From Language Implicit Structure to UNL Explicit Knowledge Infrastructure

Sameh Alansary†* **Magdy Nagi‡***
*Bibliotheca Alexandrina, Alexandria, Egypt
†Phonetics and Linguistics Department, Faculty of Arts, Alexandria University, Alexandria, Egypt
‡Computer and System Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt
†sameh.alansary@bibalex.org
‡magdy.nagi@bibalex.org

Abstract

Despite the focused attention of improvements achieved by the NLP community on various language-related issues of knowledge representation, knowledge representation has not been satisfactorily achieved. This paper focuses on five axes. The first axis deals with presenting a structured representation of linguistic knowledge of natural language sentences through the Universal Networking Language (UNL) framework. The second axis deals with applying the UNL knowledge representation system on Arabic examples illustrating how effective the UNL system in representing Arabic morphology, syntax, semantics and pragmatics. The third axis highlights the available corpora of the UNL. The fourth axis highlights the potential applications of the current knowledge representation system in the fields of information extraction, summarization and other applications that seek an understanding of natural language input. The fifth and final axis deals with the evaluation of the UNL system output.

Keywords: knowledge Representation, knowledge infrastructure, UNL, semantic representation.

1 Introduction

When computers make intelligent processing based on knowledge including reasoning, enough knowledge must be provided in a form that is accessible and understandable for computers. The collection of this knowledge is called “Knowledge Infrastructure” (KI) which is a compilation of the representation of knowledge on many levels.

Knowledge representation (KR) aims at representing knowledge in symbols in order to facilitate inference [1]. But what makes a representation of linguistic knowledge adequate; in other words, what type of information should a knowledge representation contain? There are some basic characteristics that distinguish a good knowledge representation: good coverage, comprehensibility, consistency, efficiency, facility of modification and updating, preciseness and unambiguity and explicitness

Natural languages are the most powerful knowledge representation languages that exist. It serves as their own meta-language and has a greater flexibility and expressive power than any formal (artificial) language. But, using natural languages for knowledge representation have some serious disadvantages; they are difficult for computers to process because of their inherent ambiguity and vagueness, and their expressive power turns out to be one of the greatest obstacles for automatic reasoning [2]. This is due to the gap between computer and human in the interpretation of any natural language sentence. The human has the ability to understand natural languages and do good linguistic knowledge representation while computer cannot, because of the nature of natural language. So it is important to provide computers with knowledge in explicit way. Formal languages have been used instead of natural languages because these languages have an unambiguous syntax and clear semantics and help to avoid errors in the interpretation of the represented knowledge.

This paper introduces one of the approaches to Knowledge Representation; the Universal Networking Language (UNL). The next sections demonstrate how the UNL can represent different levels of linguistic knowledge simultaneously or separately starting with the morphological level and ending with the pragmatic level. They also demonstrate how such knowledge can be utilized for intelligent processing by computers.

Section 2 discusses the UNL system development and how it is used to represent knowledge. Section 3 explores the use of UNL in representing Arabic knowledge infrastructure on different linguistic levels, illustrated with Arabic examples and an experiment on Arabic sentences. Section 4 surveys the different corpora within the UNL framework. Section 5 examines some intelligent applications that can make use of such representation. Finally, section 6 evaluates the output of the UNL system.

2 Representing Knowledge infrastructure using UNL

After discussing some basic characteristics that distinguish a good knowledge representation, this section will present a knowledge representation approach based on the UNL framework illustrating how the UNL formalism can achieve the aforementioned characteristics. The first subsection presents a brief description of the UNL system and its development. The second discusses how UNL represents the linguistic knowledge of natural sentences using its components. Finally, the third subsection discusses the different linguistic resources and engines of the UNL system that facilitate the automatic representation of linguistic knowledge.

2.1 A brief introduction to the UNL system and its development.

UNL stands for the Universal Networking Language; an artificial language for computers created to represent and process information and knowledge across language barriers. The UNL is first and foremost a knowledge representation language, it is not intended to describe or represent natural languages; but rather to represent the information conveyed by natural languages. The UNL system is an interlingua-based framework that aims at facilitating the semantic processing of natural language using computers [3], [4], [5]. The UNL program was launched in 1996 as a communication protocol, originally proposed by the Institution of Advanced Studies of the United Nations University, Tokyo, Japan under the auspices of the UNESCO to be a resource for developing a multilingual platform for information exchange. In January 2001, the United Nations University set up an autonomous non-profit organization in Geneva, Switzerland to be responsible for the development and management of the UNL Program; the Universal Networking Digital Language UNDL Foundation (http://www.undl.org/). 16 different languages joined the project until now, with each responsible for the development and maintenance of the components of its respective language module. Since 1996, the UNL program has passed through many phases of development and enhancement and crossed important milestones. More information about the earlier system can be found in [4], [6], [7] and [8].

