LILY: Language-to-Interlanguage-to-Language System Based on UNL

Sameh Alansary [*1],MagdyNagi [**2]

Bibliotheca Alexandrina, Alexandria, Egypt
[*]*Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt*
[1]sameh.alansary@bibalex.org

[**]C*omputer and System Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt*
[2]magdy.nagi@bibalex.org

***Abstract*— The Interlingua-based approach to Machine Translation (MT), with all its potentials and merits has been widely regarded as the most appealing approach; yet, very few systems have ever been able to achieve its theoretical prospects. Hence, this paper examines LILY (Language-to-Interlanguage-to-Language sYstem); an interlingua-based human-aided multilingual machine translation web service. It is expected to provide end-to-end high-quality translations through semi-automatic (human-interactive) analysis of the source text into the Universal Networking Language (UNL) and fully-automatic generation from the resulting UNL document into several different target languages.**

1 INTRODUCTION

Language is the principle medium of communication. It represents the ideas and expressions of the human mind. Currently, more than 5000 languages exist in the world which reflects the scope of the linguistic diversity. Access to information written in another language is of a great interest and the means of sharing information across languages is translation, therefore creating tools for machine translation is crucial. The developments that took place in the Information Communication and Technology (ICT) field have caused a revolution in the process of machine translation [1]. Research efforts have been exerted to explore the possibility of automatic translation from one language to another. Nowadays, there are different available tools which support translation of texts into one or more languages. Online translation is offered by Yahoo and Altavista through Babelfish. Bing Translator of Microsoft and Google Translate from Google are tools that are widely used for the translation by the members of the web community. Firefox uses Greasemonkey application to translate the text to other languages. Google Chrome Beta offers translation if the accessed web page is in a language other than the default language. There have been major initiatives from various research organizations and government agencies to develop tools for automatic translation of texts in an attempt to achieve wider outreach and bridge the gap of language diversity [1].

There are various methodologies for machine translation. However, the objective has always been to render the meaning of the original text into the translated one. In general, the process of translation consists of two different levels: the first one is the metaphrase which means “word to word” translation, meaning that each word in the original text is translated into the corresponding one in the target text. But the results are not always satisfying because the target text does not always convey the same meaning of the original text. The second level is the paraphrase; in this level the translated text carries the gist of the information of the source language, it focuses on the meaning of the source language rather than word to word mapping.

In this paper we will present a machine translation system called LILY (language-to-interlanguage-to-language system), employing a rule based machine translation methodology. LILY system can translate from any participating natural language to another different natural language. Section 2 will present the different methodologies of machine translation systems. Section 3 will present our project history and the interlingua that underlies its action (UNL). Section 4 illustrates the basic components of LILY; the system’s open-source components. First, the language resources (dictionaries and grammars). Second, the software used in building and running the system (analysis and generation engines). Each of these components is described and their current state is made clear. Section 5 describes the computational phase of the system, describing the different stages of developing both the analysis and generation grammars, pointing out the adopted linguistic theory. Section 6 will describe LILY `s interface and illustrate how this system is used. In section 7 LILY `s output will be evaluated. Finally, section 8 will conclude the paper.

2 METHODOLOGIES OF MACHINE TRANSLATIONS

There are different methods of machine translation; dictionary based machine translation, knowledge based machine translation, corpus based machine translation, context based machine translation, example based machine translation and rule base machine translation (RBMT), each will be discussed briefly in this section below.

A. *Dictionary Based Machine Translation (DBMT)*

It is considered as the first generation of machine translation (from 1940s to mid-1960s). This method depends on the words equivalent and it has been very helpful in translating the phrases not the sentences [1]. This method based on dictionary entries, which means that the words will be translated as a dictionary does – word by word, usually without much correlation of meaning between them. Dictionary lookups may be done with or without morphological analysis or lemmatization. This approach to machine translation is probably the least sophisticated.

B. *Knowledge Based Machine Translation (KBMT)*

This kind of translation is focused on "concept" lexicon representing a domain. KANT is an example of knowledge based machine translation. There are three main advantages of this architectural approach: Increased accuracy of translation; by allowing the creation of lexicons and grammars that can be as simple or as complex as the application requires, the approach supports a high degree of accuracy in both the source analysis and target generation phases and finally the separation of code and knowledge bases [2].

C. *Corpus Based Machine Translation (CBMT)*

This approach is widely used in MT because of its high accuracy level in translation. Corpus based approach machine translation systems are divided into three types. First is statistical machine translation (SMT) which was introduced in 1949, this method applies a statistical method to translate the texts such as n-gram. Second is example based MT which is based on finding analogues example. This concept was proposed by Makoto Nagao in 1981. Third is context based machine translation (CBMT), this method requires an extensive monolingual target text corpus, a full form bilingual dictionary and monolingual source text corpus [1].

D. *Rule Based Machine Translation (RBMT)*

In this methodology the linguistic rules are built to handle the morphological, syntactic and semantic behavior of the source and target language. Rule based machine translation can deal with different linguistic phenomena. An example of RBMT system is Anglabharati; this system translates from English to Hindi and other Indian languages [1]. This methodology has several approaches such as the direct approach; words of the source language are translated without passing through an intermediary representation, transfer approach; in which the source language is transformed to an intermediary representation but this representation is usually language dependent, and Interlingua approach; in which the source language is transformed to an intermediary language independent from any other natural language then the target language is generated from this intermediary language. Interlingua belongs to the third generation of machine translation, aiming to create linguistic homogeneity across the globe.

