2009-1

UNIARAB: An Universal Machine Translator System For Arabic Based On Role And Reference Grammar

Brian Nolan
Institute of Technology, Blanchardstown, brian.nolan@tudublin.ie

Yasser Salem
Institute of Technology, Blanchardstown

UniArab: An RRG Arabic-to-English machine translation software

Brian Nolan
Yasser Salem
Institute of Technology Blanchardstown, Dublin IRELAND
brian.nolan@gmail.com
yasser.salem@gmail.com

Abstract

This paper presents a machine translation system (Hutchins 2003) called UniArab (Salem, Hensman and Nolan 2008). It is a proof-of-concept system supporting the fundamental aspects of Arabic, such as the parts of speech, agreement and tenses. UniArab is based on the linking algorithm of RRG (syntax to semantics and vice versa). UniArab takes MSA Arabic as input in the native orthography, parses the sentence(s) into a logical meta-representation based on the fully expanded RRG logical structures and, using this, generates perfectly grammatical English output with full agreement and morphological resolution. UniArab utilizes an XML-based implementation of elements of the Role and Reference Grammar theory in software. In order to analyse Arabic by computer we first extract the lexical properties of the Arabic words (Al-Sughaiyer and Al-Kharashi 2004). From the parse, it then creates a computer-based representation for the logical structure of the Arabic sentence(s). We use the RRG theory to motivate the computational implementation of the architecture of the lexicon in software. We also implement in software the RRG bidirectional linking system to build the parse and generate functions between the syntax-semantic interfaces. Through seven input phases, including the morphological and syntactic unpacking, UniArab extracts the logical structure of an Arabic sentence. Using the XML-based metadata representing the RRG logical structure, UniArab then accurately generates an equivalent grammatical sentence in the target language through four output phases. We discuss the technologies used to support its development and also the user interface that allows for the addition of lexical items directly to the lexicon in real time. The UniArab system has been tested and evaluated generating equivalent grammatical sentences, in English, via the logical structure of Arabic sentences, based on MSA Arabic input with very significant and accurate results (Izwaini 2006). At present we are working to greatly extend the coverage by the addition of more verbs to the lexicon. We have demonstrated in this research that RRG is a viable linguistic model for building accurate rule-based semantically oriented machine translation software. Role and Reference Grammar (RRG) is a functional theory of grammar that posits a direct mapping between the semantic representation of a sentence and its syntactic representation. The theory allows a sentence in a specific language to be described in terms of its logical structure and grammatical procedures. RRG creates a linking relationship between syntax and semantics, and can account for how semantic representations are mapped into syntactic representations. We claim that RRG is very suitable for machine translation of Arabic, notwithstanding well-documented difficulties found within Arabic MT (Izwaini, S. 2006), and that RRG can be implemented in software as the rule-based kernel of an Interlingua bridge MT engine. The version of Arabic (Ryding 2005, Alosh 2005, Schulz 2005), we consider in this paper is Modern Standard Arabic (MSA), which is distinct from classical Arabic. In the Arabic linguistic tradition there is not a clear-cut, well defined analysis of the inventory of parts of speech in Arabic.

Keywords: Arabic Machine Translation, Role and Reference Grammar, RRG, Java programming, XML

1 Introduction

This paper reports on recent work the development of a rule-based semantically oriented Interlingua bridge framework for machine translation of Arabic language processing using the Role and Reference Grammar (RRG) linguistic model. Machine translation is a sub-field of computational linguistics that investigates the use of computer software to translate text (or speech) from one natural language to another. Our system has been developed and is able to analyse Arabic sentences in native orthography, and extract their logical structure. Through a detailed study of the Arabic language, we have been able to develop an analyser that can successfully process many of the unique features and challenges present in Arabic. This logical structure is then used in the generation phase, where the sentence(s) is translated into another language, in this case, English.

The Arabic language is written from right to left, it has complex, language-specific grammar rules, and a relatively free word order. These distinguishing features pose a major challenge in processing Arabic text for linguistic analysis. Our framework demonstrates that RRG is a feasible foundation for building multi-language machine translations systems. Arabic is a Semitic language originating in the area presently known as the Arabian Peninsula. The Arabic language is one of six major world languages, and one of the six official languages of the United Nations. The version of Arabic we consider in this work is Modern Standard Arabic (MSA). When we mention Arabic throughout this paper we mean MSA, which is a distinct, modernized form of Classical Arabic (Alosh 2005). MSA is the universal written language of the Arabic-speaking population, printed in most books, newspapers, magazines, official documents, and reading primers for children. Most of the oral Arabic spoken today is more divergent than the written Arabic language, because of dialectal interference. However MSA is the literary and standard variety of Arabic used in writing and formal speeches today (Schulz 2005).

In this paper we discuss the RRG UniArab MT research project and the Interlingua model of Arabic MT that we designed and built using Java and XML. With this we discuss the challenges inherent within Arabic MT and the part that RRG played in helping to overcome many of the challenges. The architecture of the lexicon and its design and implementation in XML is discussed, along with a presentation of the results produced by the UniArab software evaluation

2 The Role and Reference Grammar (RRG) Linguistic Model

Role and Reference Grammar (RRG) is a model of grammar that posits a direct mapping between the semantic representation of a sentence and its syntactic representation (Van Valin 2005). We claim that RRG is very suitable for machine translation of Arabic via an Interlingua bridge implementation model. RRG is a mono strata-theory, positing only one level of syntactic representation, the actual form of the sentence and its linking algorithm can work in both directions from syntactic representation to semantic representation, or vice versa. In RRG, semantic decomposition of predicates and their semantic argument structures are represented as logical structures. The lexicon in RRG takes the position that lexical entries for verbs should contain unique information only, with as much information as possible derived from general lexical rules.

