracted from the UNDL Foundation Tagset. This Tagset is a set of features in a UNL dictionary depending on the structure of the natural language. However, in order to boost the standardization of the lexical resources used in the UNL framework, the UNDL Foundation recommends adopting the following tags for some specific and pervasive grammatical phenomena.

Several of those linguistic constants have been already proposed in the Data Category Registry (ISO 12620)[3], and 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 simple features and a list of inflection rules.

B. Morphological Information

Morphological information is the structure of words. Many languages have rich morphology where a single word can function as an entire sentence in another language. For example, the Arabic word /fajaʿalnāhum/"فجعلناهم" can be translated into the English sentence 'and We made them'. The set of morphological information that MUHIT uses can handle these kinds of phenomena in different languages. This subsection will discuss some of this morphological information such as part of speech, Lexical structure and Inflections of words.

1) Part of speech feature

It is used to classify words into main classes and each class may include subclasses. The classes are nouns, verbs, adjective, adposition, adverb, affix, classifier, conjunction, determiner, interjection, numeral, particle and pronoun. The system is designed as such in order to create much flexibility in describing the different types of words. Moreover, the classes are divided into subclasses. For example, MUHIT differentiates between two types of nouns, common noun such as "صندوق" 'box' - "باب" 'door' – "ورقة" 'paper' and proper noun as "نجيب محفوظ" 'Naguib Mahfouz' - "مصر" 'Egypt' - "اليونسكو" 'UNESCO'. Pronouns are also included with a detailed description as a part of the noun class. It is divided into demonstrative pronoun such as "هذه" 'this' - "هذان" 'these', possessive pronoun as the "نا" 'our', "ي" 'my' in words like "قلمنا"

[2] For more information see: http://www.unlweb.net/wiki/Tagset
[3] http://media.dwds.de/clarin/userguide/text/concepts_ISOcat.xhtml

‘our pen’ - “قلمي” ‘my pen’ , relative pronoun such as “الذي” ‘who’ - “التي” ‘who’ - “الذين” ‘whose’, reflexive pronoun such as “هو” ‘he’ - “هي” ‘she’ and interrogative pronoun as “من” ‘who’ - “ماذا” ‘what’ - “كيف” ‘how’.

Verbs are also classified into subclasses, full verb, copula verb, and modal verb. Full verb is the main verb which is the head of a verbal phrase with full meaning. In Arabic, full verb is the most widely used type of verbs. For example, in the sentence "أكل الولد الطعام" ‘the boy ate the food’ the verb "أكل" ‘eat’ is a “full verb” as it expresses the meaning of the “eating” action. Copula is a word used to link the subject of a sentence with a predicate (a subject complement). In English, verb "to be" that conveys the meaning of “having the quality of being” is an example of copula. The Arabic verb "بدا" ‘seem’ as in "الغرفة تبدو فارغة" ‘the room seems empty’ indicates the meaning of “being” therefore acting as a copula. Auxiliary verb is a verb that gives further semantic or syntactic information about a main or full verb. It is also called helping verb, helper verb or auxiliary verb, for example the Arabic verb "كان" in the sentence "كان الولد يلعب بالكرة عندما رأيته" ‘The boy was playing ball when I saw him’ indicates that the main verb "لعب" ‘play’ occurs in the past. Modal auxiliary verb is a type of auxiliary verb that is used to indicate modality. Modal auxiliary verbs give more information about the function of the main verb that follows it and is used to indicate the attitudes on the part of the speaker towards the factual content of the utterance, e.g. uncertainty, definiteness, vagueness, possibility. Typically, it is manifested by the grammatical category of mood. For example, the Arabic verb "استطاع" ’able’ acts as a modal auxiliary verb in the sentence "استطاع الجندي أن يحمي وطنه" ‘The soldier was able to protect his country’, as it gives more information about the function of the main verb "حمى" ‘protect’, it conveys the “ability” of doing the action.

Adverbs are also classified into four different subclasses. Specifier adverb which is the adverb that specifies an adjective, a verb or another adverb such as “جدا” ‘very’ - “أيضا” ‘also’. Adjunct adverb which is the adverb that qualifies an adjective, a verb or another adverb by indicating manner, time, place or frequency such as “بإخلاص” ‘Sincerely’ - “بشكل دائم” ‘Permanently’ - “عمليا” ‘Practically’. Conjunct adverb which is considered as a connecting adverb that does not add information to the sentence and is not considered as a part of the propositional content (or at least not essential), it only connects the sentence with previous parts of the discourse such as “بالرغم من ذلك” ‘In spite of this’ - “بالإضافة إلى” ‘In addition to’ - “في الواقع” ‘In fact’. Disjunct adverb which expresses information that is not considered essential to the sentence it appears in, but which is considered to be the speaker's or writer's attitude towards, or descriptive statement of, the propositional content of the sentence such as “لسوء الحظ” ‘Unfortunately’ - “بصفة شخصية” ‘personally’.

Affixes are also classified into four types; prefixes such as the Arabic definite article “ال” ‘the’ as in “الولد” ‘the boy’ and the Arabic present tense prefix “يت،” as in “يشرب” ‘he is drinking’ or “تشرب” ‘she is drinking’. Suffixes such as Arabic feminine suffix “ة” as in “معلمة” ‘female teacher’ and the Arabic plural masculine suffix “ون” as in “معلمون” ‘male teachers’ and English present suffix “s” as in “boys”. Infixes which appear within the stem, they are very rare in English, such as "ma" in "sophistimacated" and such as adding “و” to the middle of the word “درع” ‘Shield’ to generate the plural form“دروع” ‘Shields’. Finally, Circumfixes which appears at the front and at the back of the stem, they are very rare in English, such as "a", "ed" in "ascattered".

Conjunctions are also classified into three different types; coordinating conjunctions such as “و” ‘and’ - “أو” ‘or’, Subordinating conjunctions such as “بينما” ‘While’, Correlative conjunctions that are pairs of conjunctions that work together to coordinate two items such as “كلا” ‘both’, and Complementizers that are special subordinating conjunctions that introduce complement clauses such as “أن” ‘that’.

