The Universal Networking Language in Action in English-Arabic Machine Translation

Sameh Alansary[†‡] Sameh.alansary@bibalex.org
Magdy Nagi[††‡] magdy.nagi@bibalex.org
Noha Adly[††‡] noha.adly@bibalex.org

[‡] Bibliotheca Alexandrina, P.O. Box 138, 21526, El Shatby, Alexandria, Egypt.

[†] Department of Phonetics and Linguistics
Faculty of Arts
Alexandria University
El Shatby, Alexandria, Egypt.

[††] Computer and System Engineering Dept.
Faculty of Engineering
Alexandria University,
Egypt.

Abstract

This paper presents an interlingua approach to the machine translation of lengthy documents. This approach is based on encoding the source text in the form of universal semantic networks, using the Universal Networking Language, UNL interlingua, which can then be decoded back into any natural language. This UNL technology has been applied to 1000 pages from the Encyclopedia of Life support Systems (EOLSS) in order to be translated into the six official languages of the UNESCO. This paper summarizes the overall strategy adopted with a focus on the decoding of UNL documents into Arabic.

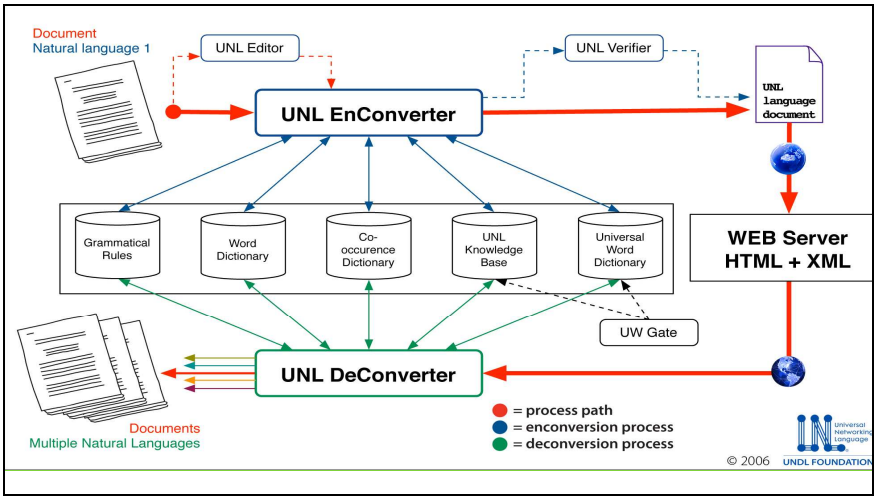
1 Introduction

There are several approaches to Machine translation (MT). Statistical machine translation has been the most widely used approach so far according to a recent survey (Lopez, 2008). A different approach is the interlingua approach, similar to UNL, which relies on transforming the source language into a language-independent representation, which can then be transformed into the target language. When multilingual translation is of interest, the interlingua approach allows building a system of Natural languages with concurrent efforts rather than the compound efforts a statistical approach would require. The challenge with the interlingua approach is designing a language-independent intermediate representation that captures the semantic structures of all languages while remaining unambiguous. Section 2 of this paper will briefly explain the semantic-based approach to machine translation used in multilingual document processing, through that, the interlingua called Universal Networking Language (UNL) will be introduced. Section 3 will briefly compare UNL with other interlinguas. Section 4 will discuss how HTML documents can be translated into different languages using UNL. Section 5 will examine the structure of the Arabic-UNL dictionary through which the Arabic language can be encoded and decoded. Section 6 presents the decoding of the UNL representation into Arabic. Section 7 presents an illustrative example of an Arabic translated sentence, and the results of the evaluation of the translation output. Finally, section 8 concludes the paper.

2 The UNL System

UNL is an electronic language that enables rewriting internet articles, written in various languages, in UNL format in order to be translated into any other Natural Language. The architecture of the UNL system (Figure 1) comprises three sets of components (Uchida, 1996, 2002, 2005):
1. *Linguistic components*: these include dictionaries with Universal Words (UWs) playing the role of UNL vocabulary, grammatical rules that are responsible for transforming Natural Languages into UNL expressions that include the relations and attributes constituting UNL syntax and a different set of grammatical rules responsible for producing well-formed sentences in the target Natural Language, and a Knowledge Base which is a hierarchical representation of the universal concepts found in natural languages;

2. *Software components*: these are two software programs for converting Natural Language texts into UNL expressions (the EnConverter), and vice versa (the DeConverter). The EnConverter is a language-independent parser that provides a framework for morphological, syntactic and semantic analysis synchronously. It is designed to perform the task of converting Natural Language into UNL format; i.e. UNL expressions. The DeConverter, on the other hand, is a language-independent generator that provides a framework for morphological and syntactic generation, and word selection for the sake of natural collocation synchronously. The DeConverter can deconvert UNL expressions into a variety of native languages, using the Word Dictionary, formalized linguistic rules and the Co-occurrence Dictionary of the respective language;
3. System interface components: these are protocols and tools that enable the flow of UNL documents over the World Wide Web.



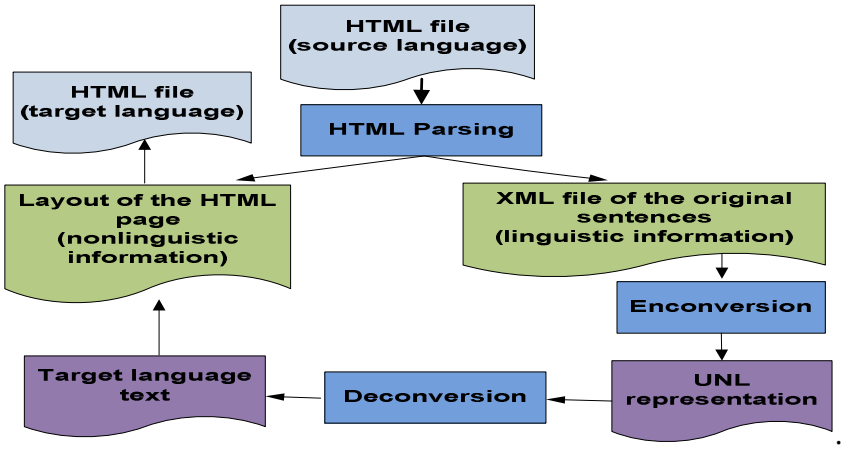
"Figure 1: The core architecture of the UNL system"