The main features of RRG are the use of lexical decomposition, based upon predicate semantics, an analysis of clause structure and the use of a set of thematic roles organized into a hierarchy in which the highest-ranking roles are 'Actor' (for the most active participant) and 'Undergoer' (Van Valin 2005). RRG characterises the relationship between syntax and semantics and can account for how semantic representations are mapped into syntactic representations. RRG also accounts for the very different process of mapping syntactic representations to semantic representations. Of the two directions, syntactic representation to semantic representation is the more difficult since it involves interpreting the morphosyntactic form of a sentence and inferring the semantic functions of the sentence from it. Accordingly, we have chosen to implement Arabic to English as the translation direction and the basis of the parse and generate functions in this version of our software.
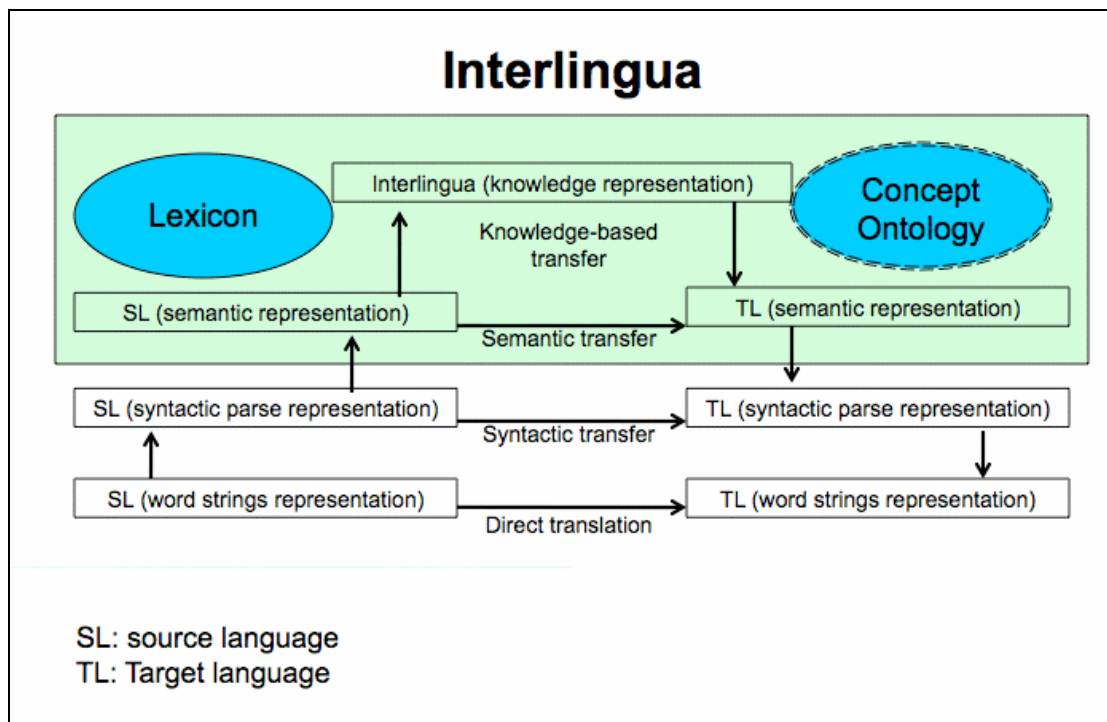
3 Interlingua approach of Arabic MT

The Interlingua approach is to develop a universal language-representation for text. In effect, in an Interlingua there is no transfer map, and the MT model thus has two main stages: input-PARSE-analysis and output-GENERATE.
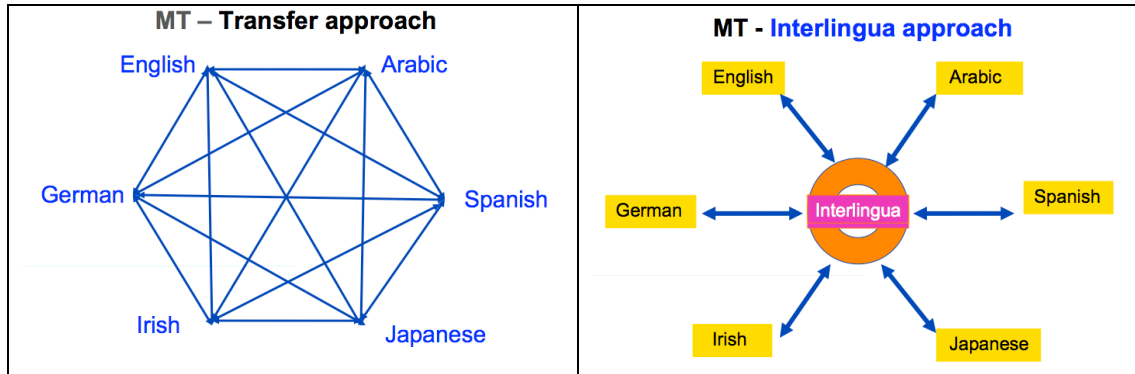
Figure 1: MT – Transfer vs. Interlingua approach

Interlingua-based MT is done via an intermediate semantic representation, based on RRG logical structures, of the source language text. An Interlingua is designed to be a language independent representation from which translations can be generated to different target languages.

Figure 2: MT – Our Interlingua approach

3.1 UniArab: Lexical representation in an Interlingua system

Transfer oriented translation systems (Figure 1) do not scale up when additional languages are added beyond the initial source (SL1) and target (TL1) language pairs, and very quickly this leads to a translation complexity problem between languages. Additionally, of course, in simple transfer-based systems there are no problems if, for

a particular language pair, there are morphosyntactic one-to-one equivalents; problems do arise, however, when there is more than one target word for a single source word.

Implementation of an Interlingua bridge architecture solves (Figure 2) the translation complexity problem as automatic language translation is made from a source language into a kernel meta representation (the input PARSE phase) and generates to a target language from the meta representation (the GENERATION phase). Ambiguity problems for an Interlingua in a multilingual system are still likely if one of the languages involved has two or more potential forms for a single given word in one of the other languages. A semantically oriented approach to MT can potentially disambiguate more easily than other strategies. For an Interlingua to be completely language-neutral, it must represent not the words of one or another of the languages, but language-independent lexical units. Any distinction that can be expressed lexically in the languages of the system must be represented explicitly in the Interlingua representation (Hutchins 2003). We use the RRG logical structures as the basis of our meta-representation in the Interlingua Bridge with a lexicon encoded in XML.