LILY system which we will be presenting in this paper is an implementation of this approach. The intermediary representation in this system is called Universal Networking Language (UNL) which is composed of Universal Words (UWs), Relations and Attributes. UWs constitute the vocabulary of the interlingua, however, they are labels that stand for abstract language-independent units of knowledge (concepts) belonging to any of the open lexical categories (nouns, verbs, adjectives or adverbs). They are represented by a unique ID number. Relations and Attributes, on the other hand, represent the interlingua's syntax. Relations stand for the links between the UWs in a given sentence, such as agt (the agent of action), obj (the object of the action)…etc. Attributes modify the network even more by encoding subjective or contextual information, examples are the attribute (@sarcastic) and the attribute (@exclamation) [3].

There are two different distinct processes in LILY; analysis and generation. Analysis involves UNLizing the incoming natural language text into the intermediate representation (UNL expression)[4],[5]. Generation involves Nlizing the intermediate representation into the natural language [6].

3 PROJECT HISTORY AND CURRENT STATUS

LILY is an interlingua-based human-aided multilingual machine translation web service. The system began at 2009 and became fully operational in 2013. LILY system was developed through two phases. The first phase included translating 600 Arabic sentences and that was LILY version 1 [7]. The second phase included translating 10,000 sentences representing 5,000 Arabic syntactic structures and the results were translated sentences of great quality, this was LILY version 2. LILY is a rule based machine translation system; Arabic linguistic resources were developed in Bibliotheca Alexandrina to analyze the Arabic language to transform it into an intermediary language (UNL) and also to generate the Arabic language from the UNL intermediate representation.

UNL which is an artificial language developed for computers attempting to replicate the functions of natural languages in communication, this language is the interlingua employed here. The UNL project has been originally proposed in 1996. The responsible organization is the Universal Networking Digital Language (UNDL) Foundation[1]in Geneva, Switzerland [3], [8], [9], [10] and [11].

[1] The official website of the foundations is available at http://www.undl.org

UNL is capable of representing the meaning of the content of natural language texts in an abstract universal format that is not influenced by the formalities of either the source or target languages. UNL aims ultimately at allowing people to generate, and have access to, information and knowledge, in their own native language by breaking down the language barriers that exclude the majority of people from gaining access to information in their native language.

As well as being able to fulfill the interlingua expectations, LILY has also combined the advantages of this classic approach with the new trend in machine translation, that is being an open-source software. Because of its vast advantages, opportunities and potentials [12], many MT systems are considering following this trend by rendering their resources and software as open source. A rule-based machine translation system is open source only when the source code of its engines and tools are distributed along with the linguistic data of the translation pairs. In addition, tools to maintain and develop the linguistic resources so that they can be used with the engines should also be distributed [12]. LILY fulfills all of the above criteria and, hence, can be positively considered an open-source MT system; moreover, not only are its components open-source, they are also free. As mentioned before in section 2, in LILY MT a system there are two processes; generation and analysis. In order to perform these processes the system depends on language resources and tools which will be described in details in the next section.

4 THE BASIC COMPONENTS OF LILY MT SYSTEM

Two main components taking part in achieving the analysis and generation processes in LILY MT system: language resources and tools. The language resources are developed for each participant language. On the other hand, tools and engines are the same for all languages. They are used to manage the linguistic resources and were developed by the UNDL foundation in cooperation with Bibliotheca Alexandrina. Fortunately, components of the UNL system are free and open-source. These components will be discussed in more detail in the following sub-sections.

A. Language Resources

The UNL System comprises two different types of language resources: dictionaries and grammars. These resources are developed and available at the UNL[arium2].

The UNL system employs a sole tagset in defining all the language resources used within the UNL framework. A tagset should describe the linguistic phenomena exhibited in a natural language, however, this may vary drastically from one language to another depending on the nature and the structure of the language. Therefore, UNL uses a universal tagset capable of describing the linguistic phenomena present in all natural languages, even if some of these phenomena were peculiar to only a handful of languages. This is crucial for generation grammars in handling such cases, and, thus, ensures adequate translatability between all participant languages. In addition, it facilitates understanding and exchanging the available language resources (dictionaries, grammar rules…etc.) across the various UNL language centers. Several of the linguistic constants used in the UNL tagset have already been proposed to the Data Category Registry (ISO 12620), and represent widely acknowledged linguistic concepts, the tagset is available at http://www.unlweb.net/wiki/index.php/Tagset.

1) Dictionaries

Each language center is responsible for compiling its analysis and generation dictionaries according to the specifications of the UNL tagset. The UNL framework distinguishes between generation dictionaries (UNL-NL) and analysis dictionaries (NL-UNL) [13]. However, both are bilingual dictionaries where lexical items of a given natural language are matched with their corresponding abstract Universal Words (UWs). The difference lies in that generation dictionaries are lexeme-based; the target language word is stored in a base form, along with inflectional rules that can generate the different word forms of that lexical item. This saves the compilation and processing time and also constitutes less of a burden on the system's resources. On the other hand, analysis dictionaries contain all the word forms of a certain lexical item since it would be quite difficult and time-consuming to predict the base form of the incoming source language word forms. Both dictionaries follow the format shown in figure 1:

```
[NLW]  {ID}  "UW"  (ATTR , ... )  < FLG ,
```

Figure 1: The general syntax of UNL-NL and NL-UNL dictionary entries

Where: NLW is the lexical item of the natural language. It can be a multiword expression, a compound, a simple word, a non-motivated linguistic entity, or a regular expression.
UW is the abstract concept representing the natural language word.