3 UNL interlingua VS. other interlinguas

The principle of interlingua is not new. The first ideas about interlingua machine translation appeared in the 17th century. Early interlingua MT systems were also built at Stanford in the 1970s. Many researchers worked in this field; such as Soudi et al (2002), Shaalan (2006), and others. Many interlingua machine translation systems have been developed recently. KANT, for example, is a Knowledge-based, accurate Natural Language Translation system (Nyberg and Mitamura 1992, 1994). UNITRAN is another system, an implementation of a principle-based approach to Natural Language translation (Dorr 1987). It might be noted that interlingua is designed to work in a specific domain and that it is quite difficult, or even impossible, to extend it to a wider domain.
Universal Networking Language (UNL) is an artificial language that can be used as a pivot language in interlingua-based machine translation systems. At first glance, the UNL seems to be a multilingual machine translation system; i.e., a kind of interlingua, into which source texts are converted before being translated into the target languages. It can, in fact, be used for such a purpose, and very efficiently too. However, its real strength is to represent knowledge, and its primary objective is to serve as an infrastructure for handling knowledge that already exists or can exist in any given language. UNL can express all kinds of information and knowledge that is conveyed by Natural Languages. Unlike other interlinguas, UNL is language- independent, universal and not limited to any specific domain; it operates on any language and any type of document.

4 Translating EOLSS using UNL

EOLSS provides a useful body of knowledge which should reach all peoples in their languages and in a way that fits their cultural backgrounds. UNL serves this purpose; it can do both, reproduce EOLSS knowledge in peoples' native languages, and enables them to explore it according to their cultural backgrounds. The UNL task is to make the entire EOLSS available in multiple languages starting with the six official languages of UNESCO. This task involves a two-step process: the first step is enconverting (encoding) the content of EOLSS from English into UNL (UNLization process); and the second is deconverting (decoding) EOLSS content from UNL into natural languages. The enconversion process will be carried out under the responsibility of the UNDL Foundation, while the deconversion into the various natural languages will be carried out by the UNL language center(s) of the respective language.
The EOLSS documents are HTML documents (http://www.undl.org/unldoc/EOLSS/E2-25-01_TXT.aspx), thus, they contain two types of information; linguistic (the text itself) and embedded non-linguistic information such as the font style (bold, italic or underlined) and size, hyperlinks, symbols and equations. In order to begin the UNLization process, natural language sentences have to be extracted first; an HTML parser has been built for this purpose. This parser generates two files: the first file contains the layout of the HTML page while the second file is an XML file that contains the original sentences present in the HTML file. The second file will be fed to the Enconverter to be transformed into UNL representation which, in turn, will be converted into a natural target language. The decoded text will then be combined with the layout file extracted by means of the HTML parser to yield the target language in the format of the source document. The whole process is outlined in Figure 2.



"Figure 2: HTML document parser"

4.1 Transforming the Content of EOLSS from English into UNL.

In order to Enconvert the English input into a universal representation in UNL format, the EnConverter scans the input sentence from left to right to perform two main stages: the first stage is automatically extracting the concepts that represent the input; while the second stage has to do with linking these concepts (nodes) together using semantic relations to build the so-called *hyper-semantic network* (Alansary, 2006b). However, in order to avoid formal ambiguities that can result in misinterpreting concepts, a semi-automatic Enconversion is adopted. This semi-automatic process starts by converting English texts into lists of nodes (morphemes) using an English Morphological analysis tool, checking the node lists by English speakers, then enconverting the node lists into UNL expressions using the English EnConverter, and finally verifying the UNL expressions using the UNL verifier and the English DeConverter. Figure (3) represents an example of an enconverted sentence[1].

[1] EOLSS- human interaction with land and water

```
[S:23]
{org}
The basic water resource is the precipitation supporting vegetation, society, and ecosystems with water.
{/org}
{unl}
aoj(precipitation(icl>process):0Y.@entry.@def,          resource(icl>functional thing):0I.@topic.@def)
agt(support(icl>maintain(agt>thing,obj>thing)):1C.@progress,  precipitation(icl>process):0Y.@entry.@def)
obj(support(icl>maintain(agt>thing,obj>thing)):1C.@progress, :01)
and:01(ecosystem:2C.@entry.@pl,      society(icl>group):1Z)
aoj:01(with(icl>having(aoj>thing,obj>thing)):2N,         ecosystem:2C.@entry.@pl)
obj:01(with(icl>having(aoj>thing,obj>thing)):2N,         water(icl>liquid):2S)
and:01(society(icl>group):1Z,    vegetation(icl>plant):1N)
mod(resource(icl>functional thing):0I.@topic.@def,     basic(mod<thing):06)
mod(resource(icl>functional thing):0I.@topic.@def,     water(icl>liquid):0C)
{/unl}
[/S]
```

"Figure 3: an example of an enconverted sentence"

5 The Technical Design of the UNL-Arabic Dictionary

The UNL-Arabic dictionary stores the meanings of words (the concepts that languages express and the context in which they would be found), the Arabic word headings and the linguistic information that will guide the encoding and decoding processes. We will focus here on word headings and linguistic information.

5.1 Word headings

In building the Arabic dictionary, entries were inserted in the form of stems to avoid adding all the possible inflectional and derivational paradigms of each lexical item. In fact, this decision has been made to suit the approach we have adopted in dealing with Arabic words, both linguistically and computationally; In other words, our design of the Head Word is based on the form needed to fulfill natural language processing tasks (Alansary 2003).

Discussing how each type of Arabic words is dealt with inside our dictionary requires a detailed generative and analytical study on Arabic language which is not the aim of this paper[2] and would be irrelevant. Instead and in order to convey a general and accurate idea about the structure of the Arabic lexical entry, this section will deal with two examples from nouns and verbs. The noun example is selected from the class of nouns that ends in "Hamza" since they show a noticeable change on both the linguistic and orthographic levels, which makes them ideal for illustrating the idea behind our stem-based approach.