The UniArab system can generate a target language through classifying every Arabic word in the input source text by creating a meta-representation of the sentence(s) input as a text in a fully populated RRG-style logical structure including its various nominals and their associated features of [def+, masc+], etc.. There are six major parts of speech in Arabic. These are verbs, nouns, adjectives, proper nouns, demonstratives, adverbs and we create a seventh for purposes of our software, which we have simply called the `other' category for Arabic words that do not fit into any of previous six categories. The major parts of speech in the Arabic language have their own attributes, and we use these attributes within the UniArab system. For example, verbs in the Arabic language agree with their subjects in gender. Arabic words are masculine and feminine; there is no neutral gender. In the UniArab system we record the gender associated with a verb in the syntax for a particular subject NP. Adjectives and demonstratives also agree with the subject in gender too. In Arabic, words come into three categories with regards to number. They are:

(1)   *Singular*, indicating one
(2)   *Dual*, indicating two
(3)   *Plural*, indicating three or more.

The UniArab system records these attributes of gender and number. It is important to understand that source language specific features may not be used, or may be significantly different, in the target language. For example, the Arabic number category of *dual* is not relevant in English. The UniArab system is directly based on RRG and uses logical structures for each verb in the lexicon.

3.2   Challenges of Arabic to English MT

Arabic words can often be ambiguous due to the three-letter root system. Most words are derived from a three-letter root that is modified to create the different derivations. In some morphological derivations one or more of the root letters is dropped, resulting in possible ambiguity. Arabic has a large set of morphological features (Al-Sughaiyer and Al-Kharashi 2004). These features are normally in the form of prefixes

or suffixes that can completely change the meaning of the word (see Figures 3 and 4). This means an MT may need to apply a thorough analysis in order to obtain the root or to deduce that in one 'word' there is in fact a full sentential proposition.

Figure 3: The root and pattern characteristics of Arabic

Figure 4: The tri-consonantal roots and word formation in Arabic

Arabic has a relatively free word order (Figure 5) and this poses a significant challenge to MT due to the vast possibilities to express the same sentence in Arabic. For the elements of subject (S), verb (V) and object (O), Arabic's relatively free word order allows the combinations of SVO, VSO, VOS and OVS. For example, consider the following word orders: (1) V N N and (2) N V N. This means that we have a challenge to identify exactly which are the **subject** and the object. An example of the RRG layered structure of the Arabic clause is presented in Figure 6.

- Arabic has **free word order**:

  - **Verb** Noun Noun

Brian rides the bus.	*brāyn yrkb ālḥāflh* براين يركب الحافلة
Brian rides the bus.	*yrkb brāyn alḥaflh* يركب براين الحافلة

  - Noun **Verb** Noun

- Arabic has SVO, VSO, VOS and OVS orders in sentences.

- No copula verb in Arabic; 'to be' or 'to have'

I am Irish.	*anā āyrlndy* أنا ايرلندي
we are students.	*nḥn tlāmyd* نحن تلاميذ
he is an engineer.	*hw mhnds* هو مهندس

Figure 5: The challenges of Arabic for MT

A simple RRG projection of an Arabic clause showing both constituents and operators within the different levels of the layered structure of the clause

SENTENCE
CLAUSE
Periphery → CORE
(ARG) (ARG) NUCLEUS ARG
ADV PP NP PRED NP
V
البارحه yesterday
لعمر to-Omar
الكتاب the-book
يُري show
لم not
جون John
V
NUCLEUS
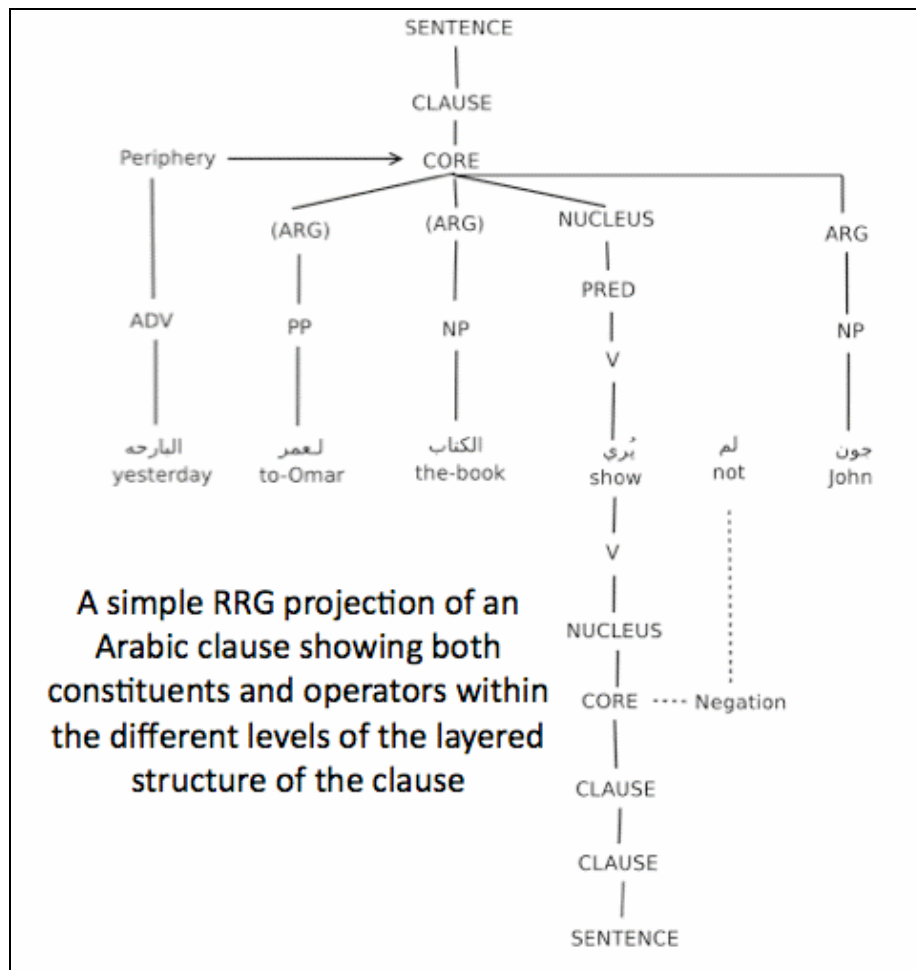CORE ···· Negation
CLAUSE
CLAUSE
SENTENCE