Numerals are also classified into six types which are: cardinal numbers which describe quantity such as “ثلاث” ‘three’, Ordinal numbers which describe position such as “أول” ‘first’, Partitive numbers which describe division such as “نصف” ‘half’ - “ثلثي” ‘third’, Multiplicative numbers which describe repetition such as “مرة واحدة” ‘once’ - “مرتان” ‘twice’, Collective numbers which describe groups such as “ثلاثي” ‘tripartite’ - “خماسي” ‘fivefold’, and distributive numbers which describe distributions such as “في أزواج” ‘In pairs’.

2) Lexical structure feature

The second morphological information is the lexical structure. It is used to classify the words into subword (bound morphemes) such as, in English affixes ("un-", "re-", "-ful", "-ness") and roots that do not occur alone ("rupt" in "interrupt", "disrupt", "corrupt", "rupture", etc),and in Arabic "س" /sin/ the future prefix. Simple words as the Arabic words “قرأ” ‘read’ - “مكتب” ‘office’ - “رائع” ‘wonderful’. Multiword expressions which are lexical structures made up of a sequence of two or more lexemes such as the English words " darkroom", " blue-green", " get over" and the Arabic words “جمهورية مصر العربية” ‘Arab Republic of Egypt’ - “محفوف بالمخاطر” ‘perilous’ - “بطريقة غير مباشرة” ‘Indirectly’.

3) 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. For example, Arabic is one of the highly inflected languages [11].For example, by assigning a feature such as M121 to the verb “استخدم” ‘use’, 146 different verb forms will be generated including the forms "يستخدم" ‘he uses’ - "استخدما" ‘both used’ - "يستخدمان" ‘both are using - "يستخدمون" ‘they are using’ - "استخدمت" ‘she used’ - "تستخدم" ‘she is using’ - "استخدمتا" ‘both (feminine) used’ - “يستخدمن” ‘they (feminine) are using’ – "استخدمي" ‘you use (imperative, feminine)’. Each paradigm contains a list of inflectional rules that are responsible for generating different word forms. Figure 1 shows a sample of the different inflectional forms of the Arabic verb "استخدم".

PAS&3PS&SNG&MCL&ACV=استخدم	PRS&NOM&1PP&ACV&PLR=نستخدم	FUT&3PS&SNG&MCL&ACC=سيستخدم
PSV&PAS&SNG&3PS&MCL=استخدم	PRS&NOM&1PP&PSV&PLR=نستخدم	FUT&3PP&MCL&DUA&NOM=سيستخدمان
PRS&NOM&3PS&SNG&MCL&ACV=يستخدم	PRS&3PS&SNG&MCL&ACV&ACC=يستخدم	FUT&3PP&MCL&DUA&ACC=سيستخدما
PSV&PRS&NOM&SNG&3PS&MCL=يستخدم	PRS&3PS&SNG&MCL&PSV&ACC=يستخدم	FUT&3PP&MCL&PLR&NOM=سيستخدمون
PAS&3PP&MCL&DUA&ACV=استخدما	PRS&3PP&DUA&MCL&ACV&ACC=يستخدما	FUT&3PP&MCL&PLR&ACC=سيستخدموا
PSV&PAS&3PP&MCL&DUA=استخدما	PRS&3PP&DUA&MCL&PSV&ACC=يستخدما	FUT&3PS&SNG&FEM&NOM=ستستخدم
PRS&NOM&3PP&MCL&DUA&ACV=يستخدمان	PRS&3PP&PLR&MCL&ACV&ACC=يستخدموا	FUT&3PS&SNG&FEM&ACC=ستستخدم
PSV&PRS&NOM&3PP&MCL&DUA=يستخدمان	PRS&3PP&PLR&MCL&PSV&ACC=يستخدموا	FUT&3PP&FEM&DUA&NOM=ستستخدمان
PAS&3PP&MCL&PLR&ACV=استخدموا	PRS&3PS&SNG&FEM&ACV&ACC=تستخدم	FUT&3PP&FEM&DUA&ACC=ستستخدما
PSV&PAS&3PP&MCL&PLR=استخدموا	PRS&3PS&SNG&FEM&PSV&ACC=تستخدم	FUT&3PP&FEM&PLR&NOM=سيستخدمن
PRS&NOM&3PP&MCL&PLR&ACV=يستخدمون	PRS&3PP&DUA&FEM&ACV&ACC=تستخدما	FUT&3PP&FEM&PLR&ACC=سيستخدمن
PSV&PRS&NOM&3PP&MCL&PLR=يستخدمن	PRS&3PP&DUA&FEM&PSV&ACC=تستخدما	FUT&2PS&MCL&SNG&NOM=ستستخدم
PAS&3PS&SNG&FEM&ACV=استخدمت	PRS&3PP&PLR&FEM&ACV&ACC=يستخدمن	FUT&2PS&MCL&SNG&ACC=ستستخدم
PSV&PAS&3PS&FEM&SNG=استخدمت	PRS&3PP&PLR&FEM&PSV&ACC=يستخدمن	FUT&2PP&DUA&MCL&NOM=ستستخدمان
PRS&NOM&3PS&SNG&FEM&ACV=تستخدم	PRS&2PS&SNG&MCL&ACV&ACC=تستخدم	FUT&2PP&DUA&MCL&ACC=ستستخدما
PSV&PRS&NOM&3PS&FEM&SNG=تستخدم	PRS&2PS&SNG&MCL&PSV&ACC=تستخدم	FUT&2PP&MCL&PLR&NOM=ستستخدمون
PAS&3PP&FEM&ACV&DUA=استخدمتا	PRS&2PP&DUA&MCL&ACV&ACC=تستخدما	FUT&2PP&MCL&PLR&ACC=ستستخدموا
PSV&PAS&3PP&FEM&DUA=استخدمتا	PRS&2PP&DUA&MCL&PSV&ACC=تستخدما	FUT&2PS&FEM&SNG&NOM=ستستخدمين
PRS&NOM&3PP&FEM&ACV&DUA=تستخدمان	PRS&2PP&PLR&MCL&ACV&ACC=تستخدموا	FUT&2PS&FEM&SNG&ACC=ستستخدمي
PSV&PRS&NOM&3PP&FEM&DUA=تستخدمان	PRS&2PP&PLR&MCL&PSV&ACC=تستخدموا	FUT&2PP&DUA&FEM&NOM=ستستخدمان
PAS&3PP&FEM&ACV&PLR=استخدمن	PRS&2PS&SNG&FEM&ACV&ACC=تستخدمي	FUT&2PP&DUA&FEM&ACC=ستستخدما
PSV&PAS&3PP&FEM&PLR=استخدمن	PRS&2PS&SNG&FEM&PSV&ACC=تستخدمي	FUT&2PP&FEM&PLR&NOM=ستستخدمن
PRS&NOM&3PP&FEM&ACV&PLR=يستخدمن	PRS&2PP&DUA&FEM&ACV&ACC=تستخدما	FUT&2PP&FEM&PLR&ACC=ستستخدمن
PAS&2PS&MCL&ACV&SNG=استخدمت	PRS&2PP&DUA&FEM&PSV&ACC=تستخدما	FUT&1PS&SNG&NOM=سأستخدم
PSV&PAS&2PS&MCL&SNG=استخدمت	PRS&2PP&PLR&FEM&ACV&ACC=تستخدمن	FUT&1PS&SNG&ACC=سأستخدم
PRS&NOM&2PS&MCL&ACV&SNG=تستخدم	PRS&2PP&PLR&FEM&PSV&ACC=تستخدمن	FUT&1PP&PLR&NOM=سنستخدم
PSV&PRS&NOM&2PS&MCL&SNG=تستخدم	PRS&1PS&SNG&ACV&ACC=أستخدم	FUT&1PP&PLR&ACC=سنستخدم
IMP&2PS&MCL&ACV&SNG=استخدم	PRS&1PS&SNG&PSV&ACC=أستخدم	
PAS&2PP&DUA&MCL&ACV=استخدمتا	PRS&1PP&PLR&ACV&ACC=نستخدم	
PSV&PAS&2PP&DUA&MCL=استخدمتا	PRS&1PP&PLR&PSV&ACC=تستخدم	
PRS&NOM&2PP&DUA&MCL&ACV=تستخدمان	PRS&3PS&SNG&MCL&ACV&JUS=يستخدم	
PSV&PRS&NOM&2PP&DUA&MCL=تستخدمان	PRS&3PS&SNG&MCL&PSV&JUS=يستخدم	
IMP&2PP&DUA&MCL&ACV=استخدما	PRS&3PP&DUA&MCL&ACV&JUS=يستخدما	