Firstly, the noun example: nouns ending in Hamza and how their stems are extracted to be subsequently stored. From an orthographic point of view there are four forms of the final Hamza; namely: "ء", "أ", "ئ"and "ؤ". Hence, every word ending in "Hamza" is stored in the dictionary without the "Hamza" so that the possible word forms can be generated from a single same stem as shown in Table 1. For example, the word "صحراء" "saħaraa?" 'Desert' is stored in the dictionary as "صحرا", from which all the different paradigms can be generated (صحراوات، صحرائها، صحراء). But, does this imply that every noun ending in "ء" should be stored in the dictionary without "ء"? It seems not, considering the difference between the words "صحراء" and "ضوء" "daw?" 'Light' for one example. Although they both end in "ء", the latter has two lexical entries (stems) stored in the dictionary, i.e. "ضو" and "أضوا". In this case, the various paradigms of the lexeme will be derived from two stems rather

[2] Alansary (forthcoming). The technical Design of the Arabic-UNL Dictionary: Challenges and Implications.

than one. A second type of words ending in Hamza can be distinguished; words like "مبدأ" "mabda?" 'principle'. Does this type of "Hamza" affect the form and number of stems? Yes, as shown in table 1, this type of word ending generally has two stems. There are, however, some exceptions e.g. the word "أبطأ" "?abṯa?" 'slower' which is stored as one lexical entry "أبطأ".

Stem1	WF1	WF2	Stem2	WF1	WF2
مبد	مبدأ- مبدئهم	مبدآن- مبدئي	مبادئ	مبادئهم	مبادئها
أبط	أبطئنا	أبطآ- أبطأهم	-	-	-

"Table 1: Stem forms of the final Hamza 'أ' nouns"

The third type of final "Hamza" is "ؤ". This type of words is stored as a single stem without the Hamza, as in "تكافؤ" "takaafu?" 'equivalence' which is stored as "تكاف". Yet, there are some exceptions such as "لؤلؤ" "lu?lu?" 'Pearls' which is stored with two stems "لؤلؤ - للآلئ". The most interesting point here is that the final Hamza is not deleted, rather, it is considered a part of the stem.

Secondly, some verb examples that show how different types of verbs are stored in the dictionary can be pointed out such as: "باع" "baaʔa" 'sold', "خاف" "khaafa" 'Feared' and "قال" "qaala" 'said'. Although all these verbs are Hollow "باع" has 3 stems (باع، بيع،بع), "خاف" has 2 stems (خاف، خف) and "قال" has four stems (قال، قول، قل، قيل) because each verb follows a different morphological pattern; "باع" follows "فَعَلَ-يفعِلُ", "خاف" follows "فَعَلَ-يفعَلُ" while "قال" follows "فَعَلَ-يفعُلُ".

So far, we have presented some examples of the variations in the forms of lexical entries and their corresponding stems. It should be clear by now that developing a successful strategy for storing the minimum number of stems capable of generating and analyzing all Arabic words is not an easy task, to say the least. If these complexities were encountered in only two examples, how complex would be extracting and storing the suitable stems for the rest of the words that exist in the Arabic language.

5.2 Augmenting Stems with Linguistic Features

Various types of information about the linguistic behavior of words are stored in the dictionary to enable generating all, and only, correct word forms as well as using these words to constitute well-formed syntactic structures.

At the morphological level, much information has been stored with each stem describing its morphological behavior in order to facilitate the selection of the correct word ending. For example, on the one hand, a code has been stored with the stem "صحرا" in order to add the final "ء" in case of singular, and "وات" in case of plural and so on. On the other hand, a list of suffixes is stored in the dictionary with information about the behavior of each and the type of words to which it can be attached.

At the syntactic level, a list of syntactic attributes has been attached to control word order inside sentences. This has been achieved by studying the range of syntactic arguments a verb takes and the range of semantic arguments the verb expresses. The difference between the verb "ميز" "mayyaza" 'favored' and the verb "تميز" "tamayyaza" 'characterized', as in examples (1) and (2) below, lies in the fact that the theme of the verb in (1) is its syntactic subject while the theme of the verb in (2) is its syntactic object. Such information represents the mapping between syntax and semantics, which is substantial in our dictionary to enable the grammar to decode UNL semantic networks into well-formed syntactic structures.

(1) tamayyaza ?al?ibn baltafawwuq تميز الابن بالتفوق
The son was characterized by his excellence

(2) mayyaza ?al?ab ?ibnahu ميز الأب ابنه
The father favored his son

6. Deconverting UNL into Arabic

The Arabic language is a morphologically and syntactically rich language; hence, its automatic generation from an interlingua is very complicated. The automatic generation of an Arabic text from the UNL interlingua, which is a hyper-semantic network, should use of all the linguistic information conveyed by the universal representation. The DeConverter generates the target sentences in a native language from UNL

expressions using deconversion rules and the dictionary of the respective language. This process starts by recognizing the main concept in the sentence. Arabic deconversion rules, then, have to select the suitable Arabic syntactic structure in which the universal representation should be generated; a topic-comment structure or a VSO structure, for example. Finally, the dictionary should provide the appropriate lexical items of the target language. In order to follow the main steps of generating Arabic language from interlingua, the next subsections will focus on the mapping between concepts and Arabic words, the syntactic stage and the morphological stage respectively.

6.1 Stages of Arabic Generation Grammar

The Arabic generation grammar is divided into three stages; namely, the lexical mapping stage, the syntactic stage and the morphological stage.