Figure 6: The layered structure of the Arabic clause

248

4 The UniArab System

UniArab is a proof-of-concept system supporting the fundamental aspects of Arabic, such as the parts of speech, agreement and tenses. UniArab stands for Universal Arabic machine translator system. UniArab is based on the linking algorithm of RRG (syntax to semantics and vice versa). The conceptual structure of the UniArab system is shown in Figure 7. The system accepts Arabic as its source language. The morphology parser and word tokenizer have a connection to the lexicon, which holds all attributes of a word. UniArab was developed in the Java programming language with the lexicon encoded in XML.

UniArab stores all data in XML format. This data can then be queried, exported and serialized into any format the developer wishes. The system can understand the part of speech of a word, agreement features, number, gender and the word type. The syntactic parse unpacks the agreement features between elements of the Arabic sentence into a semantic representation (the logical structure) with the `state of affairs' of the sentence. In UniArab we have a strong analysis system that can extract all attributes from the words in a sentence.

The structure of the UniArab system in Figure 7 breaks down into the several phases, which are described following.
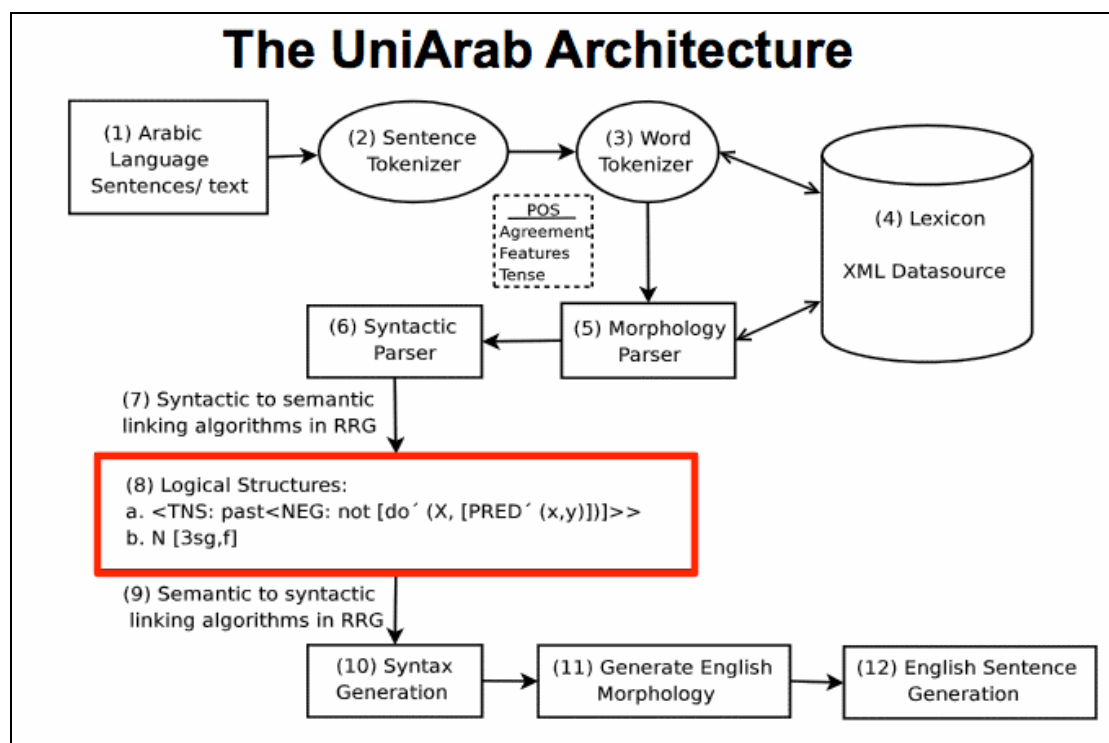
Figure 7: The conceptual architecture of the UniArab system

Phase (1) Input of Arabic language sentence: The input to the system consists of one or more sentences in Arabic.

Phase (2) Sentence Tokenizer: Tokenization is the process of demarcating and classifying sections of a string of input characters. In this phase the system splits the

text into sentence *tokens*. The resulting tokens are then passed to the word tokenizer phase.

Phase (3) Word Tokenizer: In this phase sentences are split into tokens. For example, for the Arabic sentence (4a), read from right to left, the output (4b) of phase 3 is as follows.

(4)   a.   qr'a ḫāld ālktāb   *'Khalid read the book'.*
      b.

              <sentence>
                  <word$> qr'a </word>
                  <word$> ḫāld</word>
                  <word$> ālktāb </word>
              </sentence>

Phase (4) Lexicon XML Data-source: A set of XML documents for each component category of Arabic. More details will be in sections 6 and 7.

Phase (5) Morphology Parser: Directly works with both the Lexicon and Tokenizer to produce the word order. A connection is made to the data-source of phase 4, which has been implemented as a set of XML documents. The use of XML has the added advantage of portability. UniArab will effectively work the same regardless of the operating system. To understand the morphology of each word, we first tokenize each sentence and determine the word relationships. Phase 5 of the system holds all attributes specific to each word of the source sentence.

Phase (6) Syntactic Parser: Determines the precise phrasal structure and category of the Arabic sentence. At this point, the types and attributes of all words in the sentence are known.

Phase (7) Syntactic linking (RRG) We must first develop the link from syntax to semantics out of the phrasal structure created in Phase 6, if we are to create a logical structure that will generate a target language and also act as the link in the opposite direction from semantics to syntax. The system should answer the main question in this phase, **who does what?** In this case the actor is *Khalid* and the undergoer is *the book*, as in (4) above.