Figure 1: All the rules of the paradigm M121 and the generated forms of the verb “استخدم”

Inflectional paradigms are also used to generate different word forms of nouns as in the noun “ناشر” ‘publisher’, 12 different word forms will be generated including the forms "ناشر" ‘male publisher’ - "ناشرة" ‘female publisher’ - "ناشران" ‘two male publishers’ - "ناشرتان" ‘two female publishers’ -"ناشرون" ‘male publishers’ - "ناشرات" ‘female publishers’. Each paradigm contains a list of inflectional rules that are responsible for generating different word forms. Figure 2 shows a sample of the different inflectional rules stored in the inflectional paradigm M532 which handles the regular plural in Arabic as in the word "ناشر" which generates the different word forms in different numbers, gender and with different cases.

(1) SNG&MCL:=0>"";	ناشر
(2) FEM&SNG:=0>";"ة	ناشرة
(3) MCL&DUA&NOM:=0>";"ان	ناشران
(4) MCL&DUA&ACC:=0>";"ين	ناشرين
(5) MCL&DUA&GNT:=0>";"ين	ناشرين
(6) FEM&DUA&NOM:=0>";"تان	ناشرتان
(7) FEM&DUA&ACC:=0>";"تين	ناشرتين
(8) FEM&DUA&GNT:=0>";"تين	ناشرتين
(9) MCL&PLR&NOM:=0>";"ون	ناشرون
(10) MCL&PLR&ACC:=0>";"ين	ناشرين
(11) MCL&PLR&GNT:=0>";"ين	ناشرين
(12) FEM&PLR:=0>";"ات	ناشرات

Figure 2: Paradigm M532 and the generated forms of the noun “ناشر”

Other inflectional paradigms are responsible of generating different broken plural patterns as shown in figure 3, the noun “رصيف” ‘pavement’ which is assigned to M551 to generate the plural form “أرصفة” ‘pavements’ - “صاحب” ‘owner’ which is

assigned to M556 to generate the plural form "أصحاب" 'owners' - "باب" 'door' which is assigned to M558 to generate the plural form "أبواب" 'doors' – "أرض" 'land' which is assigned to M584 to generate the plural form "أراضي" 'lands' – "رئيس" 'president' which is assigned to M619 to generate the plural form "رؤساء" 'presidents' and etc.

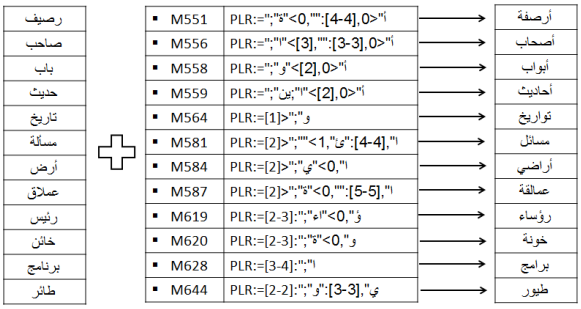

Figure 3: Different paradigms which generate different broken plural patterns

Inflectional paradigms are also used to generate different word forms of adjectives as in the adjective "اجتماعي" 'social', 12 different word forms will be generated including the forms "اجتماعي" 'describing masculine noun' - "اجتماعية" 'describing feminine noun' - "اجتماعيان" 'describing dual masculine nominative noun' - "اجتماعيتان" 'describing dual feminine nominative noun' - "اجتماعيون" 'describing plural masculine nominative animate noun' and "اجتماعيات" 'describing plural feminine nominative animate noun'. Each paradigm contains a list of inflectional rules that are responsible for generating the different word forms. Figure 4 shows a sample of the different inflectional rules stored in the inflectional paradigm M466 which handles adjective with a regular plural pattern in Arabic as in the word "اجتماعي" which generates the different word forms describing adjectives in different numbers, gender and with different cases.

