Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas

Sameh Alansary

Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Alexandria, Egypt.

Bibliotheca Alexandrina, Alexandria, Egypt.

Sameh.alansary@bibalex.org

Abstract*: Interlingua-based machine translation is probably the most attractive among the three classic approaches to MT. Early pioneers as well as current researchers experimented with this approach and produced some very stimulating methodologies to reaching such a language-independent framework. In this paper, we shall briefly review some of the most renowned endeavours in interlingua-based machine translation and bring into view how the latest of which; the Universal Networking Language (UNL) differs and compares to these other systems.*

Key words: *machine translation, interlingua, interlingua-based machine translation, universal networking language, UNL.*

1 INTRODUCTION

Generally, three classic approaches have been acknowledged in the field of Machine Translation; Direct, Transfer and Interlingua. The Direct approach is mainly a lexicon-based approach in which a computer program performs a word-for-word substitution (with some local adjustment) between language pairs using a large bilingual dictionary[1], [2].

The Transfer approach operates over three stages: analysis, transfer and generation. First, the SL text is parsed into a source-language-specific intermediate syntactic structure. Then, linguistic rules specific to the language pair transform this representation into an equivalent representation in the target language. Finally, the final target language text is generated [3].

The Interlingua approach, on the other hand, is based on "the argument that MT must go beyond purely linguistic information (syntax and semantics) and involve an 'understanding' of the content of texts"[1], [2]. Interlingua-based translation is divided into two monolingual components: analyzing the SL text into an abstract universal language-independent representation of meaning (the interlingua), and generating this meaning using the lexical units and the syntactic constructions of the target language.

2 INTERLINGUA: DEFINITIONS AND CHARACTERISTICS

The motivation behind devising an interlingua was the long-lived belief that while languages differ greatly in their "surface structures", they all share a common "deep structure". Hence arose the idea of creating a universal representation capable of conveying this deep structure while enjoying the regularity and predictability natural languages lack.

In order to be capable of representing natural language content, an interlingua should be, first, unambiguous; it should be more explicit even than the natural language it is representing [1]. Second, it should represent the full content of the input text; its morphological, syntactic, semantic and even pragmatic characteristics [4]. Third, it should be universal, capable of representing the abstract meaning of any text, belonging to any domain or language. Fourth, an interlingua should represent the content of the input alone and not be influenced by the formal representation of the content in the SL text [5]. Fifth and finally, the interlingua should be independent of both the SL and the TL; analysis should be SL-specific and not oriented to any particular TL, and likewise should be the generation [6].

The advantages of using such an approach include economy, modularity, localization, back-translation possibility and potential uses in other NLP-related areas such as cross-lingual information retrieval, summarization, rephrasing and question answering[4], [1],[5] –[7].

3 SOME WELL-KNOWN INTERLINGUA-BASED SYSTEMS

Despite its numerous advantages, the interlingua approach is probably the least used among the three classic approaches. However, many research projects have produced quite promising prototypes. The following section briefly reviews four of the most renowned interlingua-based machine translation projects.

A. DLT

DLT stands for Distributed Language Translation, a research project developed in Utrecht, The Netherlands. Preliminary research in the project began as early as 1979. In 1984, DLT entered a six-year project to build an MT system capable of translating from simplified English into French. However, in 1990, the DLT pilot project came to an end[8] after receiving a fair amount of publicity.

DLT is an interactive system developed to operate over computer networks. Translation is distributed between two independent terminals; one for the analysis and another for generation. In the DLT system, the intermediate representation (the interlingua) is a ready-made logical language with supposedly standardized rules for vocabulary and structures; i.e. Esperanto.

Semantic and Pragmatic knowledge constitute the language-independent component of the system is completely handled in the intermediate stages of forming the Esperanto representation. Language-specific information, on the other hand, is purely syntactic and is developed for a specific pair of languages, in one translation direction only; from English to Esperanto, for instance [1].

The text entered at one terminal is syntactically parsed into dependency trees. In case of syntactic ambiguity, the parser produces all possible alternative trees regardless of their semantic probability[9]. “The result is a (sometimes large) number of 'formally possible parallel translations” [10]. Then, rules replace SL words with their Esperanto equivalents (all possible alternatives), and English syntactic labels with Esperanto ones.

So far, all candidate parses are equally probable. To choose one, first, the system consults the Lexical Knowledge Bank (LKB) which is a database containing pairs of content words linked by a connector (see figure 1)[1]. Its role is to indicate which word in most likely to appear in the given context.

ĉambro a hela	‘light room’
ĉambro a komforta	‘comfortable room’
ĉambro a komuna	‘common room’
ĉambro a nuda	‘bare room’
etc.	

Figure 1: A sample from DLT's Lexical Knowledge Base (LKB)

If no exact match was found in the LKB, an algorithm called SWESIL ranks the possible alternatives according to their semantic proximity. If, after all, the system was not able to conclusively choose one by itself, a machine-initiated disambiguation dialogue presents the operator, in his native language, with the phrases or sentences requiring disambiguation, on which he/she may choose one of the possible interpretations listed on the screen [10]. Finally, the chosen tree is regularized and linearized into a plain Esperanto text as shown in figure 2 which is subsequently sent to the Decoding terminal[1].

Al multnaciaj entreprenoj asignajtis subvencioj.

Figure 2: The Esperanto representation of the English sentence "Multinationals were allocated grants"

The Decoding terminal starts by parsing the Esperanto text into a dependency tree, and replacing Esperanto lexical items by those of the target language. However, because there are usually several words that are possible translations to a single Esperanto lexical item, the Metataxor generates several target dependency trees from a single Esperanto tree.

Disambiguation, in this half of the process, requires bilingual information; an Esperanto to target language bilingual dictionary that contains Esperanto word pairs as contextual clues for the plausibility of a word in a given context (see figure 3) [9]. In addition, there can be no interaction with the receiving user.