6.1.1 The Lexical Mapping Stage

The lexical mapping stage performs the mapping between the meaning conveyed by the concepts of the intermediate representation (UNL intelingua) and the lexical items of the target language. For example, the word "part" can be translated in the Arabic language as "جزء" "guz?" or "دور" "dawr", or "ناحية" "naaḥiyah" or "منطقة" "manṯiqah" according to the context in which it appears. UNL provides a different concept for each of these words distinguishing between the different senses of the word "part", a fact that helps in overcoming the problem of lexical ambiguity during the translation process. To be more precise, the word "part" is expressed in the UNL representation by four different concepts "part(icl>section)" which is mapped with the corresponding Arabic noun "جزء", the concept "part(icl>role)" which is mapped with the corresponding Arabic noun "دور", the concept "part(icl>area)" which is mapped with the corresponding Arabic noun "ناحية" and the concept "part(icl>region)" which is mapped with the corresponding Arabic noun "منطقة". However, in many cases the situation is more complex. One of these cases is when a single concept is mapped with more than one lexical item in the target language. For example, the Universal Word "people(icl>person)" can be mapped with either "شخص" "shakhs̲", "إنسان" '?insaan', "نسمة" "nasamah" or "ناس" "naas" in Arabic; the selection between these different lexical items depends entirely on the context in which they appear. This represents a real challenge. Consider the following example:

(3) ويفتقد اليوم نحو 2 بليون **نسمة** الصرف الصحي الكافي
Wa yftaqid ?alyawm naḥwa 2 billion nasamah ?als̲arf ?als̲iḥḥy ?alkaafy.
Today, 2 billion people lack adequate sanitation.

(4) يجلب **الناس** الماء من مصدر بعيد
yaglib ?alnaas ?almaa? min masdar baʿiid.
People bring water from a distant source.

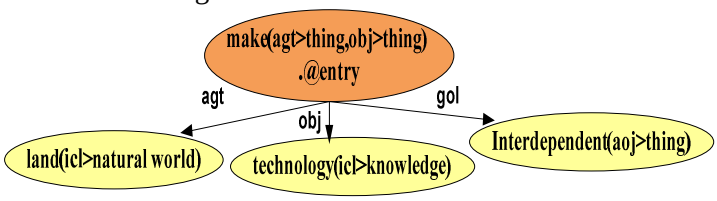
In (3), generation rules should select the lexical item "نسمة" "nasamah" because it represents a *census*, unlike (4). Consequently, much effort has been invested in providing the Arabic dictionary with the semantic features necessary to restrict the lexical and/or grammatical co-occurrences of each word.

6.1.2 The Syntactic Stage

The syntactic stage is concerned with the order of words in the node list under generation. It can be divided into two phases. The first builds the main skeleton of the sentence; a process that starts by the deconverter automatically identifying the main predicate of the sentence (marked in the semantic network by the attribute *@entry*). Then, syntactic rules are applied to insert the syntactic arguments of the predicate, which sometimes differ from the semantic relationships the predicate has in the semantic network. The second phase in the grammar deals with the generation of modifiers; in this phase, rules are formulated to generate the modifiers of each node in the network, whether they were of the same type, or different in types (mod, aoj. . . etc.). Many challenges are faced in the syntactic stage, such as the mapping between semantics and syntax, and the problem of generating nominal chunks (N_Cs). Each of these challenges will be discussed in more detail in the following sections.

6.1.2.1 The Semantics-Syntax Interface

In recent years, all the studies that addressed the interaction between syntax and semantics regarded syntax as the basis from which semantic relations can be deduced; in other words, they identify the syntactic arguments (subject, object, . . . etc.) and then determine the semantic relations between these categories (agent, theme, . . . etc.) (Alcántara and Moreno, 2004). This way of thinking was the cause of many problems, the most prominent of which is Ambiguity. The UNL system is semantically based; it is the only linguistic system, as far as we know, that starts off by laying down the semantic relations and from there pinpoints the syntactic categories. Consider the UNL semantic network in Figure 4.



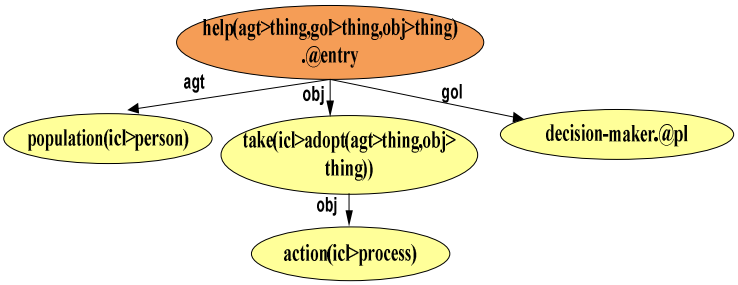
"Figure 4: The UNL graph for the English sentence 'Technology has made lands interdependent'"

The syntactic module arranges the nodes of the network in order to generate a well-formed Arabic sentence structure. A 'VSO' structure will be selected if the main entry is a verbal concept. In this case, "make(agt>thing,obj>thing)" will fill in the 'V' slot, the agent will fill in the 'S' slot, the theme or the object will fill in the 'O' slot and the goal will fill in the 'ID_O' slot. The formal rules in (5) map semantic relation with syntactic slots. The braces refer to the existing node in the node list, the quotations refer to the node to be inserted into the node-list, and "P" refers to the <priority> of the rule.

(5) a) :{ p2,^@entry,^GOL-O:::}"^15:p2:gol:"P300;
b) :{ p2,^@entry,^OBJ-G:::}"^15:p2:obj:"P200;
c) :{ p2,^@entry :::}"^15:p2:agt:"P100;

Accordingly, the distribution of the constituents of the sentence can be controlled to generate the following legible Arabic sentence: "جعلت التكنولوجيا الأراضي مترابطة" "ga؟alat ?altiknulughya ?al?araady mutaraabitah" 'Technology has made lands interdependent'.
The generation process of the sentence highlights the type and number of syntactic arguments the verb takes and determines the type and number of semantics arguments that a predicate expresses (as seen in the network in Figure 4) which reflects the interface between semantics and syntax (Alansary 2007). In some cases, the decoding process is simply a 1 to 1 mapping between semantics and syntax; however, in most cases, the situation is much more complex. We cannot simply map the goal of the verb with the indirect object in the syntactic realization in every sentence, sometimes the goal may map with the direct object in the syntactic realization. In Figure 5, the entry of the network fills in the 'V' slot, the agent in the 'S' slot, the goal in the 'O' slot and the theme in the 'ID_O' slot to generate the legible Arabic sentence: "يساعد السكان صناع القرار في اتخاذ القرار" "yusaa؟id ?alsukkaan sunnaa؟ ?alqaraar fi ?ittikhaadh ?alqaraar" 'population can help decision-makers in making decision'. The formal rules in (6) map semantic relations with syntactic slots. Rule (6a) inserts the right node if it receives "obj" relation from the left node, therefore, "اتخاذ" will be inserted. Rule (6b) has a lower priority; it inserts the node that receives a "Gol" relation from a verb that is marked in the lexicon with the feature "GOL-O". This feature informs the grammar that the 'goal' of this verb is syntactically its "obj", therefore it is mapped with the "obj" slot, and accordingly, "صناع القرار" is inserted. Rule (6c) has the lowest priority, it inserts the node that receives "agt" relation, and therefore, the node " السكان " will be inserted.