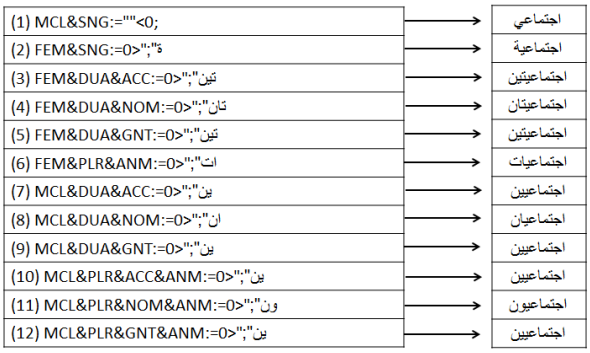

Figure 4: Paradigm M466 and the generated forms of the adjective "اجتماعي"

C. Morpho-syntactic information

Morpho-syntactic information is concerning with the grammatical categories and linguistic units that have both morphological and syntactic properties. Gender, number, person and many other features are involved in the grammatical agreement in a large number of languages, therefore they are typical morpho-syntactic features.

1) Transitivity

It is a feature of verbs which indicates the number of objects a verb requires or takes in a given instance. The transitivity of a verb can be basically classified into intransitive (NTST) and transitive (TSTD). Intransitive verbs are further classified into two types: unaccusative verb whose subject is not the agent, as the verb "تدفق" 'flow' in the Arabic sentence "تدفق الماء" 'Water flowed', the verb "fall" in the English sentence "John fell", and unergative verb whose subject is the agent, as the verb "مشى" 'walk' in the Arabic sentence "مشى الولد" 'the boy walked', the verb "run" in the English sentence "John ran". Transitive verbs are further classified into four types: direct transitive; a verb which takes a subject and a single direct object as the verb "شرب" 'drink' in the Arabic sentence "شرب الرجل الماء" 'the man drank the water', indirect transitive; a verb which takes a subject and a single indirect object, as the verb "رحب" 'welcome' in the Arabic sentence "رحب الرجل بضيوفه" 'the man welcomed his guests', ditransitive; a verb which takes a subject and two objects as in the verb "أمر" 'order' the Arabic sentence "أمر الله الناس بالحق" 'Allah ordered people the truth', or tritransitive; a verb which takes a subject and three objects as

the verb “أرى” ‘show’ in the Arabic sentence "كذلك يريهم الله أعمالهم حسرات عليهم" ‘Thus will Allah show them their deeds as regrets’. Besides intransitive and transitive verbs, there are verbs without transitivity such as copula verbs, for example, the verb “بدى” ‘sound’ in the Arabic sentence "الأمر يبدو خطير" ‘The matter sounds serious’. For more information about transitivity see (http://www.unlweb.net/wiki/Transitivity).

2) 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 “طاولة” ‘table’ - “بنت” ‘girl’ - “جريدة” ‘newspaper’, common such as “ضحية” ‘victim’ - “موديل” ‘model’ - “نكرة” ‘nobody’ and variable such as “كأس” ‘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 implies a change of the natural gender of the reference. For example, “رجل ضحية” ‘victim = man’ and “امرأة ضحية” ‘victim = woman’. whereas, in variable gender, a change of the gender does not affect the reference, we can say “كأس هذا” or “كأس هذه” both equal ‘this glass’. Gender attribute is important in generating the different word forms of both adjectives and verbs as in the adjective “نشيط” ‘active’, the word form “ نشيط"‘he active’ describes masculine noun and the word form "نشيطة"‘she active’ describes feminine noun. Similarly with the verb “كتب” ‘write’ the word form “ يكتب"‘he is writing’ indicates that the verb agent is a masculine noun and the word form "تكتب"‘she is writing’ indicates that verb agent is feminine noun. For more information about gender see (http://www.unlweb.net/wiki/Gender).

3) 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 “كتابة” ‘writing’. The second main class plural includes; dual nouns such as “كتابان” ‘two books’, paucal nouns which are nouns that refer to few of a class such as “بضع” ‘few’ as in “بضع سنين” ‘few years’, multal nouns which are nouns that refer to many of a class such as “كثير من” ‘Many of’, plural nouns such as “أطفال” ‘children’, and plural tantum which refers to plural nouns that do not have a singular form such as “توابل” ‘spices’. 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 “أكلوا” ‘they ate’ takes plural feature to indicate that its subject is plural and so on so forth. As for adjectives, “جميل" ‘describes singular masculine noun’ and "جميلان" ‘describes dual masculine noun’. For more information about number see (http://www.unlweb.net/wiki/Number).

4) 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 "هو" ‘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"أسمع" ‘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/Person).

5) Tense

It is a category 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. Tense can be broadly classified into two main categories; absolute tense and relative tense. Each category is classified to subclasses such as past tense as the verb “كتب” ‘write’ in the sentence “ كتب الولد الدرس” ‘the boy wrote the lesson’ (PAS), present tense as in “يكتب الولد الدرس” ‘the boy writes the lesson’ (PRS) and future tense as in “سيكتب الولد الدرس” ‘the boy will write the lesson’ (FUT). Relative tense is also classified into subclasses such as relative past (RPT), relative nonpast (NRPT), relative future (RFT), etc. For more information about tens see (http://www.unlweb.net/wiki/Tense).

6) Case

It is a grammatical category that indicates the grammatical function of a word such as the role of subject, of direct object, or of possessor in a phrase or clause. It can be broadly classified as: accusative case (ACC) which describes direct object of a verb, for example “ولدين” ‘two boys’ is (ACC) in “قابل الرجل الولدين” ‘the man met the two boys’. Generative case(GNT) which describes possessor, for example “ولدين” ‘two boys’ is (GNT) in “الولدين كتاب” ‘the book of the two boys’. Nominative case (NOM)which describes subject of a finite verb, for example, “ولدان” ‘two boys’ is (NOM) in “ يلعب الولدان الكرة” ‘the two boys play football’. Case attribute is important in generating different word forms of both adjectives and verbs as in the adjective “صغير” ‘little’ the word form “صغيران"describes noun in nominative case and the word form "صغيرين" describes noun in generative case. With regards to the verb “قال” ‘say’ the word form “يقولان" ‘they are saying’ describes the verb in the present tense and nominative case while the word form "يقولا" ‘they are saying’ describes the verb in the present tense and accusative case. For more information about case see (http://www.unlweb.net/wiki/Case).