Esperanto entry word	Semantic relator (an Esperanto morpheme)	Disambiguating contexts: Esperanto words (with morpheme tokens), here illustrated with approximative English glosses	French word with syntactic word class
akr'a	*a*	*dolor'o, mal'varm'o, riproĉ'o'j, vort'o'j, romp'o, eĝ'o* 'pain', 'cold', 'blames', 'words', 'break', 'edge'	vif/ADJ
akr'a	*a*	*naz'o, orel'o'j, tur'o* 'nose', 'ears', 'tower'	pointu/ADJ
akr'a	*a*	*spic'o, pipr'o, brand'o* 'spice', 'pepper', 'brandy'	fort/ADJ

Figure 3: A sample from the Esperanto to French bilingual dictionary

If no exact match was to be found in the bilingual dictionary, proximity scores are again calculated using SWESIL. The target dependency tree is finally linearized and adjusted to form a readable text to be received by the target user. The overall design of the DLT system is shown in figure 4 [1].

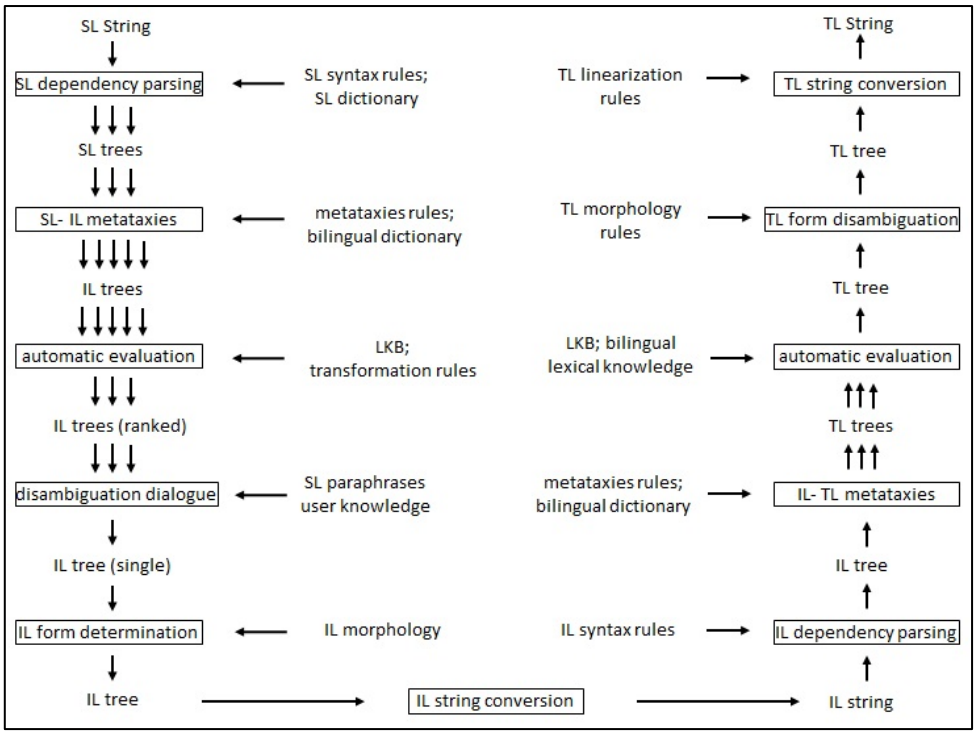

Figure 4: The overall design of the DLT system

B. UNITRAN

The name UNITRAN stands for UNIversal TRANslator; a translation system developed at Massachusetts Institute of Technology. The system operates bidirectionally between Spanish and English. However, other languages may be added by setting the parameters that fit them [11].

The UNITRAN system comprises two main components between which processing tasks are divided; the syntactic component and the lexical-semantic component. The syntactic component is based on the Government and Binding theory, it is responsible for handling the language-specific syntactic differences by accepting and producing grammatically correct sentences. The syntactic component is composed of a set of parameters associated with universal principles. These parameters are built-in in the analyzer and generator to be set according to the values of the language being processed. Thus, the analyzer and generator used are the same for all languages. The lexical-semantic component,

on the other hand, is based on the Lexical Conceptual Structure theory, it contains the information necessary to provide a conceptual form (the LCS) to underlie the source language sentence, and to match it to the appropriate target-language lexical items [11], [12].

The intermediate representation (the LCS) relies on a set of primitives that serve as the basic units of meaning such as event, state, property…etc. [12].

First, the operator sets the analyzer parameters to suit the values of the source language. For example, the "null subject" parameter has to be set to "yes" for Spanish and Italian…etc., but to "no" for English and German…etc. This is done through a menu operation.

The processing, then, begins by the syntactic component parsing a morphologically analyzed input into a tree showing the structural relations between constituents. Then, the lexical-semantic component maps each source word onto its corresponding LCS. The resulting LCS forms are subsequently merged into a single LCS (the composed LCS) which is the interlingua representation underlying the whole input sentence (see figure 5) [12].

```
Example:
  (CAUSE PERSON
    (GO-LOC PERSON
      (TO-LOC
        (IN-LOC PERSON ROOM)))
    FORCEFULLY)
```

Figure 5: The composed LCS underlying the sentence "John broke into the room"

The second stage is substitution. Each node in the composed LCS is mapped onto a target language word and the resulting LCS is mapped onto the syntactic realization of the target language sentence.

After setting the generator's parameters to meet the requirements of the target language, the generation process start by performing structural movement and generating the correct morphological forms of the target sentence's constituents. Figure 6 shows the overall design of the UNITRAN translation system [11], [12].