2.2 How can UNL represent linguistic knowledge

UNL uses components similar to those of natural languages. Linguistic knowledge can be formally represented through a semantic network made up of three different types of discrete semantic units, universal words (UWs), semantic relations, and attributes. This three-layered representation model is the cornerstone of the UNL.

Universal Words (UWs): These are the words of UNL; they are used to express the meaning of any concept, and therefore, the whole set of UWs are supposed to include the entire set of concepts known to man. The UNL assumes that UWs must correspond to and only to semantic discrete units conveyed by the natural language open lexical categories (nouns, verbs, adjectives and adverbs). Any other semantic content (conveyed by articles, prepositions, conjunctions etc.) should be represented as attributes or relations. For instance, the UWs of the sentence in (1) are the concepts expressing the meanings of the words “man”, “buy”, “beautiful”, and “car”. Currently, the UNL system uses a numerical ID to refer to concepts. These IDs are extracted from the English WordNet 3.0. The use of Word Net is discussed in [9], [10] and [11]. Hence, the numerical IDs referring to the concepts of sentence (1) are 202207206 which refers to the verb ”buy”, 110287213 which refers to the noun “man”, 102958343 which refers to the noun “car”, and 300217728 which refers to the adjective “beautiful”.

(1) The man bought beautiful cars.

Semantic Relations: The semantic relations constitute the syntax of UNL. They are three-letter symbols such as agt (agent), obj (object), ins (instrument), etc. Relations are organized in a hierarchy. This hierarchy contains four general relations: participant (ptp), for the necessary arguments (subject and complements) of verbal predicates; attribute (aoj), for the necessary arguments (subject and complement) of nominal predicates; specifier (mod) for general specifiers and finally adjunct (adj) for general adjuncts, including time, location and manner. (http://www.unlweb.net/wiki/Relations). For instance, the semantic relation between the UWs of the sentence in (1) above are, an “agt” relation between the UW (202207206) “buy” and the UW (110287213) “man”, an “obj” relation between the UW (202207206) “buy” and the UW (102958343) “car”, and a “mod” relation between the UW (102958343) “car” and the UW (300217728) “beautiful” as shown in figure (1).

Attributes: Are additional tags used to represent the information conveyed by natural language grammatical categories (such as tense, mood, aspect, number, etc). They are used to further modify the semantic network and add information that is not expressed via UWs or semantic relations. The names of attributes are always expressed in lower case letters and its syntax is defined as follows: <attribute> = “@”<attribute name>. Attributes convey three different types of information: First, information about the role of the node in the UNL graph, as in the case for ‘@entry’ which is assigned to the main node of a UNL graph. For example, the UW (202207206) which refers to the word “buy” in sentence (1) is the entry of the sentence, hence it was assigned the attribute ‘@entry’. Second, the information conveyed by bound morphemes and closed classes such as affixes (gender, number, tense, aspect, mood, voice, etc), determiners (articles and demonstratives), adpositions (prepositions, postpositions and circumpositions), conjunctions. For example, the UW (102958343) which refers to “car” will be assigned the attribute “@pl” because it is in the plural form to be: 102958343.@pl , and the UW (202207206) which refers to the verb “buy” will be assigned the attribute “@past” because the verb is in the past tense to be: 202207206.@past. Third and finally, attributes convey information on the (external) context of the utterance, i.e., non-verbal elements of communication, such as prosody, sentence and text structure, politeness, schemes, social deixis and speech acts. For example, in sentence (1), the attribute “@surprise” is assigned to the entry of the sentence as shown in figure 1.

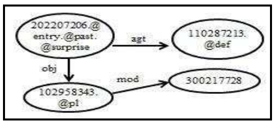


Figure 1. The UNL semantic graph representing sentence (1)

2.3 UNL Linguistic resources and engines

In addition to the three main components of the UNL that enable the knowledge representation in the form of a semantic graph and a hyper graph, the UNL system can also has other components that aid the knowledge representation; these are the linguistic resources, tools and engines. The UNL System comprises three different types of language resources: lexica, grammars and corpora, in addition to tools for manual knowledge representation and engines for automatic representation. All of these have been provided mainly through the UNLweb[1]. This section will discuss these linguistic resources and some of UNL engines and tools developed to make use of these resources.