Phase (8) Logical Structure: The creation of logical structure is the most crucial phase. An accurate representation of the logical structure of an Arabic sentence is the primary strength of UniArab. The results of the parse can be seen in the following logical structure for the verb 'read'

(5)   a. $\langle$TNS:PAST[**do'**(x,[**read'**(x,(y)])]$\rangle$
      b.   Verb 'read':   sg 3rd.m PAST qr'a
                  where :the Proper Noun is: *Khalid* sg unspec.m: ḫāld
                  and   the Noun is: *the book* sg def.m: ālktāb.

We also have the challenge of inferring the indefinite article, from the information unpacked in phase (5) and phase (6), as this does not exist in Arabic. All of the unique information for each word can thus be taken from the lexicon to aid in the creation of

a logical structure of the target language.

Phase (9) Semantic to Syntax: Assuming we have an input and have produced a structured syntactic representation of it, the grammar can map this structure from a semantic representation. In this phase the system uses a linking algorithm provided by RRG, to determine actor and undergoer assignments, assign the core arguments and assign the predicate in the nucleus. We determine the grammatical subject by analysing the agreement marking on the verb and the various nominals. The system uses the semantic arguments of logical structures.

Phase (10) Syntax Generation: The generation phase from the Interlingua Bridge meta-representation to the morphosyntax of a particular target language will, of course, depend on the characteristics of the target language. In our proof-of-concept software, we generate to grammatically correct English (see also phases 11 and 12, below). The generation phase implements the RRG semantics-to-syntax linking system.

Phase (11) Generate English Morphology: The system generates English morphology in an innovative way, generating the tenses that are not existent in Arabic but which do exist in English as well as the copula verb of 'to be' correctly, as appropriate. Our solution is to recognize the difference between morphological features and syntactic functional categories. The tense features must be determined analytically, and expressed correctly for the target language, in this instance, English.

Phase (12) English Sentence Generation: The process of generating an English sentence can be as simple as keeping a list of rules and these rules can be extended through the life of the MT system. The system will apply some operations in English such as vowel change in the lexical item of English to denote sg vs. pl, for example, m**a**n vs. m**e**n. Sometimes this accompanies affixations: br**ea**k/br**o**ke/br**oken** (=broke+en) to denote various tense and aspect distinctions.

Having described the various Interlingua phases, we now discuss in more detail, in the next section, the GENERATION from meta-representation (i.e., the logical structure) to target language.

5 UniArab - Generation

The target language generation phases in the UniArab system follow the syntactic realization model. Generation takes as input, the universal logical structure of the input sentence(s) and produces, as output, the grammatically correct morphosyntax of the target language. The UniArab system is a universal machine translator, which means that it can translate Arabic into any other natural language. The UniArab system is evaluated using Arabic as source language into English as the target language.

In the UniArab system phases 9, 10, 11 and 12 are for generation of the target languages, in our case this is English. First, the *Semantic to Syntactic* phase determines the actor and undergoer assignments, assigns the core arguments and assigns the predicate in the nucleus. The system uses semantic arguments of logical structure. In the UniArab system we keep all word attributes whether they are used in the target language or not. In this case, the gender of the noun *the book*, in Arabic is

masculine, but in English *book* has neutral gender. In Phase 10, *Syntax Generation*, and Phase 11, *Generate English Morphology*, UniArab uses target language rules to generate the syntax. The verb's logical structure indicates to UniArab how many arguments the verb takes. For example, the logical structure will be as in (6a), from the lexicon. Now the UniArab system replaces **x** with *Khalid*, and **y** with *the book*, after which it now holds the following (6b):

(6) a. **Read**: [**do'**(x, [**read'**(x, (y))])]
     b. **Read:** [**do'**(Khalid, [**read'**(Khalid, the book)])]

In the last phase, *English Sentence Generation*, the UniArab system builds the final shape of a sentence: *Khalid read the book*. Moreover, there are some special cases, like the UniArab system adding the copula verb 'to be' into the English copula sentence, or changing the source language verb's tense to an appropriate and grammatically correct tense in the target language, depending on the tense distinction in the target language. Also, the word order in the target language must be considered and applied correctly.

6 An XML-based lexicon

In order to build this system and represent the data sources, we use the Java with the XML language (Bray et al., 2008). XML has become the default standard for data exchange among heterogeneous data sources (Arciniegas, 2000). The UniArab system allows data to be stored in XML format. This data can then be queried, exported and serialized into any format the developer wishes. We choose to create our data source as XML, for optimum support on different platforms. It was also easier as we used Arabic letters, not Unicode, inside the data source, and XML fully supports Arabic. We created our search engine for the lexicon using Java. The lexicon is represented as an XML data object

6.1 Advantages of XML

XML gives us a generalized way to store data, which is not married to any particular technology. This makes it easy to store information, and retrieve and manage it later, as required. Using XML to manage information offers a number of advantages, including the following:

(7)
1. **Easily build:** A well formed data element must be enclosed between tags. The XML document can be parsed without prior knowledge of the tags. XML allows one the possibility of defining ones own application relevant tags, such as tags representing data description or data relationships, in our situation to do with lexical items.

2. **Human readable:** Using intelligible tag names make it possible for the XML to be easily read by people as well as software.

3. **Machine-readable:** XML was designed to be easy for computers to process. XML is completely compatible with Java, and is portable. Any application can process XML on any platform, as it is a platform-independent language.

252

4. **XML fully supports Arabic:** We chose to create our data-source as XML files, for optimum support of different platforms. It was also easier as we used Arabic letters rather than Unicode inside the data-source.

5. **XML search engine:** It is easy to extend the search sample to display more information about the search. Search via the Java API Document Object Model (DOM) was found to be the ideal tool for searching collections of XML documents, that is, our lexicon.

6.2 Lexicon interface

In order to allow for robust user interaction with the lexicon, we use a graphical interface to capture the information for each part of speech. The user selects the part of speech of the word to be added, and is then presented with only the attributes relevant to the selected part of speech. The interface also limits the user's selections to acceptable values and ensures that all attributes are filled.

With this technique, we minimize the risk of human errors, and therefore the information is more accurate. The graphical interface is quicker and easier when a user adds a new word in the lexicon within the XML data source. Figure 8 shows the entry interface that is implemented as part of the UniArab system.