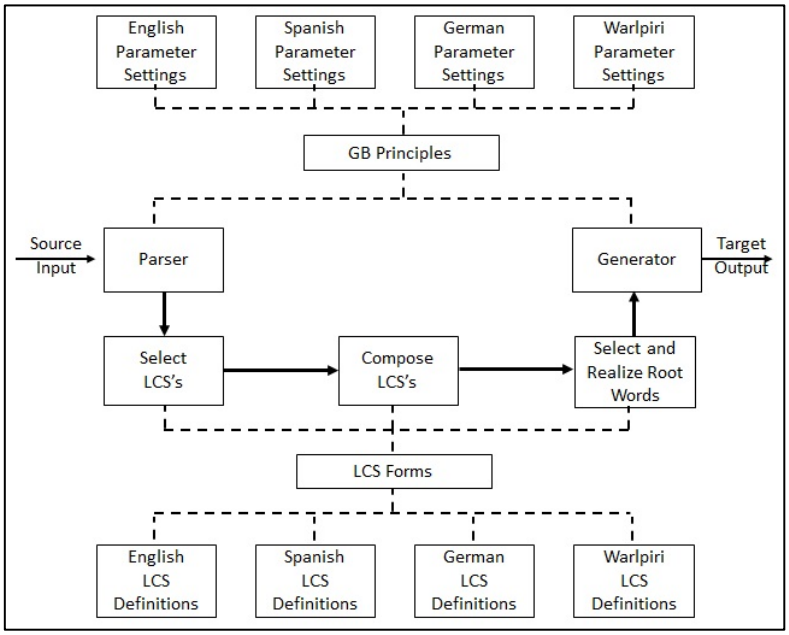

Figure 6:the overall design of UNITRAN

C. KANT

The KANT (Knowledge-based, Accurate Natural-Language Translation) system has been developed at Carnegie-Melon University (CMU) in Pennsylvania, USA in 1989 [13]. KANT is the only interlingua-based MT system to be operational commercially. It has been used in translating English technical documents into French, Spanish and German. The addition of more target languages such as Portuguese, Italian, Russian, Chinese and Turkish is under research [5]. The KANT prototype has also been used in generating Japanese and German [14].

KANT is a sublanguage translation system; it is used by large manufacturers to translate their technical documentation from English into several target languages [13]. “Though the analysis component must support generation in multiple languages, it currently handles only one source language, and therefore can tolerate a slight degree of source language dependence” [15].

The system codes for analysis and generation are language-independent whereas the specific knowledge required to process a certain language (grammars and lexicons) is developed separately for each language [13].

The first stage in the translation process is concerned with authoring the input. KANT is designed to translate only a well-defined subset of source language “constrained both by the domain from which the source texts are drawn (e.g. service information for heavy machinery), and by general restrictions” that are put on the vocabulary and structures of input language [15]. Kant’s vocabulary (non-domain specific) is limited to a basic vocabulary of about 14,000 distinct word senses while domain-specific technical terms are limited to a pre-defined vocabulary[16], approximately 60,000 words and phrases for heavy equipment manuals [17]. Structural restrictions, on the other hand, attempt to limit the use of constructions that would create difficulties in parsing such as the use of relative clauses with an explicit relative pronoun rather than reduced relative clauses [18].

In the first processing stage of knowledge-based parsing, the source text is processed using the source language grammar and lexicon to produce a Source F-Structure (a grammatical functional structure) for each sentence. Kant uses an explicit and very restricted domain model-based semantic restrictions to resolve ambiguity (e.g. phrase attachments). An example of these semantic restrictions is shown in figure 7[14].

```
(*E-CLEAN
   (is-a *EVENT)
   (agent *USER)
   (theme *PHYSICAL-LOCATION
          *PHYSICAL-OBJECT)
   (instrument *O-CLEANING-INSTRUMENT))
```

Figure 7: Kant’s semantic restrictions on the English verb "clean"

In the Interpretation stage, mapping rules map lexical items onto semantic concepts, and syntactic arguments onto semantic roles, forming the intermediate representation (see figure 8) [15]. The interlingua representation comprises information from all necessary levels of linguistic analysis; lexical, syntactic, semantic and pragmatic.

“The primary power supply component will supply the necessary 240 Volts DC to the input lead.” ⟹

```
(*a-supply
   (tense future)
   (mood declarative)
   (punctuation period)
   (source (*o-power-supply-component
             (reference definite)
             (number singular)
             (attribute (*p-primary))))
   (theme (*u-volt-dc
            (reference definite)
            (number plural)
            (attribute (*p-necessary))
            (quantity
               (*c-decimal-number
               (integer "240")
               (number-type cardinal)
               (number-form numeric)))))
   (goal_to (*o-input-lead
               (reference definite)
               (number singular))))
```

Figure 8: The interlingua representation for a sentence from a television repair manual

In the generation stage, target mapping rules indicate how the interlingua representation maps onto the appropriate Target F-Structure. The overall architecture of the Kant translation system is shown in figure 9[17].