7) Voice

It is a feature of verbs which describes the relationship between the action that the verb expresses and the participants identified by its arguments. The voice attribute can be classified into active voice, middle voice and passive voice. Not all languages have these three types. For example, Arabic classifies verbs into only two types; active and passive, for example the verb "باع" ‘sell’ in the sentence "باع الرجل البضاعة" ‘the man sold the goods’ is used in the active voice (ACV) while in the sentence "بيعت البضاعة" ‘goods were sold’ the verb is used in the passive voice (PSV). For more information about voice see (http://www.unlweb.net/wiki/Voice).

8) Mood

It is one of a set of distinctive verb forms that are used to signal modality. Imperative mood is a grammatical mood that forms commands or requests. For example, the Arabic verb "باع" ‘sell’ in the sentence "بع البضاعة" ‘sell the goods’ is used in imperative mood (IMP). Jussive mood is a grammatical mood that is used in negation; in negative imperatives, and in the hortative. For example, the Arabic verb "باع" ‘sell’ in the sentence "يبيع البضاعة" ‘he sells the goods’ can be used in jussive mood (JUS) as in the sentence "لم يبع البضاعة" ‘he didn’t sell the goods’. For more information about mood see (http://www.unlweb.net/wiki/Mood).

D. 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.

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 "مشى الولد"‘the boy walked’, divalent such as in the sentence "كتب الولد الدرس"‘the boy wrote the lesson’, trivalent such as in the sentence "أعطي الرجل جاره هدية" ‘the man gave his neighbor a gift’ and tetravalent such as in the sentence "كذلك يريهم الله أعمالهم حسرات عليهم"‘Thus will Allah show them their deeds as regrets’. In some cases the predicate is considered avalent which means that this predicate does not have arguments. For example “beautiful” (adjective) is avalent. For more information about valency see (http://www.unlweb.net/wiki/Valency).

2) Aspect

Grammatical aspects 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 الدرس" ‘start to explain the lesson’ and causative (CAU)as in the Arabic sentence"جعلهيتقبل الأمر"‘made him accept it’. For more information about aspect see (http://www.unlweb.net/wiki/Aspect).

3) Subcategorization frames

They are sets of rules used to generate syntactic structures out of the base form. Subcategorization frames determine the number and types of the necessary syntactic arguments (specifiers, complements and adjuncts) of the verb. Subcategorization frames 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 "وهب الرجل المال لجاره"‘the man gave money to his neighbor’, The Subcategorization frame is:

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

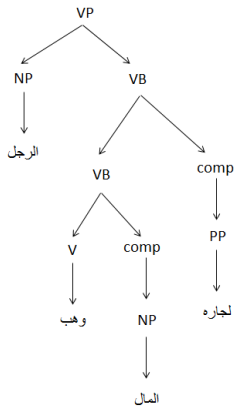

Figure 5: The syntactic tree of the sentence

indicates that the verb "وهب" '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 "ل" 'to'. As shown in figure 5.

E. Semantic information

The semantic information focuses on the relation between signifiers like words, phrases, signs, symbols and what they stand for; their denotation. This subsection will discuss some semantic information as animacy and semantic classification of words.

1) 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' - "مركب" 'boat' - "حرية"'freedom'.

2) Semantic classification of the words

MUHIT utilizes a semantic ontology. This ontology classifies the entities existing in the natural world into a semantic hierarchy. This hierarchy points out the particular type of each concept and the kind of relation it holds with other concepts in the ontology. Each entry in this hierarchy carries a set of features and attributes and all subclasses of this concept inherit the properties of that class. Ontologies are useful in NLP as they play a crucial role in the disambiguation of word senses as well as the understanding of a natural language text by determining the exact sense of a word via its position in the semantic hierarchy. CYC, WordNet and Sensus are examples of ontologies that have been used for language understanding[12].

The semantic ontology 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. For example, the words "coast", "seacoast", "sea-coast" and "seashore" are all synonyms grouped together in a single synset that refers to a unique cognitive concept which is "the shore of a sea or ocean". For example, nouns in the WordNet hierarchy are divided into several semantic fields each having a "unique beginner" as the starting node. A unique beginner is a semantic entity that probably has no hypernym and from which nouns that belong to this distinct semantic field can be pulled out. The WordNet employs a set of 25 unique beginners, 8 of which refer to tangible things or "entities", 5 denote "abstractions" and 3 are "psychological features".Verbs, modifiers and adverbs are also classified into distinct semantic hierarchies[13].

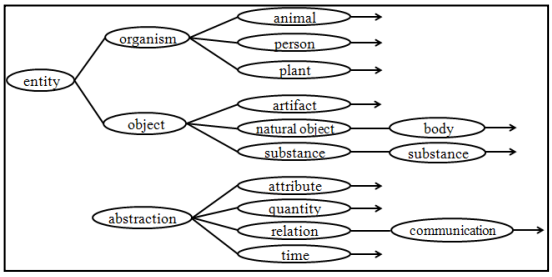

Figure 6: WordNet semantic ontology

6 HOW TO USE THE SYSTEM?

This section will present a detailed demonstration of how to use MUHIT. It is available on the UNL[lab] and it is accessible through the websitehttp://www.unlweb.net/lab/. Being an online application provides the user with a number of advantages; data is stored remotely hence requiring no disk space from the part of the user, no installation or updating is required, and most importantly providing an easy access through the internet. Moreover, the user is not required to have an account in the UNL to access this application.MUHIT is a free and open source application.

MUHIT is auser friendly system, especially designed to provide the user with the utmost convenient design, featuring only two icons for facilitation of use as shown in figure7. One is a help icon, which is found on the left side of the searching bar, upon clicking a list explaining everything about MUHIT is shown, including how it is used. The other is the search icon which is found on the right side of the search box.

Figure 7: MUHIT interface

A. What are the search options?

MUHIT provides a varity of search options for more comrehensive results. A distinct advantage that the system provides is that there is no need to choose a specific language ahead, either as a source language or as a target language, rather the system searches for the string in all existing dictionaries belonging to the different participating languagesas shown in figure 8.