2.3.1 Linguistic resources

Lexica

The UNL System contains three different types of lexical databases: dictionaries, knowledge bases and memories. There are three different types of the dictionaries: The UNL Dictionary is a flat list of UWs and their corresponding semantic features, the NL Dictionary is a list of natural language entries and their corresponding features and the UNL-NL Dictionary is a list of lexical mappings between UNL and a given natural language. There are also three different types of memories; the UNL Memory is a network of necessary and typical interactions between UWs, the NL Memory is a list of typical interactions between natural language entries and the UNL-NL Memory is a list of mappings between UNL and a given natural language. There is also the UNL Knowledge Base which is a network of necessary interactions between UWs.

Grammars

Grammars are the formalizations required to transform natural language sentences into UNL representation and vice versa. Grammars require an accurate understanding of the human language, thus, the UNL system provides the tools and methods capable of effectively decomposing the

[1] (http://www.unlweb.net/)

sentence into its basic constituents, understanding and encoding in some formal manner the intended meaning behind each constituent and the meaning reflected by its superficial grammatical form as well as its position in the sentence. Moreover, the semantic relation between each constituent and the others should also be understood and encoded. UNL grammars are divided into two main types: transformation rules and disambiguation rules; transformation rules are the rules responsible for transforming natural language into UNL while disambiguation rules are used to prevent wrong lexical choices, provoke best matches and check the consistency of the graphs, trees and lists.

Corpora

There are two types of corpora; the UNL Corpus and the NL Reference Corpus; the UNL Corpus which is composed of the documents written in UNL and provided according to the UNL document structure. The UNL corpus is used as a guide for developers to demonstrate how the linguistic knowledge of a language can be represented through UNL. The NL Reference Corpus is the corpus used to prepare and to assess grammars for sentence-based UNLization[2]. More information can be found in [12].

2.3.2 Engines and Tools

The UNLdev is the wrapper application for the development of various UNL tools and applications. There are tools for professional users (linguists, computational linguists) and non-professional users. These tools are programming software and they differ from UNL applications in that they are not tailored for non-specialists and require fair expertise in UNL. Most of these tools are shareware. The UNLdev includes three analysis software; IAN (the Interactive ANalyzer), the UNL Editor and SEAN (Shallow Enhanced ANalysis system). IAN is a tool for the semi-automatic (interactive) analysis of natural language texts and the generation of UNL Documents, 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, the UNL Editor and SEAN are language-independent and are parametrized to the natural language of the input through a dictionary and a grammar that are provided as separate interpretable files.

The UNLdev also includes the generation software EUGENE; The dEep-to-sUrface natural language GENErator. EUGENE is fully automatic; it receives a UNL input and delivers an output in natural language without any human intervention. The UNLdev also includes other tools that can be used in different applications[3].

3 Using UNL in the representation of Arabic knowledge infrastructure.

After discussing how the UNL can represent knowledge through its components, linguistic resources and engines, this section will discuss the implementation of the UNL approach on the representation of Arabic knowledge infrastructure. The greatest challenge in the field of knowledge representation is not gathering knowledge, but rather representing and structuring it in an adequate manner that can be searched efficiently, or used to infer further knowledge. These goals in their essence correspond to the task of natural language analysis. The main goal of language analysis is to obtain a suitable representation of text structure and facilitate the task of processing texts based on their contents.

As mentioned earlier, UNL represents information and knowledge in the form of a UNL graph. UNL analyzes natural language texts on various levels; morphological, syntactic, semantic and pragmatic levels, starting with the surface representation to deeper ones in order to make the structure of input sentences explicit as shown in figure 2.

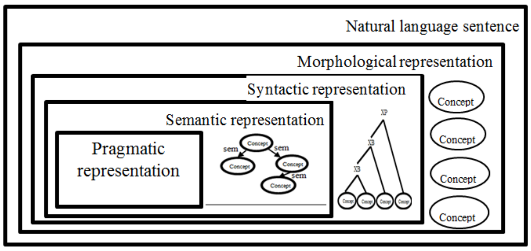


Figure 2. Linguistic levels of knowledge representation

3.1 Representation of morphological knowledge.

In order to represent morphological knowledge

[2] For more information see http://www.unlweb.net/wiki/List_of_UNL_Corpora)

[3] (http://dev.undlfoundation.org/index.jsp)

using UNL; morphological analysis of input sentences is performed. Tokenization (text segmentation) is an important step before morphological analysis. It is a dictionary-based process; the strings of the natural language input are matched against the entries existing in the dictionary and the corresponding UWs are extracted.