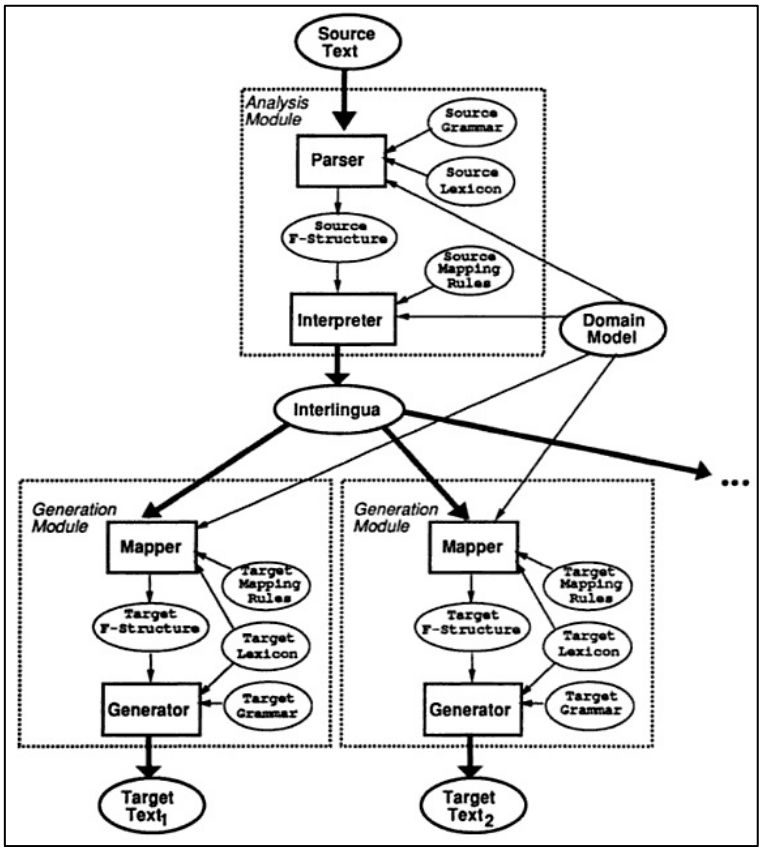

Figure 9: The run-time architecture of KANT

D. UNL

The Universal Networking Language project was launched in 1996 at the Institute of Advanced Studies of the United Nations University (UNU/IAS), Tokyo, Japan. 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 UNL; the Universal Networking Digital Language (UNDL) Foundation. In addition, 17 language centers all over the world are working on the development of the UNL resources necessary for incorporating their native language into the UNL program. Among these centers are the Arabic UNL center in Alexandria, Egypt (http://www.bibalex.org/unl), the Spanish center in Madrid, Spain (www.vai.dia.fi.upm.es) and the Russian center in Saint Petersburg, Russia (www.unl.ru).

The mission of the UNL program is to overcome the language barrier and enable all peoples to generate, and have access to, information and knowledge in their native languages and cultures by coding, storing and disseminating human knowledge, in any given domain, in a language-independent format that represents only the core content and abstracts away from the particular characteristics of the original language in which it was expressed[1].

The UNL is a formal artificial language that replicates the functions of natural language in communication, but is, nevertheless, designed for computers rather than humans; in other words, it is not intended to be an auxiliary language such as Esperanto, Ido or others. People should use UNL in “communication” in the same subtle manner they do with other procedural languages such as HTML[19].

The UNL program has passed through several stages of development[20], [21], the third and latest of which is the UNL+3 project. A three-year project to advance the long-term mission of the UNDL[2]. In this phase, the linguistic infrastructure has been developed using the X-bar theory. Accordingly, the analysis and generation processes pass over five stages, rather than the direct approach adopted in the previous approaches, to help yield more accurate results.

[1]More information about the UNDL, its ideology and its goals is available at http://www.undl.org

[2]More information about UNL+3 is available at www.unlweb.net.

The main bulk of the UNL system is language-independent. The engines' codes necessary for converting natural language input into UNL (UNLization) and converting UNL into natural target language (NLization) are the same whatever the input or output language may be.

In addition, information on the semantic abstract concepts (Universal Words or UWs) depicted by different cultures are organized hierarchically in the UNL Knowledge Base (UNL KB) and the UNL Ontology. The UNL KB constitutes a network structure where UWS are interconnected through the Relations of UNL. It comprises any relation necessary to define a given UW. Figure 10 shows the extended format of the UNL KB entries.

```
<relation name="RNAME" type="RTYPE" frequency="RFREQ">
  <source id="SID" attribute="ATT" lang="UNL" frequency="SFREQ" class="SCLASS">SOURCE</source>
  <target id="TID" attribute="ATT" lang="UNL" frequency="TFREQ" class="TCLASS">TARGET</target>
</relation>
```

Figure 10: An entry in the extended format of the UNL KB

where:

RNAME is the name of one existing UNL relation ("agt", "aoj", "obj", etc);
RTYPE is the type of the existing relation
RFREQ is the frequency of the relation TYPE between the SOURCE and the TARGET in the corpus;
SFREQ is the frequency of the SOURCE in the corpus;
TFREQ is the frequency of the TARGET in the corpus;
SID is a number used to identify the SOURCE;
TID is a number used to identify the TARGET;
ATT is one of the existing UNL attributes ("entry", "past", etc);
SCLASS is the general class of the SOURCE;
TCLASS is the general class of the TARGET;
SOURCE is the source node of the UNL relation;
TARGET is the target node of the UNL relation;

On the other hand, The UNL Ontology, formerly known as the UW System, is a tree-like structure where UWs are interconnected through hierarchical relations only: icl (is-a-kind-of) and iof (is-an-instance-of). It only contains relations that may be used for inheritance.

There is also the UNL<->NL Memory which is a set of mappings between a given natural language and UNL. It may be unidirectional (UNL-NL Memory or NL-UNL Memory) or bidirectional (UNL<->NL Memory). It is used to improve and normalize the results of the UNLization and the NLization, as it contain segments that have been previously UNLized or NLized.

Language-dependent resources, on the other hand, are developed by the language center of the respective language. They include lexicons that map natural language lexical items onto universal concepts (Universal Words or UWs) and vice versa, and the grammar rules that determine well-formedness standards. Two types of lexicons are used in the UNL system; enumerative and generative. The enumerative dictionary is used for UNLization, it contains all the word forms of a language and the corresponding UWs, The generative dictionary is used for NLization, it stores only a base form of each lexical item along with a set of generative rules that can manipulate the base form into the form required in the target language.

Translation takes place over two completely independent processes; UNLization and NLization. The language-independent UNLization tool (IAN) converts natural language input into UNL format through five phases. First, the natural language list is processed to identify the abstract concepts represented by the words in the input sentence using the language's word dictionary. Second, these constituents are parsed into a surface syntactic tree. Third, the surface tree is analyzed on a deeper level to form the deep syntactic tree. Fourth, the syntactic tree is transformed into a semantic network and finally the network is post-edited for any modifications that would make the resulting semantic network more accurate. Figure 11 shows the design of the UNLization process.

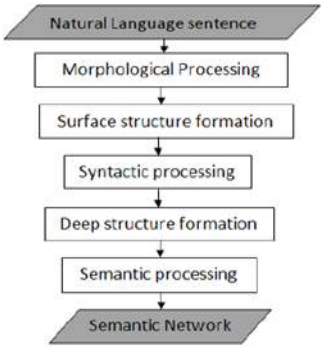

Figure 11: The design of the UNLization process

The result is a semantic network that is the input for the NLization process. Semantic networks represent all aspects of input content; semantic, syntactic, pragmatic, format…etc. Semantic information about the abstract concepts themselves is stored with each UW. As for the semantic links that tie these concepts in a given sentence, they are expressed via Relations. Relations are three letter symbols expressing an ontological relation such as "icl" (a kind of), a thematic relation such as "agt" (agent), or a logical relation such as "and" [22]. Note that these Relations are entirely semantic and are not influenced by the syntactic roles of the constituents in the sentence being processed.