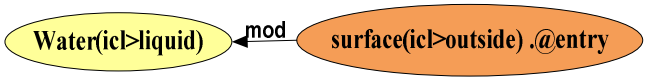
(6) a) :{ p2,^@entry,OBJ-G:::}"^15:p2:obj:"P300;
b) :{ p2,^@entry,GOL-O:::}"^15:p2:gol:"P300;
c) :{ p2,^@entry :::}"^15,@pl,N2:p2:agt:"P100;



"Figure 5: the UNL graph for the English sentence: population can help decision-makers take action"

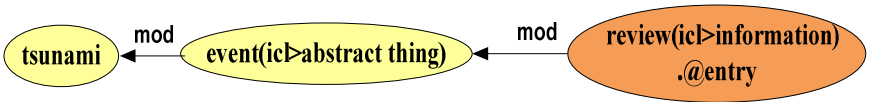
6.1.2.2 Generating Nominal Chunks

Nominal chunks are chunks extending from the beginning of the noun phrase to the head noun. All kinds of modifiers and/or specifiers occurring between the beginning of the noun phrase and the head noun are included in N_Cs. In order to generate idiomatic Arabic structures, rules should be devised to control the generation of nominal chunks (as seen in the network in Figure 6).



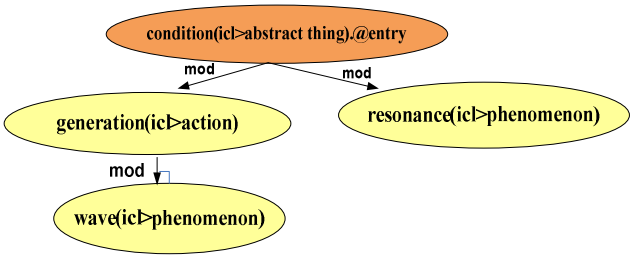
"Figure 6: The graph of the noun phrase 'Water surface'"

This small network suggests that the concept "water(icl>liquid)" restricts the concept "surface(icl>outside)". Syntactic rules take the decision whether the word "ماء" "maa?" should be inserted after the word "سطح" "saṯħ" or before it. From the previous example, we can conclude that the applied rule will generate the main concept in the nominal phrase, the one that assigns the "mod" relation, followed by the other nominal concept which receives it. This rule can be applied to generate several structures (see the semantic network in Figure 7).



"Figure 7: The UNL graph for the English phrase 'A review of tsunami events'"

The syntactic rules apply to generate the syntactic structure "استعراض أحداث تسونامي" "?istiʕraad ?aħdaath tsunami" 'A review of tsunami events' which is generated by inserting the main concept "استعراض", marked in the network by the attribute '@entry', followed by the other nominal concept which receives the 'mod' relation from "أحداث", then the final concept in the network, "تسونامي". However, the situation will be more complicated when the main concept is assigning "mod" relations to two nominal concepts at the same time (Figure 8). In this case, the question would be; which node in the network should be inserted first after the main concept?



"Figure 8: The UNL graph for the English phrase 'The resonance conditions of generation of the wave'"

In order to decode this semantic network into Arabic. The Arabic rules should be able to make the decision that the word "رنين" "raniin" 'resonance' should be inserted first, then the word "تولد" "tawallud" which is the Arabic counterpart of the concept "generation(icl>action)", this is achieved through the formal rules in (7).

(7) (a) :{@enrty,N:::}"N,10::mod:"P200;
(b) :{@entry,N:::}"N,^10::mod:"P100;

A higher priority is given to the formal rule in (7a) which states that if the left node assigns a "mod" relation and the word class of this node is an N, insert the Arabic counterpart of the nominal concept "generation(icl>phenomenon)" – "تولد" – which receives a "mod" relation. Thus, the Arabic rules generate (8) but do not generate (9). The output in (9) is ambiguous because "رنين" can be interpreted as referring to "موجة" "mawgah" 'wave', which is an incorrect analysis, or as modifying "حالات" "ħalaat" 'conditions' which is the correct analysis suggested by the semantic network in Figure 8.

(8) حالات الرنين لتولد الموجة
ħalaat ?alraniin litawallud ?almawgah.
The resonance conditions of generation of the wave

(9) حالات تولد موجة الرنين!
ħalaat tawallud mawgat ?alraniin.
The resonance conditions of generation of the wave

6.1.3 The Morphological Stage

The morphological stage specifies how words are formed and adjusts the gender, number, person and definiteness agreement. In this section, we will highlight the main morphological features used to improve the quality of the morphological structure of words in UNL-based Arabic machine translation systems. The morphological rules adopted in the current system are adequate enough to deal with the Arabic morphological phenomena that come to light in different grammatical contexts. The next subsection will shed light on the morphological generation of word forms, agreement, and reference resolution.