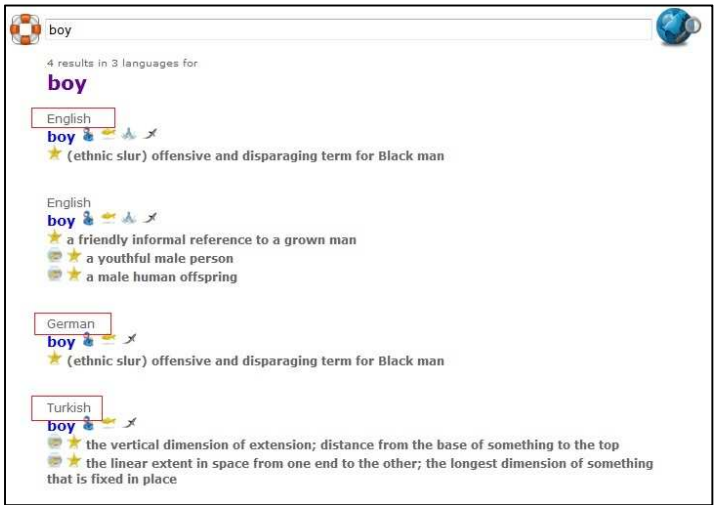

Figure 8: Results of regular search in English

Hence, the search does not only include base forms, which are the typical headwords of most dictionaries, but all existing inflected forms as well as shown in figure 9. This means that, in order to find cross-language synonyms for "eat", you may type "eats", "ate", "eaten", "eating", or any other form and still get the results. Therefore, Presenting a bigger chance for

finding the intended search results especially for users that are unfamiliar with the source language since it is difficult sometimes to detect the base form especially from irregular inflections.

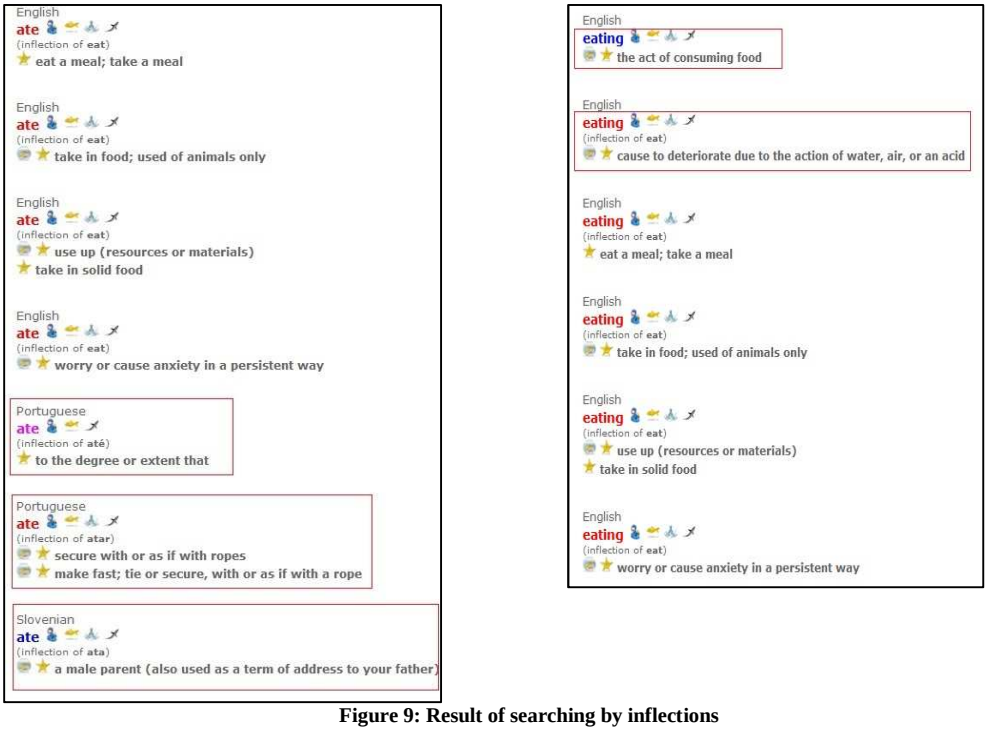

Figure 9: Result of searching by inflections

Along typical search option and searching by inflections, MUHIT has adopted the regular expression system to provide a concise and flexible means for matching strings of text, such as particular characters, words, or patterns of characters. The adopted regular expressions follow the PCRE library (http://www.pcre.org/).

Two of the applied regular expressions for wildcards are (_) and (%). While (_) is used for one single character, (%) is used for any number of characters even zero. For instance,x_z will return any string beginning with x and ending with z with one character in between. For instance if the search is conducted with the letter “b” , “y” and “_” in the middle the results will be the words “bly” from Afrikaans and “bay”, “boy” and “buy” from English as shown in figure 10.

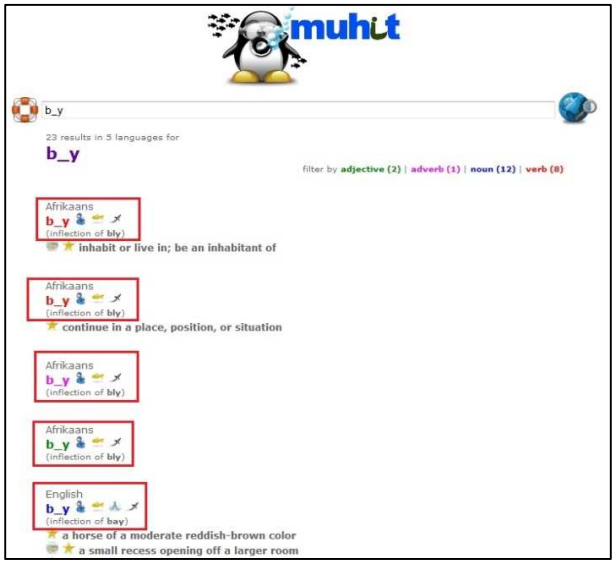

Figure 10: Results of searching by wildcard (_)

The input “x%z” will return any string beginning with x and ending with z with any characters in between. For instance if the search is conducted with the letter “b” , “y” and “%” in the middle the results will be any word that begins with “b” , ends with “y” and contains any number of characters in between. Figure 11 shows that the input “b%y” results includes the words “bakkery”, “Barry”, “battery” and “bely” from Afrikaans language.