On the other hand, grammatical information such as Tense, Aspect, Person, and Number…etc. is encoded in the form of Attributes of linguistic features. Attributes are tags that annotate a particular word in the semantic network such as "@past", "@progressive"…etc. Attributes also express contextual and subjective information such as "@discontented", and "@insistence". Some information about the formatting of the original co-text is also encoded in Attributes such as "@parenthesis" and "@title"[3][22].

Linguistic features on the other hand are extracted from the UNL tagset. The UNL tagset is a standardized repository containing tags for some specific and pervasive grammatical phenomena. Many of those linguistic constants have been proposed to the Data Category Registry (ISO 12620), and represent widely accepted linguistic concepts. This tagset helps standardize and harmonize the UNL resources (lexicons and grammars) so as to make them as understandable and exchangeable as possible. Tags in the tagset are used to mark each entry in a language lexicon with all the linguistic information it carries; such as number, gender, semantic typology, register, etc. The final UNL semantic network will look as shown in figure 12 while figure 13 shows the UNL expression of the same sentence. The UNL expression is the formal representation of semantic networks.

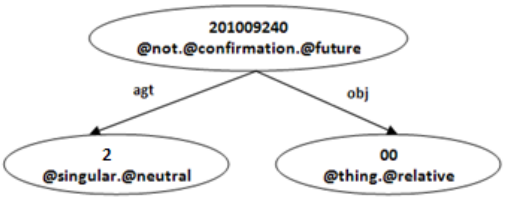

Figure 12: The UNL graph representing the sentence "you won't say that will you?"

```
agt(201009240:XC.@future.@not.@confirmation.@entry,  00:DF.@2.@singular)
obj(201009240:XC.@future.@not.@confirmation.@entry,   00:DM.@thing.@relative)
```

Figure 13: The UNL expression for the UNL network in figure 12.

The NLization process begins after receiving the UNL expressions of the text to be translated. The language-independent NLization tool (EUGENE) uses the target language word dictionary to transform it into a natural language sentence. The NLization process passes through five phases similar to the UNLization process but in the reverse order. First, the UNL network is edited in order to make it more suitable for translation. Second, the network is transformed into a deep syntactic structure from which the surface structure is extracted in the third phase. In the fourth phase, the tree structure is linearized into a list structure and finally this list is post-edited for morphological adjustments to produce a well-formed comprehensible natural language sentence. Figure 14 shows the design of the NLization process.

[3]The complete set of UNL specifications and components is also available at http://www.undl.org/

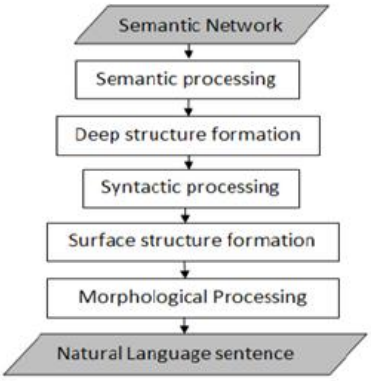

Figure 14: The design of the NLization process

The overall architecture of the UNL translation process is shown in figure 15.

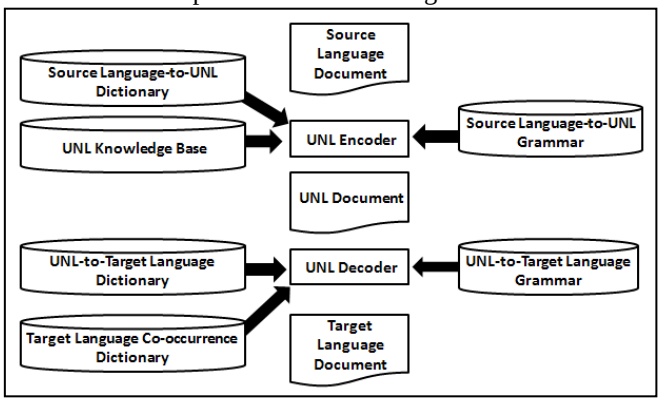

Figure 15: The overall architecture of the UNL system

The processes of UNLization and NLization may follow several different paradigms, as follows:

- Language-based UNLization (based mainly in a NL-UNL dictionary and NL-UNL grammar)
- Knowledge-based UNLization (based mainly in the UNL Knowledge Base)
- Memory-based UNLization (based mainly in the UNL-NL Memory)
- Statistical-based UNLization (based mainly in statistical predictions derived from UNL-NL corpora)
- Dialogue-based UNLization (based mainly in the interaction with the user)

The actual UNLization and NLization are normally hybrid and may combine several of the strategies above.

In the current phase of UNL development; UNL+3, specifications have been modified in order to cover even more linguistic phenomena, and to handle some of the problems in the earlier stages. The new project also offers a free and open virtual learning environment (VALERIE) for those wishing to contribute to the development of such a massive project[4] in addition to the UNLarium which is an open-source web-based development environment where registered users are able to create, edit, share, search, export and download lexical and grammatical resources that have been provided by other users and in other languages[5][23], [24].

Although Machine Translation is one of the possible and more obvious and promising uses of UNL, it is not the only area in NLP where UNL can prove useful. As it offers a complete understanding of natural language content, UNL can serve areas such as summarization and text simplification. Moreover, by providing a language-neutral representation of meaning, UNL can dramatically improve our ability to search for and find information, thus, helping areas such as information retrieval and others.