6.1.3.1 Handling word forms

The Arabic verbs are divided into different categories, each category has a morphological behavior different from the others. For example, the form of the defective verb "أتى" "?ataa" 'come', with an initial hamza changes according to the number and the gender of its subject, and according to the tense. Therefore, seven forms have been designed to represent all the possible paradigms of the verb "أتى" ; these are [آت], [أت], [ئت], [أتا], [آتي], [أتى] and [أتي]. Each of these forms has a specific code that enables the grammar to pick the appropriate form according to the syntactic structure of the sentence and the tense of the verb. In addition, a given affix will be added to the HeadWord depending on the subject of the verb to generate the final realized form. Backtrack rules and transfer rules are also used to control the selection of the correct stem that represent the lexeme. A striking feature of Arabic morphology is that verbs of the same type do not necessarily behave in the same manner morphologically. For example, both the verbs "أخذ-أكد" "?akkada-?akhadha" 'confirmed-took' have the hamza in the initial position but exhibit different morphological behavior. The morphological rules are powerful enough to generate correctly the verbs "يأخذ" "ya?khudh" and "يؤكد" 'yu?akkid' in the present tense and not, for example, an incorrect form such as "يأكد" "ya?kd". The reason behind this variation in the morphological behavior is that the morphological pattern of the two verbs is different.

6.1.3.2 Achieving Agreement

Agreement is achieved using the attributes stored in the lexicon with each entry. Morphological rules should achieve verb-subject agreement, noun-adjective agreement and number agreement. The scope of the agreement covers not only contiguous concepts but also discontiguous concepts that are not linked semantically together. Consider the examples (10) and (11):

(10) البيئة المادية التي اعتبرت ثابتة
?albii?ah ?almaaddyyah ?allaty ?uʕtubirat thaabitah.
The physical environment that was perceived as unchanging.

(11) جعلنا الموضوع مهتمين
gaʕalnaa ?almawduuʕ muhtammiin.
The topic raised our attention.

In (10), the word “ثابت” “thaabit” ‘unchanging(aoj>thing)’ is related semantically to the verb “اعتبر” “?iʕtubira” ‘perceive(icl>become aware)’, but it is not related semantically to the word “بيئة” “bii?ah” ‘environment(icl>surrounding)’. However, they are related together syntactically, therefore, they must agree in gender.

6.1.3.3 Reference Resolution

The scope of morphological generation includes the rules dealing with generating reference. In (12) the pronoun “ها” “ha” ‘its’ is generated as a reference to its antecedent “ثقافة” “thaqaafah” ‘culture’, which is a singular feminine noun. In (13) the masculine pronoun “ه” ‘its’ is generated to suit its antecedent “العلم” “alʕilm” ‘science’, which is a singular masculine noun. See the formal rule in (14).

(12) طمح الزعماء الروحانيون التقليديون في كل أرض وزمن أن يوفقوا التمثيل الكوني والمهارات التقنية التي طورتها الثقافة الخاصة مع القيم الأخلاقية والمعايير المثالية السلوك التي وضعتها تلك الثقافة نفسها.

tamaħa ?alzuʕamaa? ?alrawħaniyyuun ?altaqlidiyyuun fi kol ?ard w zaman ?an yuwaffiqu ?altamthiil ?alkawny wa ?almaharaat ?altiqaniyyah ?allaty tawwarathaa ?althaqaafah alkhaassah maʕa ?alqiyam ?al?akhlaaqiyyah w ?almaʕaayiir ?almithaaliyyat ?alsuluuk ?allaty wadaʕatha telka ?althaqaafah nafsihaa.
Traditional spiritual leaders, in every land and age, sought to reconcile the cosmological representations and technical skills developed by their own particular culture, with the ethical values and ideal standards of behavior devised by that same culture.

(13) dhalika ?alʕilm nafsuh . ذلك العلم نفسه That same science.

(14) a) :{<aoj,R1}{>aoj,NLNK,^NFLNK:NFLNK}P200;
b) :{<aoj,R2}{>aoj,NLNK,^NMLNK:NMLNK}P200;

In addition, the morphological rules are able to generate another kind of reference; the reference that is already expressed in UNL by a universal concept. For example, in (15) there is a ‘mod’ relation between the word “تراث” “turaath” ‘legacy(icl>property)’ and the pronoun ‘they(icl>person)’, which has three possible forms in the dictionary “ه”, “ها” and “هم”. Morphological rules are able to select the correct pronoun “ها”, which refers to the feminine non-human antecedent “ثقافة’. In (16), however, there is a ‘mod’ relation between the word “عمل” “ʕamal” ‘work(icl>activity)’ and ‘they(icl>person)’. The rules in this case can generate the reference “هم” that refers to the masculine plural human antecedent “علماء” ‘ʕulamaa?’.

(15) أعادت الثقافات تقييم تراثها
?aʕaadat ?althaqaafaat taqyiim turaathihaa.
Cultures re-evaluated their heritage.

(16) لم يعد يدعي العلماء أن عملهم ليس له تأثير حالي
lam yaʕud yaddaʕy ?alʕulamaaʕ ?anna ʕamalahum laysa lahu ta?thiir ħaaly
Scientists can no longer claim that their work will have no immediate effects

6. An Illustrative Example of the Arabic Generation Process from UNL interlingua

Figure 9 demonstrates an example of the generation process to deconvert a given UNL expression into Arabic. In the lexical mapping phase, different lexical entries have been retrieved from the dictionary. In the syntactic phase, word order rules have been applied to generate well- formed sentence structures. The morphological phase finalize the generation of word forms.

UNL expression

```
{unl}
aoj(lead to(aoj>thing,obj>thing):2X.@entry,          activity(icl>action):0K.@topic.@pl)
obj(lead to(aoj>thing,obj>thing):2X.@entry,          deterioration(icl>phenomenon):35)
obj(deterioration(icl>phenomenon):35,     quality(icl>attribute):3Z)
mod(quality(icl>attribute):3Z,     ground water:3M)
aoj(such as(aoj>thing,obj>thing):0W,       activity(icl>action):0K.@topic.@pl)
qua(activity(icl>action):0K.@topic.@pl,   a lot of(qua<thing):02)
aoj(economic(aoj>thing):0B,       activity(icl>action):0K.@topic.@pl)
obj(such as(aoj>thing,obj>thing):0W,       :01)
and:01(mining(icl>industry):2P.@entry,   construction(icl>way,obj>concrete thing):27)
and:01(construction(icl>way,obj>concrete thing):27,          deforestation(icl>action):1S)
and:01(deforestation(icl>action):1S,          application(icl>use):14)
obj:01(application(icl>use):14,     fertilizer(icl>substance):13)
{/unl}
```