[2] The main development environment of UNL: http://www.unlweb.net/unlarium/.

ATTR is the list of features of the NLW, these are set according to the UNL tagset. It also includes the inflectional rules in the generation dictionary.
FLG is the three-character language code according to ISO 639-3.
FRE is the frequency of NLW in natural texts. It is used in natural language analysis (the analysis of the source language).
PRI is the priority of the NLW. It is used in natural language generation (the generation of the target language). Figures 2 and 3 show examples from the Arabic analysis and generation dictionaries, respectively.

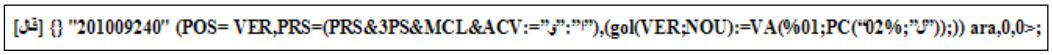

Figure 2: The entry for the Arabic verb "قال" 'say' in the UNL-Arabic generation dictionary

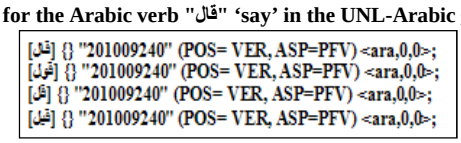

Figure 3: The entries representing the word forms of the Arabic verb "قال" 'say' in the Arabic-UNL analysis dictionary

Currently, 17 participant institutions are compiling the analysis and generation dictionaries of their languages. All current dictionaries are available in the UNL[arium] environment. Logged in users can browse the dictionaries, view their current state of development, statistics of their components, even they can download and export a whole dictionary. For more detailed discussion of UNL lexicons see [13].

2) Grammars

Grammar rules are also developed in the UNL[arium] environment according to the specifications of the UNL unified tagset. These rules are responsible for translating UNL expressions into natural language and vice versa. Hence, they are generally unidirectional; from the source natural language into UNL (UNLization rules) and from UNL into the target natural language (NLizationrules).

Rules can be morphological, semantic, syntactic, phonetic or pragmatic. They are divided into two basic types; transformation rules and disambiguation rules. Transformation rules are used to analyze the source language and transform it into the intermediary representation (UNL representation) and also to generate the target natural language out of the UNL. They consist of seven different types of rules (LL, TT, NN, LT, TL, TN and NT), as indicated below:

LL- List Processing (List-to-List)
LT - Surface-Structure Formation (List-to-Tree)
TT - Syntactic Processing (Tree-to-Tree)
TN - Deep-Structure Formation (Tree-to-Network)
NN - Semantic Processing (Network-to-Network)
NT - Deep-Structure Formation (network-to-tree)
TT - Syntactic Processing (tree-to-tree)

The transformation should be carried out progressively, i.e., through a transitional data structure: the tree, which could be used as an interface between lists and networks. On the other hand, disambiguation rules (D-rules) are optional and they are used to restrict the applicability of transformation rules by assigning priorities. They are used to prevent wrong lexical choices, to prompt best matches or to check the consistency of the graphs, trees and lists. Disambiguation rules are divided into three types; network disambiguation rules, tree disambiguation rules and list disambiguation rules.
Similar to the dictionary, 17 participating institutions are working on the development of their analysis and generation grammars. All current grammars are available for browsing, downloading and exporting in the UNL[arium] with statistics about the components of the grammar of each language.

B. Tools and Engines

Two processes are required in order to perform the translation. First is the analysis process, in which the natural language texts are translated into UNL using a specially designed engine called The Interactive ANalyzer (IAN); which is a natural language text analysis engine. It is the same for all languages, it simply employs the grammar rules, the NL-UNL dictionary of the source language to analyze the input and generate its corresponding UNL expressions. It operates semi-automatically; word sense disambiguation is still carried out by a language specialist, nevertheless, the system can filter the candidates using an optional set of disambiguation rules. Syntactic processing, on the other hand, is carried out automatically using the natural language analysis grammar. IAN is also available as free and as open-source software. Second is the generation process, in which the UNL expression is translated back into natural language using the generation engine the dEep-to-sUrface natural language GENErator engine (EUGENE). Similar to the UNLization

engine, EUGENE is language-independent, it simply uses the target language grammar rules and UNL-NL dictionary in order to decode the incoming UNL document and generate it in natural language format.

5 COMPUTATIONAL TECHNIQUES

As discussed before LILY application is an interlingua based machine translation system; depending on an intermediary language (UNL) to perform the translation process. An analysis grammar was built to analyze the source language with different linguistic levels to transform it to the interlingua representation using the analysis dictionary designed for this task which is mentioned in section 4-A-1. Another grammar was built to generate the target language using the generation dictionary that is designed for this task. The linguistic rules handle the morphological, syntactic and semantic behavior of the source and target language. However, there is a stage that is common between the two grammars that is responsible for the transformation process based on the X- bar theory. The transformation is carried out progressively through a transitional data structure: the tree, which could be used as an interface between lists and networks. The following subsections A, B and C will discuss the linguistic parsing algorithm and the different levels of the development of each grammar (analysis and generation).