The Universal Networking Language has already proved its efficiency in several projects such as encoding the contents of 25 English documents from the Encyclopedia of Life Support Systems (EOLSS) in UNL, and successfully generating their Arabic, French, Japanese, Russian and Spanish counterparts [25],[26]–[27]. The output of this project

[4] Available at www.unlweb.net/valerie/

[5] Available at (www.unlweb.net/unlarium/)

has been evaluated qualitatively and statistically and the results were significantly higher than those of Google, Babylon and Sakhr'sTarjim[28]. In addition, UNL has been used in creating a prototype language-independent Library Information System (LIS) that provides the resources necessary for the generation of books' metadata into at least six languages other than the original Arabic[29], [30].

4 DISCUSSION

All of the previous projects exhibit intriguing approaches to defining an abstract language-independent format for knowledge representation. However, they vary in the degree of language-independency, the complexity through which they achieve such a representation and in their capabilities. Most of the previous systems incorporate a stage of syntactic parsing which leads to a semantic mapping of the resulting syntactic tree. UNL also uses such stages but instead of twostages, UNL uses five gradual stages to analyze the natural language sentence morphologically, on the surface syntactic level, on the deep syntactic level and semantically, and vice versa in generation. This leads to far greater accuracy in the understanding of natural language.

"The interlingua approach necessarily requires complete resolution of all ambiguities in the SL text so that translation into any other language is possible" [14]. Hence, a large section of processing stages in most interlingua-based MT systems is devoted to disambiguating the input to form the interlingua, and in some cases, disambiguating the intermediate representation to derive the output (as in the DLT system). In other cases, the system enforces very strict limitations on input language and imposes precise semantic restrictions on the abstract concepts to avoid prolonged processing (as the case in Kant). As a last resort, some systems turn to interactive communication with the user(s) such as DLT. UNL has largely avoided such intricate procedures by using a quite unambiguous intermediate representation. In UNL, a word can never have more than one conceptual representation; they are clearly distinguished by the ID number that represents their exact contextual meaning. For example, "bank" meaning "a financial institution that accepts deposits and channels the money into lending activities"is clearly differentiated from "bank" meaning "sloping land (especially the slope beside a body of water)" by means of the Universal Words "108420278" and "109213565", respectively. However, the UNL system does employ disambiguation techniques on the word, syntactic tree and semantic network levels, but these techniques are entirely optional, when not used, the system can still output acceptable results.

Moreover, most interlingua-based MT systems miss one aspect of meaning or another such as UNITRAN that does not incorporate the notion of grammatical aspect [12]. UNL, on the other hand, tries to convey all aspects of meaning in its intermediate representation; semantic, syntactic, pragmatic, subjectivity, format …etc. Yet, UNL developers acknowledge that the "subtleties of intention and interpretation make the "full meaning" [. . .] too variable and subjective for any systematic treatment". Hence, it avoids the mistake of "trying to represent the "full meaning" of sentences or texts, targeting instead the "core" or "consensual" meaning that is most often attributed to them". It is also not committed to replicate the lexical and the syntactic choices of the original input and can be, therefore, regarded as an interpretation rather than a translation[6].

Mapping natural language onto an unambiguous conceptual representation is indeed quite challenging, which is why several projects attempted to curb the difficulty by either limiting input texts to specific domains (such as Kant) or controlling input language vocabulary and structures (such as Kant and DLT's prototype). UNL, however, does not put any kind of restriction on input texts or language; nevertheless, "much of the subtlety of poetry, metaphor, figurative language, innuendo and other complex, indirect communicative behaviors is beyond the current scope and goals of the UNL". Instead, it focuses on "direct communicative behavior" which accounts for "much or most of human communication in practical, day-to-day settings"[7].

Although an interlingua should, in theory, be universal, no interlingua-based system has ever intermediated between more than 10 languages. Still, UNL's mission is to eradicate language barriers by intermediating between all natural languages and has already started by incorporating 17 languages. UNL, as a non-profit project, would make possible the instant generation of various target-language versions of such a vital source of knowledge such as the internet, upon request, if webpages were to contain a UNL representation of its content along with the original language.

UNL is not simply an intermediate representation; it is a full-scale language for machines. This means that its uses go beyond the task of translation as mentioned earlier. Besides, it can represent any imaginable concept because, unlike a system such as UNITRAN which builds concepts from a limited set of primitives [12], it makes use of dozens of semantic Relations and Attributes to exactly convey the intended meaning. Another system such as the DLT uses a regularized "human" language "with its own lexical items and syntactic rules" which "caused translation in the DLT system to be sometimes viewed as, in fact, two translation processes rather than one" [1].

---

[6] This excerpt is taken from http://www.unlweb.net/wiki/index.php/Introduction_to_UNL

[7] From http://www.unlweb.net/wiki/index.php/Introduction_to_UNL

Unfortunately, due to its challenging nature, most interlingua-based system never makes it beyond the research phase. UNL, on the other hand, is no more a pilot project; it was launched in 1996 and has been ever since subject to constant developments and enhancements under the auspices of the UNDL foundation and the United Nations. The most recent development (the UNL+3) recruits even more participants by offering a free learning environment and promotes integration by providing an open-source environment for developers to share their resources. Besides, UNL has also been successfully used in numerous projects and its output is constantly subject to evaluation.

5 CONCLUSION

A language-neutral representation of meaning has always been the dream of MT researchers. Although it is one of the oldest approaches in the field, very few systems have ever attained international recognition. This paper describes three of the most referred to systems as pioneers in devising an interlingua-based system, briefly examining their designs and characteristic features and how a more modern fourth system; UNL, has succeeded in mending some of their imperfections that impeded reaching the ultimate goal of bringing down the language barriers.