Arabic is a highly inflectional language; in Arabic, the subject of verbs may be expressed in the verb form itself in the form of a suffix. For example, the suffix “ون” attached to the verb form “يدرسون” “they are studying” expresses the plural masculine subject “هم” ”they”, the suffix “ن” attached to the verb form “تدرسن” “they are studying (feminine)” expresses the plural feminine subject “هن” “they”. The inflectional form of the verb “تدرسن” in the dictionary in (a) expresses both the concept of the verb itself and the concept of the subject.

a) [تدرسن] {} “UW1” (verb, active, present, plural, 3rd person, feminine) <Ar,0,0>;

Moreover, in the dictionary, grammatical attributes are assigned to each word form; these grammatical attributes provide information about the person, gender, number of the subject of verbs and also information about the voice and tense of verbs as in “تدرسن” as well as the gender and number of nouns. Thus, in order to represent the knowledge in Arabic sentences, morphological analysis rules should interpret the grammatical attributes assigned to the word forms and transform them into UNL format, as in figures 3 and 4:

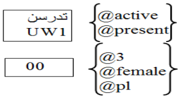


Figure 3. The UNL attributes of “تدرسن”

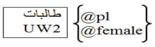


Figure 4. The UNL attributes of “’طالبات”

In figure 3, the attribute “@present” refers to the present tense and “@active” refers to the active voice. The morphological analysis also makes the concept of the subject explicit and expresses it in UNL format as “00” and assigns to it the necessary UNL person, number and gender attributes: @3, @female, and @pl respectively. In figure 4, the attribute “@pl” refers to the plural number while @female refers to the feminine gender.

3.2 Representation of syntactic knowledge.

The purpose of syntactic analysis is to determine the structure of the input text. UNL can represent the surface as well as the deep structures of a sentence. Syntactic analysis on the surface level is a first step towards the deep level.

Surface structure analysis

The surface structure analysis phase is responsible for transforming the list structure (natural sentences) into a tree structure. In this phase, small constituents or trees are constructed for the small phrases (usually noun phrases) in the sentence and then combined to form a bigger tree gradually until the whole sentence is analyzed. For instance, when analyzing a sentence such as“أكل محمد تفاحة كبيرة عندما جاء أحمد” ‘Mohammad ate a big apple when Ahmad came’, the nodes of the smallest constituents in the sentence will be linked first. This sentence is composed of two smaller phrases; the phrase “أكل محمد تفاحة كبيرة” ‘Mohammad ate a big apple’ and the phrase “عندما جاء أحمد” ‘when Ahmad came’. Thus, each will be analyzed individually before being connected.

First, the nouns “محمد” ‘Mohammad’ and “تفاحة” ‘apple’ will be projected to the intermediate constituent “NB” as they are heads of noun phrases. Then, the adjective “كبيرة” ‘big’ will be linked to the intermediate constituent “NB””تفاحة” ‘apple’ to build a bigger “NB”. As there are no other modifiers for the constructed “NBs” “تفاحة كبيرة” ‘big apple’ and “محمد” ‘Mohammad’, they will be projected to the corresponding maximal projection “NP”.

These “NPs” are preceded by a verb in the Arabic input sentence. These two “NPs” are the arguments for the verb “أكل” ‘ate’ which is a transitive verb. The first NP “محمد”‘Mohammad’ is the specifier (VS) of the V “أكل” ‘ate’ and the other “NP” “تفاحة كبيرة” ‘big apple’ is the verb’s complement (VC).

The Arabic follows the word order “VSO”, however, according to the X-bar theory[4], the order of constituents in the input sentence cannot be changed. In order to solve this problem, the following steps were carried out as shown in figure 5: Firstly, the NP “تفاحة كبيرة” ‘big apple’ (which is not followed by a verb) will be linked to an empty node in order to form the intermediate constituent “VB”. Secondly, the NP “محمد” ‘Mohammad’ will be linked to the newly formed “VB” to form the maximal projection “VP”. Thirdly and finally, since the verb “أكل” ‘ate’ is tensed, it will be an “IP” or Inflectional Phrase and will be placed in the “I” position as shown in figure 5 and the “VP” and the “I” will be linked together to build the

[4] http://www.unlweb.net/wiki/X-bar_theory

intermediate constituent “IB”. Since there are no more constituents in this “IB”, the “IB” will be projected directly to an “IP” [13].