A. Linguistic Parsing Algorithm

The transformation from the string text to a semantic graph in LILY MT system is carried out through the X bar theory; the syntactic module should start drawing the syntactic trees for phrases and sentence structures that are part of the corpus, according to the X- bar theory which is a specific implementation of constituency grammars: it is a method of sentence analysis that divides the sentence into constituents, but it states some very specific rules for doing so. The topmost node (S, in the diagram below) is called XP (X-phrase) and is considered to be the maximal projection of a head X. This means that the whole process must be understood bottom-up (from a head to its projections) instead of top-down.

The "X" is actually a variable that must be replaced by any of the possible heads: noun (N), verb (V), adjective (J), adverb (A), etc. In that sense, there is no real XP, but NP's, VP's, JP's, etc. A VP (verbal phrase) is the maximal projection of a verb (V); a NP (noun phrase) is the maximal projection of a noun (N); and so on. The use of the "X" and therefore "XP" comes from the fact that one of the claims of the theory is that all these phrases (NP, VP, JP, etc.) share the same underlying structure. Projections are always binary, i.e., the tree cannot bring more than two branches at a time because this is not possible in X-bar. In order to avoid this, the head may have intermediate projections before the maximal projection. These intermediate projections are called XB (from X-bar), and again must be replaced by the specific categories of the head (NB is the intermediate projection of N, VB is the intermediate projection of V, etc.). The X-bar abstract configuration is depicted in the diagram below:

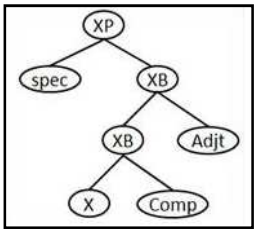

Figure 4: The schema of the X-bar theory

In figure 4, X is the head; the nucleus or the source of the whole syntactic structure, which is actually derived (or projected) out of it. The letter X is used to signify an arbitrary lexical category (part of speech). When analyzing a specific utterance, specific categories are assigned. Thus, the X may become an N for noun, a V for verb, an J for adjective, a P for preposition, etc. comp (i.e., complement) is an internal argument, i.e., a word, phrase or clause which is necessary to the head to complete its meaning (e.g., objects of transitive verbs), adjt (i.e., adjunct) is a word, phrase or clause which modifies the head but which is not syntactically required by it (adjuncts are expected to be extra-nuclear, i.e., removing an adjunct would leave a grammatically well-formed sentence), spec (i.e., specifier) is an external argument, i.e., a word, phrase or clause which qualifies (determines) the head. XB (X-bar) is the general name for any of the intermediate projections derived from X and XP (X-bar-bar, X-double-bar, X-phrase) is the maximal projection of X (http://www.unlweb.net/wiki/X-bar_theory). For a clearer picture of this theory and how it used in the transformation process, it will be applied to the example "يلتقي بالرئيس المعزول" in the following two subsections.

B. Developing the Analysis Grammar

This section is concerned with the UNL-ization from Arabic language into interlingua representation (UNL). The analysis grammar for Arabic reference corpora which consist of 10.000 Arabic sentences have been already built to represent the content to the intermediate representation (UNL expression). Arabic analysis grammar has common

modules such as; the tokenization and segmentation, dealing with polysemous, morphological analysis, syntactic analysis and semantic analysis modules. The following sub-sections will describe each of the common modules.

1) Tokenizing and Segmenting the Input

The tokenization algorithm is based on the entries of the dictionary; the system tries to match the strings of the natural language input against the entries existing in the dictionary. If there are no matches, the string is considered a temporary entry (TEMP). The tokenization algorithm starts from left to right trying to match the longest possible string with dictionary entries, and it assigns the feature TEMP to strings that are not found in the dictionary. For instance, any URL such as "www.undlfoundation.org" should be considered as TEMP. The tokenization algorithm blocks the segmentation of tokens or sequences of tokens prohibited by disambiguation rules which are mentioned in section 4-A-2. The disambiguation rules is concerned with the word segmentation, For example, D-rules can prevent segmenting the word "عاصمتي" 'my capital' into [TEMPع] + [V اصمتي] 'keep silent' by applying the rule (a), "[Nعاصم] 'protector'+ [TEMP ت]+ [PRON ي] 'my' by applying the rule (b), or [Jعاص] 'disobedient' + [V متي] ''by applying the rule (c). Given the fact that the dictionary includes [V اصمتي] 'keep silent', [Nعاصم] 'protector'+ [TEMP ت]+ [PRON ي] 'my', [TEMPع], [Jعاص] disobedient' and [V متي]. Alternatively, D-rules select the final segmentation [عاصمتTEMP] + [PRON ي] 'my'.

(a) (^"ل"^,"أ"^,"ف"^,"و",^BLK)(V)=0;
(b) (N)(^PUT,^ACC,^POD,^BLK,^STAIL)=0;
(c) (J)(^PUT,^BLK,^STAIL)=0;

2) Dealing with Polysemy

One of the most crucial points in any machine translation system is dealing with polysemy, The phenomena of polysemy is an immanent feature of natural language and is manifested at all language levels. The disambiguation rules in LILY MT are responsible for assigning the most appropriate meaning to a polysemous word within a given context. For example, the word "شخصية" 'personality' in "القوة والشخصية"' power and personality' should be assigned to the lexical item referring to the meaning of 'personality' not to the meaning of 'personal', relying on the parallel structure of coordination. However, in handling sentences such as "جاءت المعالجة سريعا" 'the processing was fast', the system will not be able to predict whether the word "معالجة" 'processing' refers to the meaning of 'processing' or 'feminine doctor'.