Lexical mapping

```
================== WORD LISTS : 14 ==================
[عديد ] "a lot of(qua<thing)" (ST,19,R2) <A,0,0>;
[عملي] {} "activity(icl>action)" (ST,N,10,N1,PL1,R1.1,R1,NAN) <A,0,0>;
[أنشط] {} "activity(icl>action)" (ST,N,N2,N2.1,R2) <A,0,0>;
[نشاط] {} "activity(icl>action)" (ST,N,N1,PL1,R2) <A,0,0>;
[تدهور {}[تدهور "deterioration(icl>decline in quality)" (ST,19,N1,R2) <A,0,0>;
[سماد] {}  "fertilizer(icl>substance)" (ST,N,N1,R2,NAN) <A,0,0>;
[أسمد] {} "fertilizer(icl>substance)" (ST,N,N2,N2.1,R2,NAN) <A,0,0>;
[أد]  {} "lead to(aoj>thing,obj>thing)" (ST,V,3V,V4,1?,2G,BE,21+O,F8.1) <A,0,0>;
[أدى] {} "lead to(aoj>thing,obj>thing)" (ST,V,3V,V1,1?,2G,BE,21+O,F8.1) <A,0,0>;
[أدي]{}  "lead to(aoj>thing,obj>thing)" (ST,V,3V,V2,1?,2G,BE,21+O,F8.1) <A,0,0>;
[وَد] {} "lead to(aoj>thing,obj>thing)" (ST,V,3V,V5,1?,2G,BE,21+O,F8.1) <A,0,0>;
[وَدي] {}"lead to(aoj>thing,obj>thing)" (ST,V,3V,V3,1?,2G,BE,21+O,F8.1) <A,0,0>;
```

Syntactic Phase

```
================== NODE LIST:00 ==================
/<</[أد|lead to(aoj>thing,obj>thing)]
================== APPLIED RULE ==================
:{p1,@entry,21,^22,^23:::}"^A2,^C3,^@p1:+p1:obj:"P100;
================== NODE LIST:00 ==================
/<</[أد|lead to(aoj>thing,obj>thing)]/[تدهور|deterioration(icl>decline in quality)]/>>/
================== APPLIED RULE ==================
:{p1,@entry,21,^22,^23:+2.3AOJ>:::}"^A2,^C3,@p1,N2:+p1,+2.3AOJ<:aoj:"P100;
================== NODE LIST:00 ==================
/[<<]/أد/|أنشط|/تدهور/>>/
```

Generated Structure

أد/ أنشط/اقتصادي/مثل/اضاف/سماد/از ال/غاب/ تشييد/تعدين/تدهور/جود/مياه/جوفي

Morphological Phase

```
    lnode>>>>>
    [أد]  {} "lead to(aoj>thing,obj>thing)"
(2.3AOJ>,2.2.2OBJ>,p1,@entry,>aoj,>obj,ST,21,3V,V4,1?,2G,BE,21+O,F8.1)<A,0,0>;
    mode>>>>>
    [أنشط {} "activity(icl>action)" (p1,2.3AOJ<,@topic,@pl,<aoj,<aoj,>qua,<aoj,ST,19,N2,N2.1,R2)
<A,0,0>;
    ================== APPLIED RULE ==================
    ?L{2.3AOJ>,21,V4,3V,2G,1?,^@past,^@complete,^@present,^@progress,^@will:::}{2.3AOJ<,@pl,
NAN,^SCOPE:::}P220;
    ==========================================
    >>>>> lnode
    [وَدي] {} "lead to(aoj>thing,obj>thing)"
(2.3AOJ>,2.2.2OBJ>,p1,@entry,>aoj,>obj,ST,21,3V,V3,1?,2G,BE,21+O,F8.1)<A,0,0>;
    >>>>> mode
    [أنشط] {} "activity(icl>action)" (p1,2.3AOJ<,@topic,@pl,<aoj,<aoj,>qua,<aoj,ST,19,N2,N2.1,R2)
<A,0,0>;
```

Generated Sentence

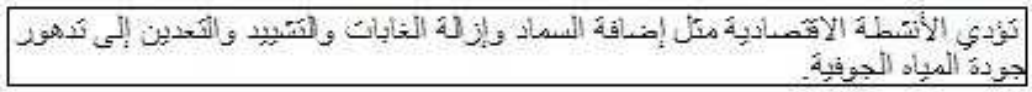
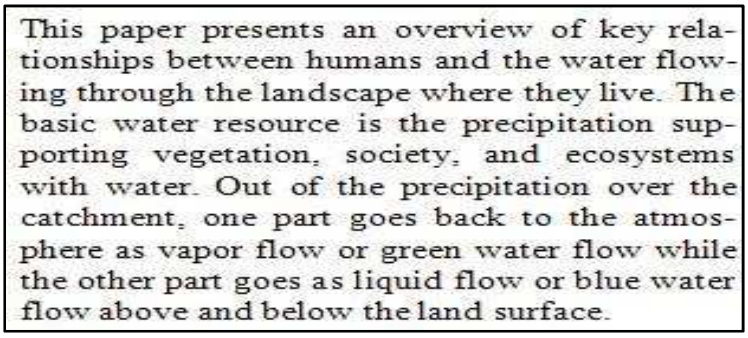


Figure 9: An example of the deconversion process

7. Evaluating the Deconversion's output

There are various means for evaluating the performance of machine translation systems. The oldest is the use of human judges to assess the quality of a translation, but human evaluation is very time-consuming and therefore not always an option. In addition, when using human evaluation objectivity is always in question. The other method is Automatic Evaluation. The greatest advantage of using such a method is that the scoring is objective, automated means of evaluation include BLEU, NIST and METEOR. The use of automatic evaluation metrics became quite widespread in the Machine Translation (MT) community, mainly because such metrics provide an inexpensive and fast way to assess translation quality. Most efforts focus on devising metrics based on measuring the closeness of the output of MT systems to one or more human translation(s); the closer it is, the better. The most commonly used MT evaluation metrics in recent years has been BLEU (Papineni et al 2002), an *n*-gram precision metrics that demonstrates a high correlation with human judgment of system adequacy and fluency. The Recall technique has become extremely important in assessing the quality of MT output, as it reflects to which degree the candidate translation covers the entire content of the reference translation (Lavie et al 2004). The Arabic UNL language center has generated Arabic translations for 25 documents (more than 13000 sentences) from the Encyclopedia of Life Support Systems (EOLSS) using the UNL technology. Figure (10) shows an example of a short paragraph from the English document: "*Human Interaction with Land and Water*". Figure (11) shows the Arabic output generated from the enconverted version of figure (10). The output in figure (11) represents machine quality translation that was, afterwards, subject to post-editing to raise its quality to publishing standards. Figure (12) shows the Arabic output after post-editing.