In the second phrase “عندما جاء أحمد” , “جاء أحمد” ‘Ahmad came’ will be treated in the same manner“أكل محمد” ‘Mohammad ate’ has been treated; as an “IP”. Then, the “IP” “جاء أحمد” ‘Ahmad came’ and the adverb “عندما” ‘when’ will be linked together to build the adverbial phrase “AP”, as shown in the figure 5. The Adverb “عندما” ‘when’ acts as a conjunction between the two parts of the sentence so the “IP” will be projected to the intermediate constituent “CB”. Finally, the “CB” “أكل محمد تفاحة كبيرة” ‘Mohammad ate a big apple’ and the “AP” “عندما جاء أحمد” ‘when Ahmad came’ which is the specifier of the “CP” will be linked together to form the maximal projection “CP” as shown in figure 5.

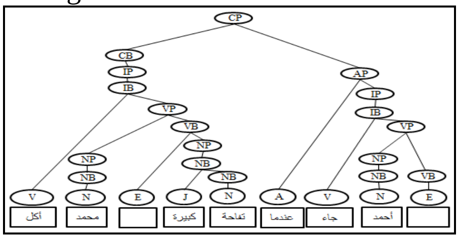


Figure 5. The surface structure tree of “ أكل محمد تفاحة كبيرة عندما جاء أحمد"

Deep structure analysis

The task of the deep structure analysis phase is to restore the dependency relations between constituents that are isolated in the surface structure. This task can be carried out over three steps: movement, de-arborization and re-categorization. The first step will be achieved over the tree in figure ٥; the verb “أكل” ‘ate’ and “جاء” ‘came’ will be moved to the empty position of “V” in the “VP” and consequently, the position of “I”; in tree in figure 6 as “V” will be left empty. The empty nodes will be deleted after movement as shown in figure ٦.

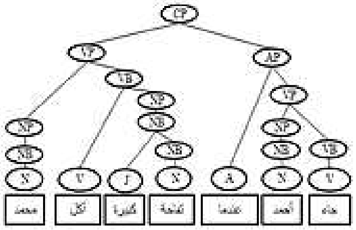


Figure 6. The syntactic structure of “ أكل محمد تفاحة كبيرة عندما جاء أحمد” after movement.

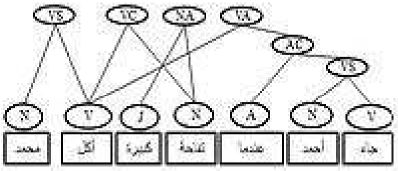


Figure 7. The head-driven structure of “ أكل محمد تفاحة كبيرة عندما جاء أحمد”.

The second step is de-arborization which is the step responsible for converting the tree structures into head-driven structures. The maximal projection “CP” will be de-arborized to the “VP” (أكل محمد تفاحة كبيرة) ‘Mohammad ate a big apple’ and the “VA”, between the head of the “VP” (أكل) ‘ate’ and the adverbial phrase “AP” as one unit (عندما جاء أحمد) ‘when Ahmad came’ ((أكل، (عندما جاء أحمد)) ‘(ate and (when Ahmed came)). The maximal projection “VP” will be de-arborized into the “VB”“أكل تفاحة كبيرة” ‘ate a big apple’ and the “VS” “أكل محمد” ‘Mohammad ate’. The “AP” will be de-arborized into “AC” between the adverb “عندما” ‘when’ and the “VP” “جاء أحمد” ‘Ahmad came’. Now, we have a “VS” between (أكل، محمد) ‘Mohamad ate’, a “VA” between (أكل، عندما جاء أحمد) ‘eat, when Ahmed came’, a “AC” between (عندما، جاء أحمد) ‘when, Ahmed came’, “VB” between (أكل,تفاحة كبيرة) ‘ate, big apple’ and a “VP” between (جاء, أحمد) ‘Ahmed, came’, and an “NB” between (تفاحة،كبيرة) ‘big , apple’ as shown in figure 7.