3) Morphological Analysis

Attributes are used to represent information conveyed by natural language grammatical categories (such as tense, mood, aspect, number, etc.). The set of attributes, which is claimed to be universal, is defined in the UNL Specs (http://www.unlweb.net/wiki/Attributes). The attributes module can handle determiners, pronouns, prepositions and verb forms. It is responsible for substituting certain words or morphemes with attributes, as in the case of quantity quantifiers ("كثير", "قليل", "أي", "كل"…etc.) which will be deleted from the natural language input text and substituted respectively by the attributes "@multal, @paucal, @any, @all …..etc." to be assigned to the following word. The person, number and gender of the pronoun are also described by UNL attributes. For example, "@3, @2, @1, @male, @female, @pl".

4) Syntactic Analysis

The syntactic module in the analysis grammar is divided into two phases; the first phase is responsible for the surface structure formation (List–to-tree rules) and the second phase which is responsible for revealing the deep structure out of the surface structure (tree-to-tree rules), the two sections below will describe this two phases in more details.

- *Surface-Structure Formation*

In this phase, rules are used to parse the tokenized input sentences into a tree structure based on X-bar theory which has been mentioned in section 5-A. This phase starts by composing small trees for the small phrases in the sentence and then combining these small trees together to form a bigger tree. List-to-tree rules are responsible for building the trees for language structures; ordering of rules is required; rules for building noun phrase trees should be followed by rules for building verb phrase trees. In order to have a comprehensive idea, let us consider the following example " يلتقي بالرئيس المعزول" 'he meets the deposed president'. In this example the verb "يلتقي" 'meets' doesn't express a single concept but in fact two; the verb itself and the implicit male subject "هو". This can be detected from the dictionary which describes the grammatical attributes of person, gender, number of the subject, voice and tense of the verb. Each grammatical attribute should be transformed to the suitable UNL attribute as discussed in section 4-A-1. Also each word should be assigned to each suitable concept as shown in figure 5 which is the first step. Moreover, the pronouns, preposition and the determiner assigned with the ID "00" as shown in figure 5.

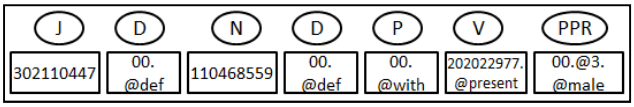

Figure 5: Each word assigned to its grammatical attributes and its concept

The list-to-tree rules are responsible for transforming the list structure in figure 5 to a tree structure. In which, small constituents are combined to gradually form a bigger tree until the whole sentence is analyzed. For instance, the noun "رئيس" 'president' will be projected to the intermediate constituent N-bar (NB) as the head of the noun phrase " الرئيس المعزول" 'the deposed president'. The definite article that precedes the adjective "معزول" 'deposed'; which represents the agreement between the adjective and its depicted noun, will be suppressed. The adjective "معزول" 'deposed' will be projected to the intermediate constituent J-bar (JB) as it is combined with an empty node. Then, the intermediate constituent (JB) "معزول" 'deposed' will be projected to the maximal projection "JP' as it is combined with an empty specifier and linked to the intermediate constituent (NB)"رئيس" 'president' to form a bigger "NB". Similarly, the definite article that precedes "رئيس" 'president' will be projected to the intermediate constituent D-bar (DB) which will be projected to the maximal projection "DP" and linked to this bigger intermediate constituent "NB" to reach the maximal projection "NP". The pronoun "هو" 'he' will be projected to the maximal projection "NP" as shown in figure 6.
The NP "الرئيس المعزول" 'the deposed president' will be linked to the preposition "ب" 'with' to build the intermediate constituent "PB" which will be projected to the maximal projection "PP" as shown in figure 7. Then the PP will be linked to the V "يلتقي" 'meet' to build the intermediate constituent "VB". Finally, the intermediate constituent will be linked to the "NP""هو" 'he' to build the maximal projection "VP" as shown in figure 7.