"Figure 10: Example of an English text from the EOLSS"

"Figure 11: The deconverted Arabic text"

"Figure 12: The translated text in publishing quality"

Two methods have been used to evaluate the output of machine translation using UNL, the first is the qualitative evaluation: the English document "*Tsunami"* has been human-translated taking into consideration the UNL expression. The comparison between the deconverted text and the human-translated text has been performed according to some linguistic criteria that evaluate the output on three levels; the syntactic level which includes word order, order of modifiers, case marking and insertion of prepositions and particles. The semantic level that includes the comprehensibility of sentences. The morphological level includes the well-formedness of word forms. And the results were 70-75% syntactically correct, 85% semantically correct, and 90% morphologically correct (Alansary et al 2007).

The second method is the statistical approach that was adopted to evaluate the output of machine translation using UNL. 500 sentences have been selected randomly form the EOLSS and translated by two human translators in addition to another translator who post-edited the machine output by making the minimal changes required. Most efforts focused on devising metrics based on measuring the resemblance between the Arabic deconverter output and one or more human translations. Three matrices were used BLEU, F1 and F mean. The results of this evaluation of the UNL system were compared to other three English to Arabic translation systems; namely, Google, Sakhr's Tarjim and Babylon. UNL translation achieved the best scores in this evaluation, followed by Google, Sakhr and then Babylon. These results were statistically significant at 95% confidence. For the details of this evaluation, cf. Adly and Alansary (2009).

8. Conclusion

This paper presents an approach for machine translation using UNL as an interlingua that intermediates between different languages. It focuses on Generating Arabic from the intermediate representation in the form of semantic networks. This approach has been applied to 25 documents from the Encyclopedia of Life Support Systems (EOLSS). And, it has been proved that the UNL system can make the dream of language-independent semantic analysis a reality, thus breaking language barriers between different nations. The generation of Arabic from the UNL semantic representation has resulted in a translation that outperformed other systems. To our knowledge, this trial can be considered as the first leap towards generating natural language from an interlingua in the history of Natural Language Processing.

References

Alansary, S., Nagi, M. and Adly, N. 2006a. *Generating Arabic text: The Decoding Component of an Interlingual System for Man-Machine Communication in Natural Language,* the 6th International Conference on Language Engineering, 6-7 December, Cairo, Egypt.

Alansary, S., Nagi, M. and Adly, N. 2006b. *Processing Arabic Content: The Encoding Component of an Interlingual System for Man-Machine Communication in Natural Language*", the 6th international conference on language engineering", 6-7 December, Cairo, Egypt.

Alansary, S. 2003. Building a Computational Lexicon for Arabic, *the 17th ALS Annual Symposium on Arabic Linguistic,*. 9-10 March 2003, Alexandria, Egypt.

Alansary, S., Nagi, M. and Adly, N. 2007. A Semantic-Based Approach for Multilingual Translation of Massive Documents. Alansary, S., Nagi, M., Adly, N.: A Semantic Based Approach for Multilingual Translation. In The 7th International Symposium on Natural Language Processing (SNLP), Pattaya, Thailand.

Alcántara M., Moreno A., 2004 *Syntax to Semantics Transformation: Application to Treebanking***.** Workshop Frontiers in Corpus Annotation at HLT-NAACL, 2-7 May, Boston.

Adly, N. and Alansary, S. 2009. *Evaluation of Arabic Machine Translation System based on the Universal Networking Language*, the 14th International Conference on Applications of Natural Language to Information Systems "NLDB 2009", 23-26 June, Saarland University, Saarbrucken, Germany.

Dorr, B. 1987. *UNITRAN: A Principle-Based Approach to Machine Translation.* AI-Technical Report 1000, Massachusetts Institute of Technology, Cambridge, MA, USA.

Lopez, A., 2008. *Statistical Machine Translation*, In ACM Comp. Surveys, Vol. 40.Lavie, A., Sagae, K., Jayaraman, S. 2004 *The Significance of Recall in Automatic Metrics for MT Evaluation*. In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), pp. 134–143, Washington.

Nyberg E. H., Mitamura T. 1994. *Evaluation Metrics for Knowledge-Based Machine Translation*, Proceedings of COLING-94, August 5-9, Kyoto, Japan.

Nyberg E. H., Mitamura T. 1992. *The KANT system: fast, accurate, high-quality translation in practical domains*, Proceedings of the 14th conference on Computational linguistics, 1992 - Volume 3.

Papineni, K., Roukos, S., Ward, T., Zhu, W. 2002 *BLEU: a Method for Automatic Evaluation of Machine Translation.* In 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318, Philadelphia

Shaalan .K, Monem .AA, Rafea A. and Baraka H. 2006. *Mapping Interlingua Representations to Feature Structures of Arabic Sentences*, The Challenge of Arabic for NLP/MT. London.

Soudi, A., Cavalli-Sforza, V., & Jamari, A., 2002. A Prototype English-to-Arabic Interlingua-based MT System," Proceedings of the Workshop on Arabic Language Resources and Evaluation - Status and Prospects, 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain.

Uchida, H, 1996. *UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration*. UNU/IAS/UNL Center. Tokyo, Japan.

Uchida H., Zhu M., 2002 *Universal Word and UNL Knowledge Base"***,** *ICUKL*, Goa, India.

Uchida H., Zhu M., 2005. *UNL2005 for Providing Knowledge Infrastructure***,** *SeC2005 Workshop*, Chiba, Japan.