Figure 8: The Lexicon Interface of UniArab

7 Lexical representation in UniArab

Lexical frames represent the language-dependent lexicon. We use an XML data source to represent the UniArab lexicon. The lexicon creates pointing references to corresponding conceptual frames with associated attributes for each word. These frames also have relations which link them to verb class frames, which are organized hierarchically according to the particular language, here, Arabic and English.

In Phase 3 of our Interlingua Bridge PARSE→GENERATE framework, the UniArab system tokenizes a sentence into words, and then sends each word to the search engine within the Lexicon to query the category of each word plus determine all attributes associated with that word. The Lexicon returns the corresponding category and its attributes. The Morphology Parser, Phase 5, receives the word metadata and ensures that the properties of the words are consistent. The verb attributes, in particular, are of critical importance in correctly extracting sentence logical structure

further down the processing chain, helping to answer the basic question `*Who does what to whom*?'

In free word order sentences of Arabic, multiple orders are possible including VSO, VOS or SVO (Figure 9). The attributes of the verb define the gender of the subject. Given a masculine gender of the verb, for example, the Syntactic Parser will look for a masculine proper noun to make the actor for this sentence. If there is more than one masculine proper noun in such a case, then Modern Standard Arabic defines the first proper noun as the actor. The Morphology Parser will, in future research, be extended so that it can deal with words that are defined in multiple categories, deciding which should be processed. Meanwhile the Syntactic Parser, so far, has only been implemented for extracting word order, though it will be extended to deal with word ambiguities in future versions.
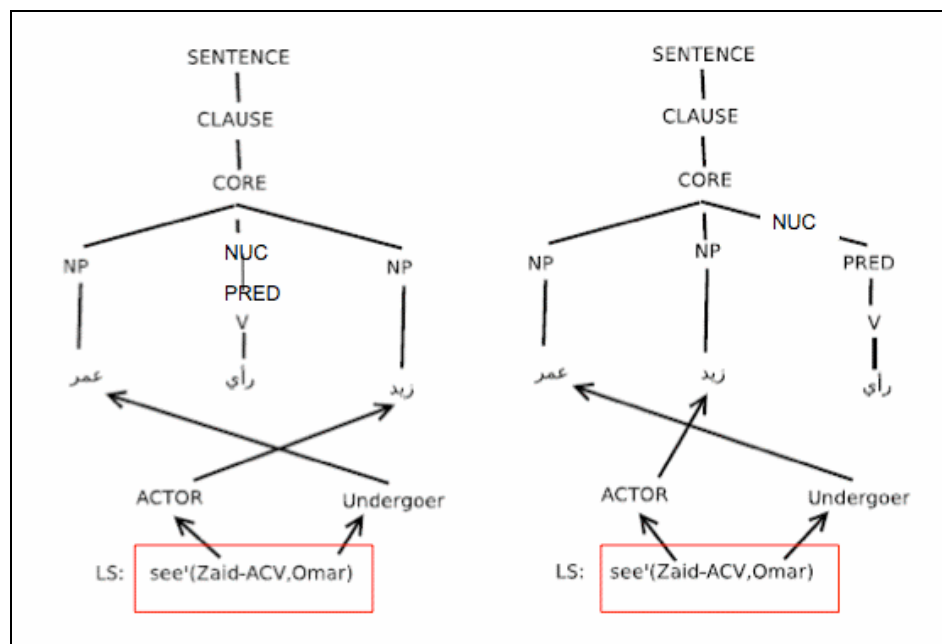
Figure 9: The linking of the Arabic clause under free word order

7.1 Lexical properties

The structure of the Lexicon including the properties stored for each word category is indicated in Figure 10. For all categories, an Arabic word is stored along with its English representation. There is an isomorphic mapping, importantly at the semantic level via the Interlingua Bridge (RRG) logical structures, from the source to the target language of non-complex sentences that UniArab processes up to now. A level of word ambiguity is supported in the structure, with each possible case stored as a separate record. All search results will be passed to the Morphology Parser to decide which is taken.

Since the verb is the key component when analysing using RRG, each verb has an associated logical structure (Figure 11), which is later used to determine the logical structure of the full sentence. The tense of the verb is also stored within its metadata along with the person.

Figure 10: Fragment - RRG lexical entry in XML for verb read

The verb type also stores the gender, which in Arabic must be either masculine or feminine; there is no neutral gender. The number property in Arabic can be singular, dual or plural. These properties help the Syntactic Parser analyse the sentence, since there must be agreement with the subject and verb, among other rules.

Figure 11: RRG lexicon in XML

We show a Java code fragment, in Figure 12, which determines the appropriate gender marking on an argument.

8 UniArab Evaluations

Evaluation of MT software is necessary in order to improve system performance and analyse potential problems and, of course, its accuracy and effectiveness. In the evaluation of UniArab we considered many different aspects of the MT system including quality of translation, time for translation, ability to add a new word in the lexicon of the system and resource utilization.

The evaluation of MT systems is a difficult task. This is not only because many different metrics are involved, but also because translation is itself difficult. The first important aspect for a potential test is to determine the translational capability.

Therefore, we needed to draw up a complete overview of the translational process, in all its different aspects.