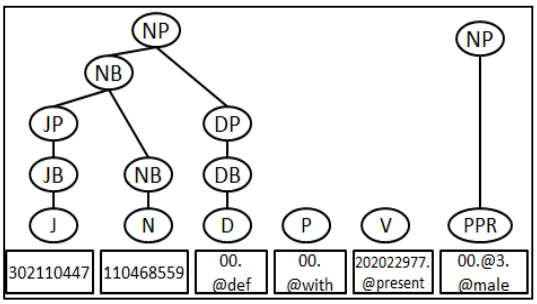

Figure 6: The formation of the surface structure tree

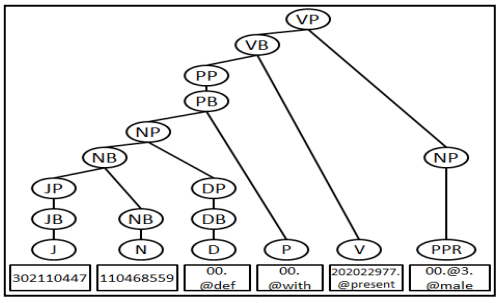

Figure 7: Final surface structure tree

- *Deep-Structure Formation*

The tree-to-tree rules (TT) are used for processing trees. These rules are used for revealing the deep structure out of the surface structure. To reach this, every maximal projection which is not linked with another node will be suppressed. Then, the maximal projection "VP" will be de-arborized into "VB" "يلتقي بالرئيس" 'meet the president' and "VS" "هو يلتقي" 'he meets'. The verb "يلتقي" 'meet' is subcategorized as permitting a prepositional phrase headed by "ب" 'with', accordingly, the preposition phrase will be suppressed linking the verb to the "NP" "الرئيس المعزول" 'the deposed president' by the syntactic role "verb complement" (VC) as shown in figure 8. The NP "الرئيس المعزول" 'the deposed president' will be de-arborized linking the noun "رئيس" 'the president' and the adjective "معزول" 'deposed' by Noun Adjunct (NA). Finally, the noun "رئيس" 'president' and the determiner "ال" 'the' will be linked by Noun Specifier (NS) as shown in figure 8 which is considered as the input for the semantic analysis module.

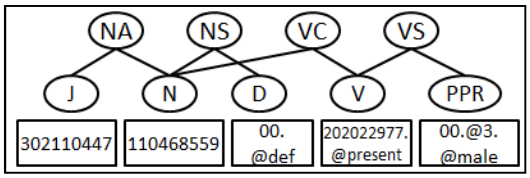

Figure 8: The deep structure of "يلتقي بالرئيس المعزول"

5) Semantic Analysis

In this module, rules have been built to derive the semantic network from the syntactic graph in figure 8. The output of the tree-to-tree phase will be the input of this module or the tree-to-network phase. The syntactic roles should be mapped with semantic relations; accordingly the VS will be mapped with "agent (agt)", the VC with "object (obj)", the NA with "modifier (mod)". Finally the NS between the noun "رئيس" 'president' and the determiner "ال" 'the' will be replaced by the attribute "@def" on the node "رئيس" 'president' as shown in Figure 9.

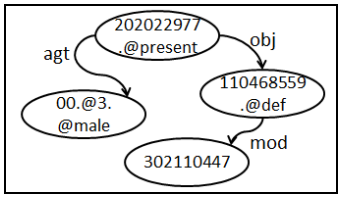

Figure 9: The intermediate representation (UNL Semantic representation)

C. Developing the Generation Grammar

This section is concerned with the NL-ization from the interlingua representation (UNL) into Arabic. Arabic grammar has been especially developed to conduct the generation process. The process of generation may be seen to some extent as a mirror image of the analysis process. Similar to the process of natural language analysis, generating well-formed natural language sentences requires passing through a set of grammar modules which are the lexical mapping module, semantic-syntactic module, the syntactic module, and the morphological generation module.

1) Lexical mapping module

The lexical mapping stage performs the mapping between the meaning conveyed by the concepts of the intermediate representation (UNL interlingua) and the lexical items of the target language through the generation dictionary mentioned in section 4-A-1. For example, the concept '202022977' in figure 9 can be translated into the Arabic verb "التقى" 'meet', the concept '110468559' can be translated into the Arabic noun "رئيس" 'president' and the concept '302110447' can be translated into the Arabic adjective "معزول" 'deposed'. UNL provides a concept for each Arabic word differentiating between the different senses of the words, an option that helps in overcoming the problem of lexical ambiguity during the translation process.

2) Semantic-syntactic Module (network-to-tree)

This module is responsible for mapping the semantic relations to their syntactic equivalents. As an example; the semantic graph generated in figure 9 which represents a verbal phrase that requires mapping rules to map the semantic relations agt, obj and mod to their counterpart syntactic relations; Verb specifier (VS), Verb Complement (VC) and Noun Adjunct (NA). The generated syntactic relations will be processed in the following subsections.

3) The Syntactic Module

The syntactic module is responsible for transforming the deep syntactic structure into a surface syntactic structure. The Syntactic module is divided into two phases; the tree-to-tree phase and the tree-to-list phase. The tree-to-tree phase is responsible for gathering individual syntactic relations and forming higher constituents but the tree-to-list phase is responsible for linearizing the surface tree structure into a list structure, section a and b below will describe the two phases in more details.

- *The tree-to-tree phase*

In the tree-to-tree phase, rules are responsible for building the surface syntactic structure of the sentence by building the intermediate constituents (XBs) which are combined to form the maximal projections (XPs) and combined finally to form the sentence structure. For example, the syntactic relations VS, VC, and NA will be combined to form the maximal projection VP, the specifier of the verb which is "00 .@3.@male" as shown in figure 8 will be projected to the intermediate projection NB then to the maximal projection NP as shown in figure 10, NA between "رئيس" 'president' and "معزول" 'deposed' will be projected gradually to the intermediate projection NB, because the word "رئيس" 'president' was assigned to the attribute @def (definite) " ال"'the' was attached to the tree to represent the specifier of maximal projection NP"الرئيس المعزول" 'the deposed president' according to the schema of X-bar theory as shown in figure 11.