Figure 11: Results of searching by wildcard (%)

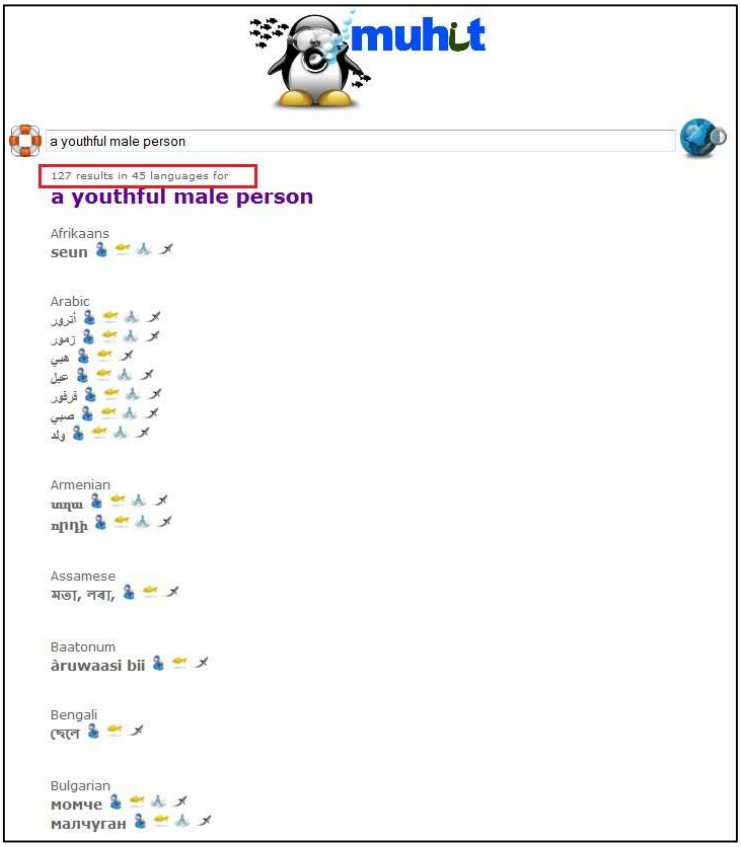

Figure 12: Results of searching by concept

Figure 12 shows the possibility to search with the concept. The results provide the number of languages that have translated this concept. For example figure 12 shows that 45 languages have translated the concept “a youthful male person” and it also shows the different synonyms in each language.

B. MUHIT results

As mentioned before MUHIT is designed with the intention of creating an organized enviroment in order to facilitate the task of searching. Thus, each result appearing for the search word is accompained with either three or four icons on its left. Each icon is responsible for providing the user with certain information. The information displayedarethe author of the entry, The features of the entry, such as part of speech, number, gender etc., The inflections of the entry, if any. For each sense of the entry, it is also provided: The set of synonyms in the same language, The set of synonyms in different languages.

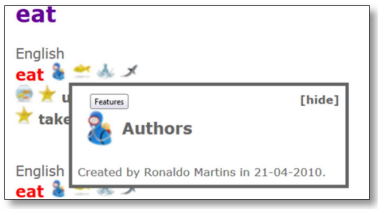

Figure 13: The author and date of creation

The first icon shows the author of the entry and the date of adding it as shown in figure 13.The second icon shows the lingustic description of the entry; the lexical category, the part of speech, the lexical structure, more linguistic information depending on the part of speech, the number of the inflectional paradigm and the subcategorization frame if exists as shown in figure 14.

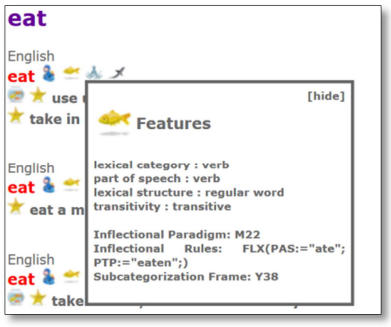

Figure 14: The lexical description of the entry

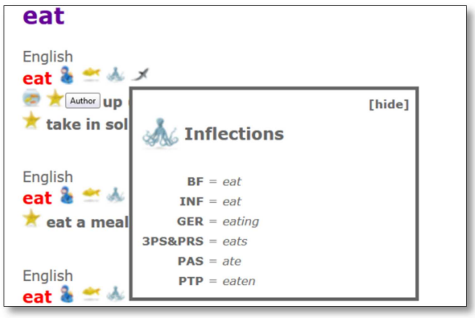

Figure 15: The inflected forms of the search word

The third icon shows the infelected forms of the search word as shown in figure 15.The fourth icon is the report problem icon to report any error about the existing entries as shown in figure 16, in attempt of participation from the part of the user in updating the system.

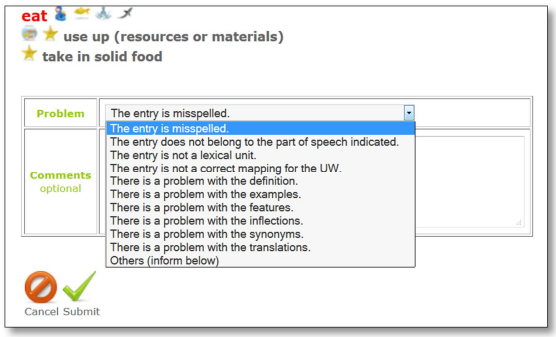

Figure 16: The content of the report problem icon

Besides all the linguistic information provided for each search word, two icons appear on the right side of each single sense of each search word. The first is responsible for displaying the synonyms of the word. The sets of synonyms are also browsable, which makes it possible to navigate further in order to find correspondences for related words in the same language and in different languages as well as shown in figure 17. The second icon shows the avilable different translations for each sense in different languages, which provides a high degree of accuarcy since each translation correspond to a particular senseas shown in figure 18.

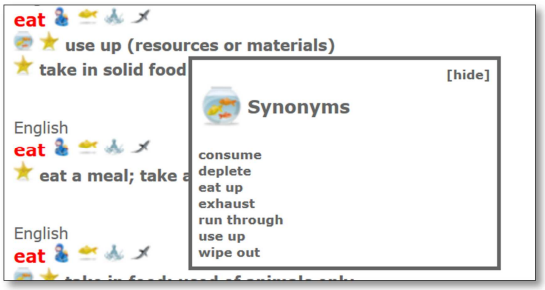

Figure 17: The synonyms of the search word

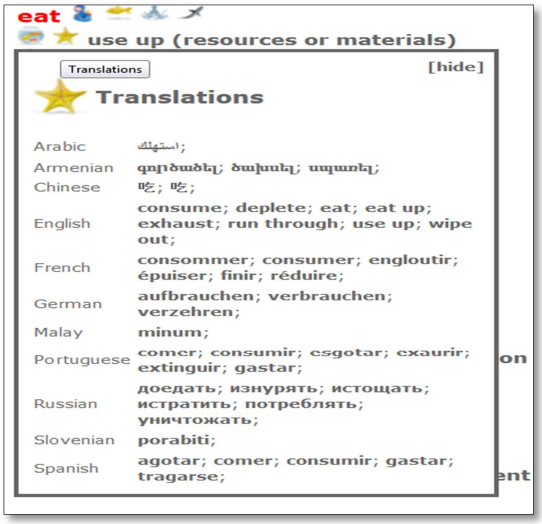

Figure 18: The available translations of the search word

All the options mentioned above have combined both the advanges of a monolingual dictionary and a multilingual one. On the one hand, it provids a detailed linguistic description, that is usually found in learners dictionary only, for each result.On the other hand, it provides a translation for each single sense.