A good translation has to effectively capture the meaning. This involves establishing the size of the translation task, is it machine legible and if so, according to which standards? Current general function MT systems cannot translate all texts consistently. Output can have very poor quality. It is to be mentioned that the '*subsequent editing required*' increases, as translation quality gets poorer (Turian et al. 2003).

```java
public static String verbGender ="NON";
public static String translate ="";
public static String LS1 ="";
public static String LS0 ="";
//public
//static String toBe1="";
static String [] allAttribute2;

public static void GenerationLS1(String [] allAttribute, String [] ArabicSentence)
{
    allAttribute2 =  allAttribute;
    int numSubject =0;
    //who do what
    for(int n = 0 ; n < allAttribute.length; n++)
    {
        if (allAttribute[n].equals("Verb"))
        {
            // there is a verb
            if (allAttribute[n+1].equals("M"))
            {
                //the gender of verb = M
                verbGender = "M";
                //Add LS to string LS
                LS0 = allAttribute[n+2];
            }
            else if (allAttribute[n+1].equals("F"))
            {
                // the gender of verb = F
                verbGender = "F";
                LS0 = allAttribute[n+2];
            }
            else if (allAttribute[n+1].equals("NoGender"))
            {
                // the gender of verb = F
                verbGender = "NoGender";
                LS0 = allAttribute[n+2];
            }


        }
    }// end loop who do what
```

Figure 12: Java code fragment that determines the appropriate gender marking

Given the scale of the lexicon implemented in this work so far, we evaluate the effectiveness and accuracy of UniArab by comparison of output results against an ideal output produced by hand by a native Arabic L1 speaker. We created variants of Arabic sentences that represent all possible structures of the sentences that UniArab can translate. We then make a comparison between human-translated and machine-translated versions. At the moment, the lexicon is categorised into seven parts of

speech. We have designed the GUI so that when adding a specific word to the lexicon, only the related options are presented to the user for that part of speech. This minimises errors when entering data. As our research extends, we expect to modify the categorisation of the lexicon to allow for more complicated word types.

UniArab does not process ambiguous words or complex sentences, so far, in this research. This research focussed first on discovering whether the logical structure of a sentence, based on RRG can be used for translation. Hence, we decided to limit the scope of the project to exclude ambiguity resolution, since this is work in a new area that has not been investigated before. We fully expect to expand the system to allow it to cope with ambiguity in the future. The system's reliability and accuracy depends on the content of the lexicon in the XML data source and cannot handle words not in the lexicon. However, it manages this intelligently by determining the 'x' and 'y' argument slots in the logical structure and inserting the (unknown) Arabic nominal into the correct slot. This native Arabic word in then carried through to the English translation. to handle unknown words. UniArab does not process single words, even if those words are in its lexicon, because UniArab is built on the logical structure of verbs. The missing or unknown word can then be easily inserted into the lexicon.

Therefore, for the processing of unrecognised Arabic words, where a word is not available in the lexicon, but the logic structure is recognised, then UniArab will output a correctly structured translation, but with the unknown Arabic word in its position within the English sentence (Figure 13). This makes the system resilient to slight misspellings (in nominals), which can be recognised and corrected by the human translator.
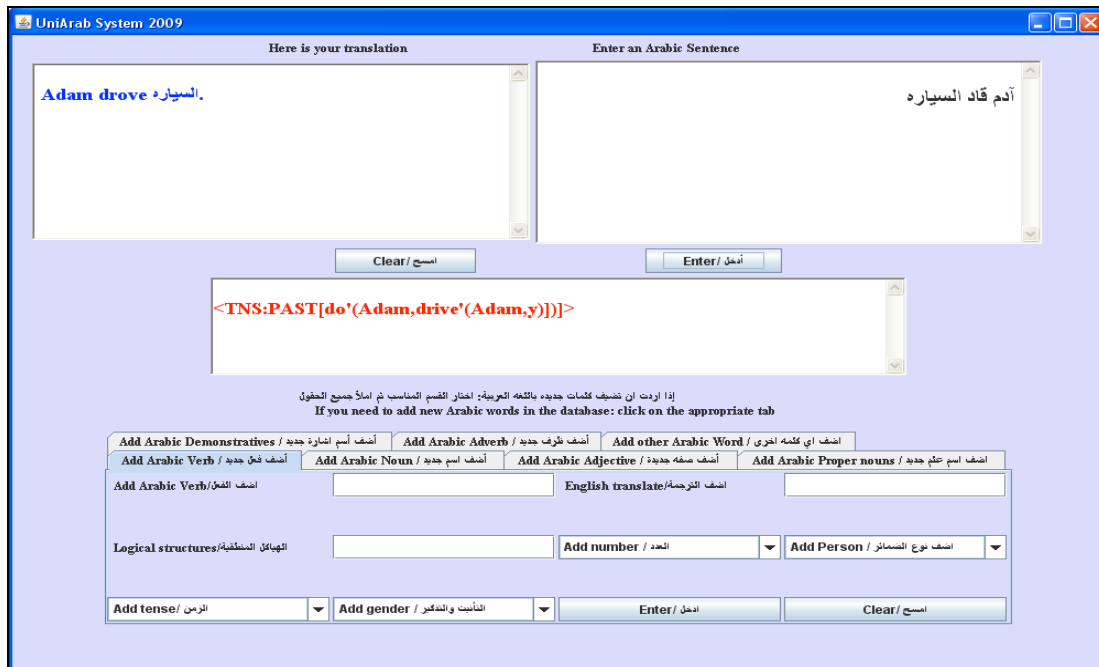
Figure 13: Processing unrecognised Arabic words

In our comparison with other translation systems we have used non-complex sentences. While UniArab is limited to non-complex sentences and has appropriate coverage within these, we believe it is essential to reach high quality translation of

these sentences in the first instance, in order to be able to expand to high quality translations of more complex sentences. We can see that the existing tools from Google and Microsoft cannot even achieve reasonable translations of simplex sentences, so how can we expect them to give high quality translations of larger text? We have found that small errors in the initial analysis of a sentence can cause huge errors in the final translation, so high quality analysis is very important.

We have MT processing of non-complex sentences in Arabic and their equivalent translations in English. By non-complex we mean any clause that does not have a juncture relation, of any kind, in RRG terms. We have covered a representative broad selection of verbs across intransitive, transitive and ditransitive constructions in simplex sentences in active voice. Complex sentences are beyond the research scope to date, but we intend to address this in the next version. However, we do address copula-like nominative clauses in Arabic. We tested UniArab in many ways. We tested single sentences and multiple sentences. UniArab easily deals with more than one sentence as input and its output matches. That is, UniArab can accept and translate a text consisting of many sentences. Additionally, we entered random sentences together in one input or as individual sentences.

In our testing and evaluation of UniArab, we subjected the UniArab System to a series of tests in a wide range of sentence categories. For each test we compared the results obtained through UniArab to those obtained when using translation engines from Google and Microsoft. We also presented a human-translated equivalent to each. In contrast, the Google and Microsoft translators gave mixed results. In many cases, sentence meaning was lacking, and even some basic constructs could not be translated.