The third step is re-categorization which is the step responsible for transforming the arborized constituents which are “NB”, “VB” and “VP” to the syntactic roles “NA”, “VC” and “VS” as shown in figure 7. The “NB” (تفاحة كبيرة) will be transformed into a “NA”; noun adjunct as the adjective is an optional modifier. The “VB” (أكل تفاحة) ‘ate an apple’ will be transformed to a “VC”; as the noun “تفاحة” ‘apple’ is a thing which can be eaten. The other “VP” (جاء أحمد) ‘Ahmad came’ will be transformed to a “VS”; as the verb “جاء” ‘came’ requires a human to do the action.

3.3 Representation of semantic knowledge.

The fourth level of linguistic representation is the level of semantic representation. This level is responsible for mapping the syntactic roles in figure 7 with their equivalent semantic relations or UNL attributes. The head-driven structure in figure 7 contains six syntactic roles; two “VSs”, a “VC”, a ”NA”, a “AC” and a “VA”. Since the verb “أكل” ‘ate’ is a transitive verb that requires an agent and an object, the syntactic roles VS and VC will be respectively mapped to agent (agt) and object (obj) semantic relations. Then, the “NA” between the noun “تفاحة” ‘apple’ and the adjective “كبيرة” ‘big’ will be mapped to a modifier (mod) relation. The “AC” between the adverb “عندما” ‘when’ and the “VS”(جاء أحمد) ‘Ahmad came’ will be deleted and the adverbial time attribute will be assigned to the “VS” (جاء أحمد) ‘Ahmad came’. The “VA” of “ أكل (جاء أحمد)” ‘ate (Ahmad came)’will be mapped with a (tim) semantic relation to refer to the time of eating as the constituent (جاء أحمد) ‘Ahmad came’ carries an adverbial time attribute. It can be noticed that the (tim) relation is between “أكل” ‘ate’ and (جاء أحمد) ‘Ahmad came’ not between “أكل” ate‘’ and “جاء” ‘came’ as in the other relations in the sentence this is because (جاء أحمد) ‘Ahmad came’ is an adverbial phrase which is represented in the UNL as a hyper graph or scope as discussed in section 3. Figure 8 represents the final semantic graph.

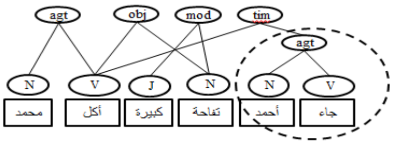


Figure 8. The semantic graph of “ أكل محمد تفاحة كبيرة عندما جاء أحمد”

3.4 Representation of pragmatic knowledge.

The last level of representation is the pragmatic level. UNL has the ability to represent information about the external context of the utterance via attributes as mentioned in section 3. For example: “@polite” is used in polite requests such as “ من فضلك افتح الباب” ‘please, open the door’ or for showing respect to a person from a higher class as in “سيادتكم” ‘your excellency’. However, progress in this level is still behind the previous ones.

After discussing the representation of knowledge on different linguistic levels we can summarize all of the above mentioned illustrations in the template of linguistic knowledge representation that was mentioned earlier in figure 2. Figure 9 shows a natural language sentence “افتح الباب بسرعة”‘open the door quickly’ represented explicitly in the different levels of linguistic knowledge; morphological, syntactic, semantic and pragmatic levels.

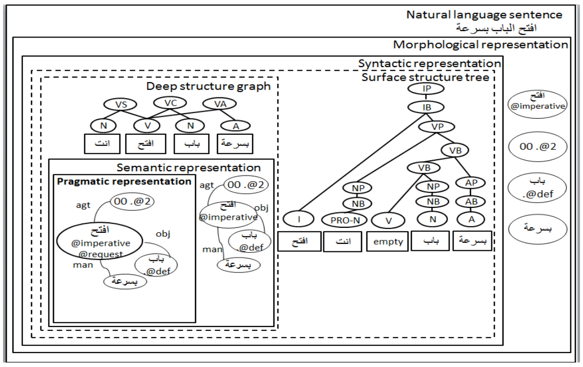
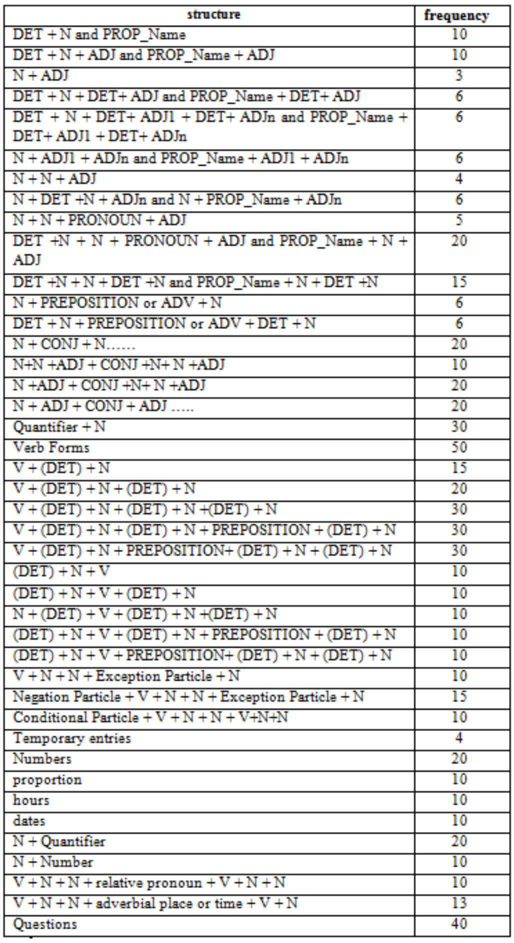