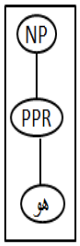

Figure 10: The maximal projection for VS

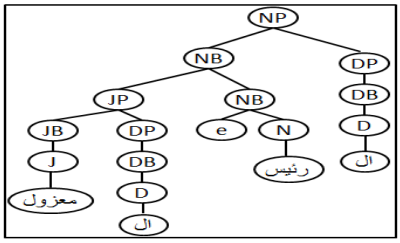

Figure 11: The maximal projection for NA

The preposition “ب”'with' was inserted to the NP constituent in figure 11 to constitute the complement of the main verb “التقى” ‘meet’ as shown in figure 12. The verb complement will in turn be combined with the verb “التقى” ‘meet’ to form the intermediate projection VB “التقى بالرئيس المعزول” ‘he meets the deposed president’ which in turn will be combined with an empty adjunct (e) as show in figure 13 to form a larger intermediate constituent VB. Finally, the resulting VB is combined with the specifier (spec or VS) to build the final maximal projection of the phrase VP as shown in figure 13.

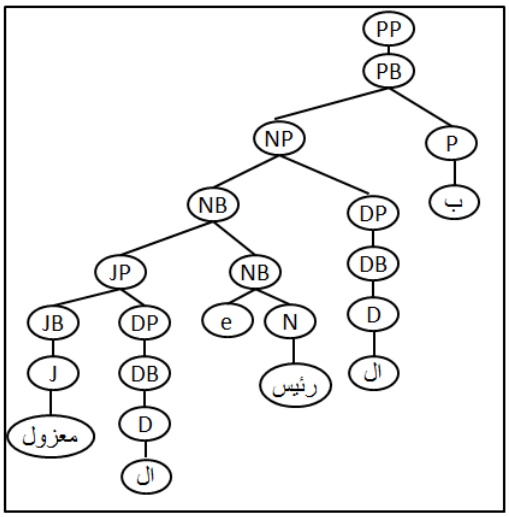

Figure 12: The maximal projection for VC

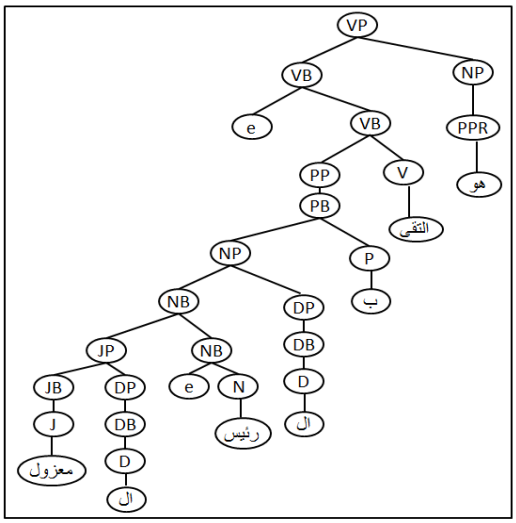

Figure 13: The syntactic tree of final VP

- *The tree-to-list phase*

In the tree-to-list phase, rules are responsible for transferring the surface syntactic structure into a list structure and also adding the required spaces. Moreover, in this stage the specifier of the verb which is the pronoun “هو”‘he’ can be suppressed because its occurrence is preserved in the information stored and assigned to the form of the verb which will be generated in the morphological generation stage.

4) The Morphological Generation Module

This module is responsible for converting the attributes represented in the interlingua into the suitable natural language words or affixes. For example, the attribute @present is represented in the Arabic language by either the present male prefix “ي”or the present female prefix “ت”but here in our example because this attribute is assigned to the verb “التقى” which is related to the male pronoun specifier “هو”, the system can detect that the correct prefix to be generated is the masculine present prefix “ي” not “ت”and finally the Arabic generated sentence will be “يلتقي بالرئيس المعزول”.

6 LILY’S INTERFACE

This section will present LILY`s system interface, describing all its contents, explaining how they are used and the option it provides. The description will be accompanied by screenshots for more elaborate explanation. The interface consists of two textboxes. The one on the left is the input textbox; where the source text is entered. The second one that is on the right, it is called the output textbox; where the target language text will appear. Below these textboxes there are two combo boxes; where the user can choose the source language from the combo box on the right side and the target language from the combo box on the left side as shown in figure 14. In order to obtain the translated text, the user would press the “Translate” button placed under the combo boxes after inserting the text. Besides, LILY system provides two options. The “Fully automatic” option provides the user with the typical translation results, while the “Human interactive” option allows the user to edit and choose the intended meaning for each word. Moreover, this system provides the user with a unique option, since that the user could see the intermediary representation by clicking on the blue link that appears below the output textbox as shown in figure 15. The results will be shown in another textbox. Below the output textbox, four icons will appear with the translation results. The first is responsible for saving the UNL, the second is responsible for saving the NL, the third is a like icon and finally the fourth is a dislike option as shown in figure 15.