REFERENCES

[1]W. J. Hutchins, and H. L. Somers, An Introduction to Machine Translation, (chapter1, 4, 17) (chapter 1 p.8) London Academic Press Limited, 1992.
[2]W. J. Hutchins, Machine translation: past, present, future. Ellis Horwood Series in Computers and their Applications, Ellis Horwood, Chichester, UK, 1986.
[3]S. Nirenburg and Y. Wilks, “Machine Translation”, Advances in Computer, vol. 52, pp. 160-189, 2000.
[4]A. Lampert, “Interlingua in Machine Translation”, Technical Report, 2004.
[5] D. Bonnie J., E. Hovy and L. Levin, "Machine Translation: Interlingual Methods", Encyclopedia of Language and Linguistics. 2nd ed., Brown, Keith (ed.), 2004.
[6]W. J. Hutchins, “Machine translation: a brief history”, Concise history of the language sciences: from the Sumerians to the cognitivists. E. F. K. Koerner and R. E. Asher (eds.), Pergamon, pp.431-445, 1995.
[7]H. Uchida and M. Zhu, “Interlingua for Multilingual Machine Translation”, in Proceedings of the Machine Translation Summit IV, Kobe, Japan, July 20-22, 1993.
[8]T. Witkam, “History and Heritage of the DLT (Distributed Language Translation) project”. [Utrecht, The Netherlands: private publication, 2006].
[9]D. Maxwell, K. Schubert and T. Witkam (eds.), New Directions in Machine Translation, (chapter 8), Foris Publications, Dordrecht, Holland, 1988.
[10]T. Witkam, "DLT — An Industrial R&D Project for Multilingual MT", in Proceedings of Proceedings of the 12th International Conference on Computational Linguistics (COLING 1988), Budapest, 1988.
[11]D. Bonnie J., “UNITRAN: An Interlingua Approach to Machine Translation”. in Proceedings of the 6th Conference of the American Association of Artificial Intelligence, Seattle, Washington, 1987.
[12]D. Bonnie J., “A cross-linguistic approach to translation”. in Proceedings of 3rd International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, Linguistics Research Center, University of Texas, Texas, 11-13 June, 1990.
[13]E. H. Nyberg, T. Mitamura and J. Carbonell, “The KANT Machine Translation System: From R&D to Initial Deployment1”, in Proceedings of LISA (The Library and Information Services in Astronomy) Workshop on Integrating Advanced Translation Technology, Hyatt Regency Crystal City, Washington D.C., June 3-4, 1997.
[14]T. Mitamura, E. H. Nyberg III and J. G. Carbonell, “An Efficient Interlingua Translation System for Multi-lingual Document Production”, in Proceedings of Machine Translation Summit III, Washington D.C., The United States, July 2-4, 1991.
[15]Deryle W. Lonsdale, A. M. Franz and J. R. R. Leavitt. “Large Scale Machine Translation: An Interlingua Approach”. in Proceedings of the 7th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Austin, Texas, The United States. 1994.
[16]E. H. Nyberg and T. Mitamura, “The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains,” in Proceedings of the International Conference on Computation Linguistics, (COLING 1992), Nantes, France, July, 1992.
[17]T. Mitamura, E. H. Nyberg 3rd and J. G. Carbonell, “Automated Corpus Analysis and the Acquisition of Large, Multi-Lingual Knowledge Bases for MT1”, in Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, Japan, July 14-16, 1993.
[18]T. Mitamura, “Controlled Language for Multilingual Machine Translation1”, in Proceedings of Machine Translation Summit VII, Singapore, September 13-17, 1999.

[19]H. Uchida, UNL: Universal Networking Language – An Electronic Language for Communication, Understanding, and Collaboration, UNU/IAS/UNL Center, Tokyo, Japan. 1996.
[20]S. Alansary, M. Nagi and N. Adly, "Generating Arabic text: The Decoding Component of an Interlingual System for Man-Machine Communication in Natural Language", in Proceedings of the 6th Egyptian Society of Language Engineering Conference, (ESOLEC 2006), 6-7 December, Cairo, Egypt. 2006.
[21]S. Alansary, M. Nagi, and N. Adly, "Processing Arabic Content: The Encoding Component of an Interlingual System for Man- Machine Communication in Natural Language", in Proceedings of the 6th Egyptian Society of Language Engineering Conference, (ESOLEC 2006), Cairo, Egypt, 2006a.
[22]H. Uchida and M. Zhu, "UNL2005 for Providing Knowledge Infrastructure", in Proceedings of the Semantic Computing Workshop (SeC2005), Chiba, Japan, 2005.
[23]S. Alansary, M. Nagi and N. Adly, "UNL+3: The Gateway to a Fully Operational UNL System", in Proceedings of the 10th Egyptian Society of Language Engineering Conference (ESOLEC 2010), Ain Shams University, Cairo, Egypt, December 15 – 16, 2010.
[24]S. Alansary, "A Practical Application of the UNL+3 Program on the Arabic Language", in Proceedings of the 10th Egyptian Society of Language Engineering Conference, Ain Shams University, Cairo, Egypt, December 15 – 16, 2010.
[25]S. Alansary, M. Nagi and N. Adly, "A Semantic-Based Approach for Multilingual Translation of Massive Documents", in Proceedings of The 7th International Symposium on Natural Language Processing, (SNLP), Pattaya, Thailand, 2007.
[26]S. Alansary, M. Nagi and N. Adly, "Machine Translation Using the Universal Networking Language (UNL)" , in Proceedings of the 8th International Conference on Language Engineering, Ain Shams University, Egypt, December 18 - 19 2008.
[27]S. AlAnsary, M. Nagi and N. Adly, "The Universal Networking Language in Action in English-Arabic Machine Translation", in Proceedings of 9th Egyptian Society of Language Engineering Conference on Language Engineering, (ESOLEC 2009), Cairo, Egypt December 23-24, 2009.
[28]N. Adly, and S. Alansary, "Evaluation of Arabic Machine Translation System based on the Universal Networking Language", in Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems, (NLDB 2009), Saarland University, Saarbrücken-Germany, June 24 - 26 2009.
[29]S. Alansary, M. Nagi and N. Adly, "A Library Information System (LIS) Based on UNL Knowledge Infrastructure". in Proceedings of the Universal Networking Language Workshop In conjunction with 7th International Conference on "computer science and information technology - 2009" (CSIT-2009), Yerevan – Armenia, September 28th - October 2nd , 2009.
[30]S. Alansary, M. Nagi and N. Adly, "Towards a Language-Independent Universal Digital Library", in Proceedings of the Second International Conference on Universal Digital Libraries, (ICUDL 2006), Alexandria, Egypt, 17-19 November, 2006.