Figure 9. The linguistic knowledge representation of “افتح الباب بسرعة”

The above example “ أكل محمد تفاحة كبيرة عندما جاء أحمد” is an experiment on one of the 600 sentences collected to cover 42 Arabic structures as shown in figure 10. The structures have been collected from the Arabic books of syntax.



structure	frequency
DET + N and PROP_Name	10
DET + N + ADJ and PROP_Name + ADJ	10
N + ADJ	3
DET + N + DET+ ADJ and PROP_Name + DET+ ADJ	6
DET + N + DET+ ADJ1 + DET+ ADJn and PROP_Name + DET+ ADJ1 + DET+ ADJn	6
N + ADJ1 + ADJn and PROP_Name + ADJ1 + ADJn	6
N + N + ADJ	4
N + DET +N + ADJn and N + PROP_Name + ADJn	6
N + N + PRONOUN + ADJ	5
DET +N + N + PRONOUN + ADJ and PROP_Name + N + ADJ	20
DET +N + N + DET +N and PROP_Name + N + DET +N	15
N + PREPOSITION or ADV + N	6
DET + N + PREPOSITION or ADV + DET + N	6
N + CONJ + N......	20
N+N +ADJ + CONJ +N+ N +ADJ	10
N +ADJ + CONJ +N+ N +ADJ	20
N + ADJ + CONJ + ADJ	20
Quantifier + N	30
Verb Forms	50
V + (DET) + N	15
V + (DET) + N + (DET) + N	20
V + (DET) + N + (DET) + N +(DET) + N	30
V + (DET) + N + (DET) + N + PREPOSITION + (DET) + N	30
V + (DET) + N + PREPOSITION+ (DET) + N + (DET) + N	30
(DET) + N + V	10
(DET) + N + V + (DET) + N	10
N + (DET) + V + (DET) + N +(DET) + N	10
(DET) + N + V + (DET) + N + PREPOSITION + (DET) + N	10
(DET) + N + V + PREPOSITION+ (DET) + N + (DET) + N	10
V + N + N + Exception Particle + N	10
Negation Particle + V + N + N + Exception Particle + N	15
Conditional Particle + V + N + N + V+N+N	10
Temporary entries	4
Numbers	20
proportion	10
hours	10
dates	10
N + Quantifier	20
N + Number	10
V + N + N + relative pronoun + V + N + N	10
V + N + N + adverbial place or time + V + N	13
Questions	40

Figure 10. The Arabic corpus (600 sentences covers 42 structures).

4 Available UNL Corpora

Several corpora have been built using the UNL strategy mentioned earlier. In 2012, the UNL

Reference Corpus (UC) was built; the corpus has been used to prepare and to assess grammars for sentence-based NLization.

In 2010, the project IGLU was built, it aims at UNL-izing the definitions of 27,255 entries extracted from an abridged version of the WordNet3.0. Results are expected to be incorporated into the UNL Knowledge Base which codifies the most systematic part of the meaning conveyed by natural language words, and to constitute a UNL-ization memory to be used in future mappings between English and UNL. IGLU contains 30,342 distinct sentences and 141,577 open-class tokens corresponding to 27,255 entries of the WordNet3.0. The corpus can be exported and downloaded from the UNL$^{\text{arium5}}$.

Another UNL corpus has also been built with the aim of translating into the Universal Networking Language (UNL) the integral text of *Le Petit Prince* (a French novella).