Figure 14: The interface of LILY MT

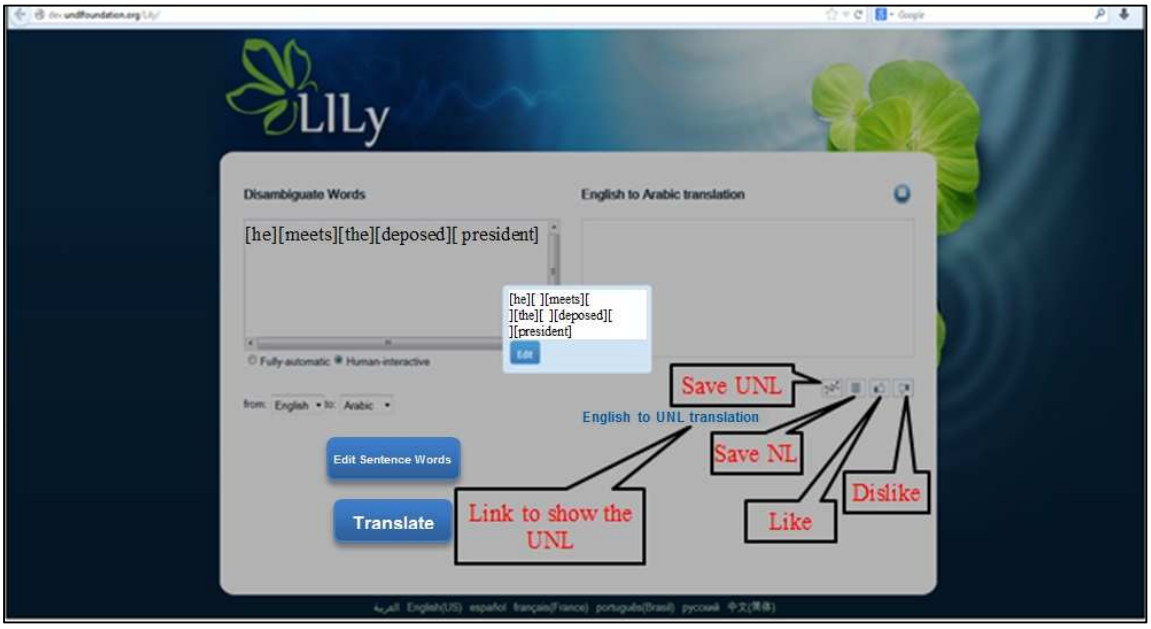

Figure 15: The interface while choosing the Human – interactive option which allows the user to interact with the system and disambiguate the senses, link to appear the UNL and the other four icons

7 EVALUATING THE RESULTS

In this paper we have presented LILY system as a MT system capable of translation between natural languages using an intermediary representation (UNL), we have focused on the Arabic linguistic resources used for the analysis process from Arabic to UNL and the generation process from UNL to Arabic. However, the output of the analysis stage (UNL expressions) has been evaluated using a manually semantically annotated corpus to determine the quality of the automatically generated semantic networks. Precision and Recall were used as evaluation measures. Precision measurement of the semantically annotated Arabic sentences was 0.984 while recall measurement was 0.98.

Likewise, the output of the generation stage; Arabic generated sentences, has been evaluated using a corpus manually translated by specialists. Precision and recall measurements were also used in evaluating the output of the generation stage. Precision measurement of the semantically annotated Arabic sentences was 0.984 while recall measurement was 0.98.

8 CONCLUSION

LILY is an outstanding MT system, to create dynamic equivalence between the translated and original language text and stepping towards achieving the far-fetched dream of computer linguists. In this article, the infrastructure of LILY interlingua MT system is discussed. The Machine Translation components involved in LILY, including the language resources, tools, are presented and they are all provided in an open-source form and for free at www.unlweb.net. As for the intermediary language (UNL), it promises to fulfil the interlingua potential and make the dream of language-independent knowledge representation come true.

REFERENCES

[1] S. Tripath, J. Krishna Sarkhel, *"Approaches to machine translation"*. Annals of library and information studies. Vol.57, PP.388-393, 2010.

[2] E. Nyberg, T. Miltamura, J. G. Cabnell, *"The KANT Machine Translation System: From R&D to Initial Deployment"*. Computer science department – paper 339,1997.

[3] S. Alansary, M. Nagi, N. Adly, "UNL+3: The Gateway to a Fully Operational UNL System", in *Proc. of 10th Conference on Language Engineering*, Cairo, Egypt, 2010.

[4] S. Alansary, M. Nagi, N. Adly, "Processing Arabic Text Content: The Encoding Component in an Interlingual System for Man-Machine Communication in Natural Language,"in *Proc. of 6th Conference on Language Engineering*, Cairo, Egypt, pp. 221-258,2006.

[5] S. Alansary, M. Nagi, N. Adly, "UNL+3: *Understanding Natural Language through the UNL Grammar Work-bench"*. Conference on Human Language Technology for Development (HLTD 2011), Bibliotheca Alexandrina, Alexandria, Egypt, May 2 - 5 2011.

[6] S. Alansary, M. Nagi, N. Adly, "Generating Arabic Text: the Decoding Component in an Interlingual System for Man-Machine Communication in Natural Language," in *Proc. of 6th International Conference on Language Engineering*, Cairo, Egypt, pp. 259-280,2006.

[7] S. Alansary, "A Formalized Reference Grammar for UNL-based Machine Translation between English and Arabic", in *Proc. of 24th international conference on computational linguistics (COLING 2012)*, Mumbai, India, 2012.

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

[9] H. Uchida, M. Zhu, "The Universal Networking Language beyond Machine Translation", UNL Foundation, 2001.

[10] C. Jesús, A. Gelbukh, E. Tovar (eds.), *"Universal Networking Language: advances in theory and applications"*. Mexico City: National Polytechnic Institute, 2005.

[11] H. Uchida, M. Zhu, "UNL2005 for Providing Knowledge Infrastructure," *Proc. Of Semantic Computing Workshop (SeC2005)*, Chiba, Japan, 2005.

[12] M. L. Forcada, "Open-source machine translation: an opportunity for minor languages," *Proc. of Strategies for developing machine translation for minority languages*, 2006.

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