BIOGRAPHY

Dr. SamehAlansary is associate professor of computational linguistics in the Department of Phonetics and Linguistics and acting as head of Phonetics and Linguistics Department, Faculty of Arts, Alexandria University. He obtained his MA in Building Arabic Lexical Databases in 1996, and his PhD from Nijmegen University, the Netherlands in building a formal grammar for parsing Arabic structures in 2002. His main areas of interest are concerned with corpus work, morphological analysis and generation, and building formal grammars. He is also the head of Arabic Computational Linguistics Center in Bibliotheca Alexandrina. He is supervising and managing the Universal Networking Language project in Library of Alexandria since 1-6-2005 till now. Dr. Alansary is the co-founder of the Arabic Language Technology Center (ALTEC), an NGO aims at providing Arabic Language resources and building a road map for Arabic Language Technology in Egypt and in the Middle East. He Has many scientific works in Arabic Natural Language Processing published in international conferences and periodicals, and a member in many scientific organizations: (1) Egyptian Society of Language Engineering, Cairo, (2) Arabic Linguistic Society - USA, (3) Association of Computational Linguistics - USA – Europe, (4) Universal Networking Language foundation, United Nations, Geneva, Switzerland. He has obtained Alexandria University Award for promoting scientific research (2008) and for recent and advanced practices (2012).

أنظمة الترجمة الآلية المعتمدة على لغة وسيطة: لغة الشبكات الدلالية في مقابل اللغات الوسيطة الأخرى

د. سامح الأنصاري
قسم الصوتيات واللغويات، كلية الآداب،جامعة الإسكندرية، الإسكندرية، مصر

خلاصة:

تعرض هذه الورقة بعض المحاولات الأكثر شهرة في مجال الترجمة الآلية المعتمدة على لغة وسيطة من ضمنها نظام لغة التواصل العالمية كلغة وسيطة للترجمة الآلية كأحد تطبيقاته مع مقارنة هذا النظام بالأنظمة الأخرى وبيان الفرق بينهم وبيان مدى كفاءة نظام لغة التواصل العالمية في مجال الترجمة الآلية لما له من مكونات وموارد تمكنه من ذلك. وتتناول المقدمة لمحة عن تاريخ وتطور الترجمة الآلية وتوضح الاتجاهات الثلاثة المختلفة للترجمة الآلية مع التركيز على الاتجاه المعتمد على لغة وسيطة باعتباره أفضل اتجاهات الترجمة والذي تحقق أعلى النتائج. فتتناول الورقة هذا الاتجاه للترجمة الآلية بشكل مفصل وتعرض التعريف به وخصائصه ومميزاته عن باقي الاتجاهات الأخرى. ثم تتناول بعض أنظمة الترجمة الآلية المشهورة المبنية على الاتجاه المعتمد على لغة وسيطة. ومن هذه الأنظمة نظام DLTوهو مشروع تم تطويره في هولندا وكانت مدته سته أعوام من عام ١٩٨٥. وهو نظام تفاعلي متعدد اللغات يعمل عبر الشبكات الحاسوبية، ويعتمد هذا النظام أساسا على استخدام لغة الاسبرانتو كلغة وسيطة. وتوضح الورقة شرح تفصيلي لطريقة عمل نظام DLT مع مخطط توضيحي لمكونات النظام. ونظام آخر تناولته الورقة وهو نظام UNITRAN وهو اختصار لكلمة المترجم العالمي. هذا النظام يترجم عبر مجموعة متنوعة من اللغات وليس لغتان فقط أو عائلة من اللغات وهي الإنجليزية والألمانية والإسبانية. ويعتمد على نقل اللغة المصدر إلى لغة وسيطة ثم ترجمتها إلى أي لغة هدف. وتوضح الورقة طريقة عمل نظام UNITRAN ومكونات النظام من خلال شكل توضيحي له. وتناولت الورقة نظام ثالث من أنظمة الترجمة وهو KANT وهو جزء من مركز الترجمة الآلية بجامعة كارنيجى ميلون ويستخدم لترجمة دليل المستخدم للمعدات الثقيلة فهو يعمل على ترجمة نصوص في مجال معين، ويترجم بين اللغة الفرنسية والإسبانية. ويعتبر نظام KANT هو النظام الوحيد بين الأنظمة السابقة الذكر الذي تم استخدامه في إطار تجاري. وتوضح الورقة مكونات النظام بالتفصيل وطريقة عمله. وأخيرا نظام لغة التواصل العالمية والذي يعد من أحدث أنظمة الترجمة الآلية فهو لا يعد نظام للترجمة فحسب بل هو نظام للتوسط بين جميع لغات العالم ومن أحد تطبيقاته الترجمة الآلية بالإضافة إلى العديد من التطبيقات الأخرى. وقد بدأت لغة التواصل العالمية في معهد الدراسات المتقدمة في طوكيو عام ١٩٩٦ ثم تم تأسيس منظمة لغة التواصل العالمية في جنيف عام ٢٠٠١ لتكون مسؤولة عن تطوير هذا النظام. ويعتبر نطام لغة التواصل العالمية تقنية جديدة في مجال الترجمة الآلية المعتمدة على لغة وسيطة فلغة التواصل العالمية هي لغة للحاسوب تمكنه من فهم اللغات الطبيعية والتعامل معها لأن مكوناتها تشبه مكونات اللغات الطبيعية من مفردات ونحو وسمات تعبرعن قصد المتكلم وغيرها من المعلومات غير اللغوية. ويعتمد نظام لغة التواصل العالمية على التمثيل الدلالي للمعارف ومن ثم نقلها من لغة إلى لغة أخرى.

يركز البحث على بيان قوة نظام لغة الشبكات الدلالية في الترجمة الآلية متعددة اللغات ويبين الفرق بينه وبن الأنظمة الأخرى للترجمة الآلية موضحا نقاط القوة والضعف في كل منهم.