MUHIT is a multilinguil system in which the search word is matched with every existing string belonging to all participating languages, for more detailed information the system shows the number of results, for example searching for the string “b%y” resulted 803 different lexical entries as shown in figure 19. Also the results show how many languages have this string (b%y),it was found in 14 languages,and the number of each part of speech. Each part of speech is displayed in different colour to make it easier in recognizing the categorization;adjective is green, adverb is pink, nouns is blue ans verb is red as shown in figure 19.

Figure 19: The number of results and languages

7 UPDATING MUHIT

Dictionaries constitute a dynamic and everlasting effort of describing the repertoire of natural languages. The UNDL Foundation has been investing a lot in creating resources for UNL-based projects but there is always much more to be done in order to include every language and every word. MUHIT is an open intative. Hence, it is designed in a way that makes the participation easily accessible. The participation is done on different levels.

Since UNL intends to be diversity-preservative, which means that we intend to go from Swahili to Tamil without using English or any other language other than UNL itself as a pivot language, it has provided an easy access to increase the number of the participating languages. In order to add a new language there are two options. The user may join the UNLweb[4], the UNDL Foundation language resources management system, and help in creating dictionaries and grammars;

[4] The UNLweb: http://www.unlweb.net/unlweb/

or he may sponsor a given language, by making a donation to the UNDL Foundation to help in increasing and extending our natural language resources, which are actually available to anyone, not only to the UNL community.

Besides, the user could also participate on another level. The user is welcomed to improve the quality of the resources. If a user is confronted with a result that he doesnot find satisifying he may contribute to the dictionary by reporting a problem which is easily done just by clicking over the report problem icon  which appears at the right side of each entryas mentioned in section 6-B. Furthermore, if something is missing in an existing language, the user could join any of the existing projects, in that way he will be able to contribute and upgrade the dictionary.In order join a project, users have to be approved in VALERIE (http://www.unlweb.net/valerie/), the Virtual Learning Environment for UNL. Non-accredited users have access to several facilities of the UNL^{arium}, but are not allowed to add entries in any project.

8 APPLICATIONS

Such a multilingual lexical database could serve as strong starting point for many applications. This section presents some of the applications that can be developed depending on MUHIT, benefiting from its characteristics and features.

The main application that can be built using MUHIT is the cross language search. Cross–language search aims at facilitating information access across languages. It is built upon more than fifty years of research and development in machine translation and more than ten years of research in cross–language information retrieval (CLIR) Cross–language search integrates CLIR to provide the full function of finding information in languages different from the users' queries.Using MUHIT as a multilingual lexical database would constitute an extraordinary resource for cross–language information retrieval or cross–language word search. MUHIT is intended mainly for cross-language word search. This means that MUHIT can help users to find and use information in their native or non–native languages.

Machine Translation is difficult not only because each language differs from the other(even those from the same family) both structurally and lexically, but also because natural language is ambiguous(again, both structurally and lexically) and is always evolving. Due to the linguistic phenomena and the differences across languages, merging bilingual lexicons into a single multilingual repository is non-trivial. A lexical resource that aims at providing multilingual translation equivalents must be well-designed to address these issues [14]. MUHIT can also be used in multilingual machine translation systems. These systems make it possible to translate words and sentences from and to a large number of languages, containing a lot of information about languages, their grammar and other information which can be useful for anyone who has to translate texts or is just interested in languages.

9 CONCLUSION

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

REFERENCES

[1] M. Janssen,"SIMuLLDA : a Multilingual Lexical Database Application using a Structured Interlingua", Doctoral dissertation,Utrecht University, June, 2002.

[2] M. Janssen. Lexical vs. Dictionary Databases: design choices of the MorDebe system. Papers in Computational Lexicography - COMPLEX, Budapest, Hungary, 2005.

[3] M. Agnes andD.B.Guralnik. Webster's New World College Dictionary, IDG Books Worldwide, 4thed, Houghton Mifflin Harcourt, 2001.

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

[5] The UNLweb website: http://www.unlweb.net/muhit/index.php?muhit=help, (accessed in October 2013).

[6] H. Uchida, M. Zhu, T. G. Della Senta, "A Gift for a Millennium", November 1999.

[7] S. Alansary, M.Nagi, N.Adly, UNL+3: The Gateway to a Fully Operational UNL System. In Proceedings of 10th International Conference on Language Engineering, Cairo, Egypt, 2010.

[8] J. Cardeñosa, A. Gelbukh, E. Tovar (eds.): Universal Networking Language: advances in theory and applications.(Research on Computer Science, 12). Mexico City: National Polytechnic Institute. 443pp, 2005.

[9] The UNDL Foundation website: www.undl.org

[10] R. Martins, V.Avetisyan, "Generative and Enumerative Lexicons in the UNL Framework,"in Proc. Of Seventh International Conference on Computer Science and Information Technologies (CSIT 2009), 28 September - 2 October, 2009, Yerevan, Armenia Proceedings of CSIT 2009.

[11] S. Alansary, “A UNL based approach for building an Arabic computational lexicon,” in Proc. Of the 8th international conference on informatics and systems (INFOS 2012), Cairo, Egypt, may, 2012.
[12] M. Obitko. Ontologies description and applications, Research Report No. 126/01 Czech Technical University, Praue, 2001.
[13] C. Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
[14] L. T. Lim, “Multilingual Lexicons for Machine Translation,”in Proc. of The Eleventh International Conference on Information Integration and Web-based Applications and Services, Kuala Lumpur, Malaysia, 2009.