In 2005, a corpus of the UNL Documents of 30 articles from the Encyclopedia of Life Support Systems have been compiled. These UNL Documents are provided by the UNL Center of UNDL Foundation for developers to improve their language NLization tools. The UNL expressions in this corpus was produced by a semi-automatic process[6].

In 2004, the Cratylus Project aimed at translating, manually, from English into UNL, the integral text of *Cratylus* written by the Greek philosopher Plato.

Several other UNL corpora were built such as the UNESCO corpus in 2003, the UNL News corpus in 2002, the International Telecommunication Union (ITU) corpus in 2001, the Great Barrier Reef in 1999, the World Cup History corpus in 1998 and The Tower of Babel corpus in 1997.

5 UNL Applications

After achieving an accurate knowledge representation, many potential applications can be developed such as information extraction, interlingua-based machine translation, cross-language search, summarization and rephrasing systems. Each one of these will be described in more detail in the following.

Information extraction: The structured representation of knowledge achieved and illustrated in figure (9) can serve as a good environment for the information extraction processes. Such representation allows for building structured knowledge from unstructured texts which is one of the uses of information extraction systems.

Interlingua-based machine translation: The most obvious use of the Universal Networking Language is interlingua-based translation. By converting the source language into a language-independent knowledge-based format, UNL enables the generation of an equivalent source text in any natural language. This depends mainly on the existence of robust dictionaries and grammars of both the source and target languages.

Cross-language search: Another use that is equally important is the use of knowledge representation in order to search contents and retrieve the most relevant results to the user's query. A search engine that is based on semantic representation can understand what the user is searching for and search all web pages written in any language and retrieve the results in the user's native language.

Text summarization: By representing the linguistic knowledge in such a way, we can determine the main clauses or words carrying the content of a sentence and the secondary or modificatory ones.

Text rephrasing: The UNL system can generate an output that is not different from the input in language but rather in length, complexity or stylistic choices. Of course the output can also be different in language if the user wishes so.

6 Evaluation

The output of the UNLization process for Arabic language has been evaluated using a manually semantically annotated corpus in order to determine the quality and correctness of the automatically generated semantic networks which should convey exactly the intended meaning of the natural language sentences. The data set contained 600 Arabic sentences. Precision and Recall were the evaluation measures used to evaluate the output. Precision measurement of the UNLized Arabic sentences was 0.983 while recall measurement was 0.979.

7 Conclusion

The UNL system is considered a good approach to represent knowledge of different linguistic levels and build knowledge infrastructure that can be

[5] UNLWEB>UNLARIUM>CORPUS>IGLU>EXPORT

[6] This corpus can be accessed at (http://www.undl.org/unl-eolss/unldoc.html).

used in the applications of intelligent text processing. Knowledge representation using the UNL have been applied to 600 Arabic sentence and the representation of different levels was illustrated starting from the morphological level passing through syntactic and semantic levels and ending in pragmatic level. The output has been evaluated precision and recall, Precision was 0.983 while recall measurement was 0.979. The paper ended by some intelligent applications that can be built depending on the knowledge representation.

References

[1] P. Martin, 2002. Knowledge representation in RDF/XML, KIF, Frame-CG and Formalized-English, Distributed System Technology Centre, QLD, Australia.

[2] J. F. Sowa, 2000. Knowledge Representation – Logical, Philosophical and Computational Foundations. Brooks/Cole .

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

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

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

[6] I. Boguslavsky, J. Cardeñosa, C. Gallardo, L. Iraola, 2005. The UNL Initiative: An Overview. CICLing 2005, pp.377-387.

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

[8] S. Alansary, M. Nagi, N. Adly, 2008. Machine Translation Using the Universal Networking Language (UNL) , 8th International Conference on Language Engineering, Ain Shams University, Egypt.

[9] V. Dikonov, I., 2008. Boguslavsky, Universal Dictionary of Concepts // MONDILEX First Open Workshop “Lexicographic Tools and Techniques“, Moscow. pp. 31-41.

[10] S. Boudhh, P. Bhattacharyya, 2009. Unification of Universal Words Dictionaries using WordNet Ontology and Similarity Measures, Seventh International Conference on Computer Science and Information Technologies (CSIT 2009), Yerevan, Armenia.

[11] R. Martins, V. Avetisyan, 2009. Generative and Enumerative Lexicons in the UNL Framework, Seventh International Conference on Computer Science and Information Technologies (CSIT 2009), Yerevan, Armenia.

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

[13] A. Carnie, 2002. Syntax: A Generative Introduction, 2002, Blackwell Publishers, part 3, chapter 8.