This, perhaps, is due to their focus on translating long sentences and paragraphs via statistical means rather than using semantically oriented linguistic structured to drive the translation. We highlighted this by comparing them to UniArab for longer compound sentences and found that they did indeed convey more of the meaning. These results suggest that RRG is a promising candidate for Arabic to English machine translation, and as the grammar is developed, the system should begin to cope with more complicated sentences. For non-complex sentences (intransitive, transitive and ditransitive) it clearly outperforms existing systems for the production of grammatically correct translations.

In summary then, with respect to our evaluation, given the proof of concept work implemented so far, we were very careful to rigorously test and evaluate the performance of UniArab, and its accuracy in the fast production of grammatically correct sentences in the target language. We created a testbed of sets of sentences in Arabic to represent all of the possible combinations of structures and possibilities for the sentences that we wanted UniArab to be able to translate. We then executed UniArab for these and compared our results with that of a human L1 Arabic translator. We also tested the Google and Microsoft automatic machine translation services with our data set of sentences to compare our UniArab results against all of these, with some very interesting and surprising results.

Our testbed of grammatically correct sentences in Arabic and their equivalent translations in English have a good coverage and we tested UniArab with these. We additionally tested inputs of both single sentences and multiple sentences (as in a

paragraph of Arabic text). UniArab is designed to easily deal with more than one sentence as input and its output correctly and grammatically matches.

9 The accuracy of the translations

In this section we review the accuracy of the translations and compare the results of our system, UniArab based on the RRG linguistic model, with results from Google and Microsoft. While not rehearsing the complete set of evaluations here, in summary, our testing has include the following (8) testing criteria:

(8)   UniArab Evaluation tests
1. Evaluation test-1: Copula with present progressive
2. Evaluation test-3: Verb noun - one argument in different tenses
3. Evaluation test-3: Generating the English copula verb 'to be'
4. Evaluation test-4: Free word order (**V** N N - first possibility)
5. Evaluation test-5: Free word order (**V** N N - second possibility)
6. Evaluation test-6: Free word order (N **V** N - third possibility)
7. Evaluation test-7: Pro-drop sentence
8. Evaluation test-8: Intransitive sentences
9. Evaluation test-9: DTV word order
10. Evaluation test-10: DTV word order (with prepositional phrase)

We provide indicative sample outputs, plus a screen capture that demonstrates the actual results, for a number of these in Appendix 1.

9 Summary

In this paper we have presented an Arabic-to-English machine translation system called UniArab, based on our implementation of an Interlingua Bridge framework that was programmed in Java with the lexicon built in XML, and which is based on the Rule and Reference Grammar model. We detailed the design of the system and how it was built to accommodate specifics of the Arabic language and the generation of English translations. We presented a high-level view of the system framework and defined our evaluation criteria for measuring system performance. We also talked about the challenges of machine translation, with a specific focus on those specific to the Arabic language. The main topic of investigation is the development of a framework for translating Arabic to English based on RRG. The framework is designed to demonstrate the capabilities of RRG as a base for machine translation. This work has shown that RRG facilitates the translation process from a specific source human language to other target languages.

10 References

Al-Sughaiyer, I. A. and Al-Kharashi, I. A., 2004. *Arabic morphological analysis techniques: A comprehensive survey*. JASIST, 55(3): 189–213.

Alosh, Mahdi. 2005. *Using Arabic: A Guide to Contemporary Usage*. Cambridge: Cambridge University Press.

Arciniegas, F., 2000. *XML Developer's Guide*. McGraw-Hill Companies.

Attia, M., 2004. Report on the introduction of Arabic to ParGram. In *Proceedings of ParGram Fall Meeting, Dublin, Ireland*.

Attia, M. A., 2008. *Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation*. PhD thesis, University of

Manchester.

Bateson, M. C., 2003. *Arabic Language Handbook*. Georgetown University Press.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., and Yergeau, F., 2008. *Extensible Markup Language (XML) 1.0 (Fifth Edition)*.

Brown, P., Pietra, S. D., Pietra, V. D., and Mercer., R., 1993. The mathematics of statistical machine translation: Parameter estimation. In *Computational Linguistics*, volume 19 i2, pages 263–311.

Holes, C., 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press.

Hutchins, W. J. 2003. *Machine translation: General overview*. Oxford: Oxford University Press.

Hutchins, J. and Somers, H. L., 1992. *An introduction to machine translation*. Academic Press.

Izwaini, S. 2006. *Problems of Arabic machine translation: evaluation of three systems*. The British Computer Society (BSC), London, pages 118–148.

Nolan, Brian (ed.). 2004. *Linguistic Theory and Practice: Description, Implementation and Processing*. Proceedings of the RRG 2004 Dublin Conference. (Available on RRG web site.)

Owens, J., 2006. *A Linguistic History of Arabic*. Oxford University Press.

Pustejovsky, James. 1995: *The Generative Lexicon*. Cambridge, MA: MIT Press

Ramsay, A. and Mansour, H., 2006. Local constraints on Arabic word order. In *5th International Conference on NLP, FinTAL 2006*, pages 447–457.

Ryding, K. C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Salem, Yasser., Arnold Hensman and Brian Nolan. 2008a. Implementing Arabic-to-English Machine Translation using the Role and Reference Grammar Linguistic Model. In *Proceedings of the Eighth Annual International Conference on Information Technology and Telecommunication (IT&T 2008)*, Galway, Ireland, October 2008. Available at:
http://www.ittconference.ie/prevconf/2008/Conference%20Proceedings%20ITT08.pdf

Salem, Y., Hensman, A., and Nolan, B., 2008b. Towards Arabic to English machine translation. *ITB Journal*, 17:20–31.

Salem, Y. and Nolan, B., 2009a. Designing an XML lexicon architecture for Arabic machine translation based on Role and Reference Grammar. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR 2009), Cairo, Egypt*.

Salem, Y. and Nolan, B., 2009b. UNIARAB: An universal machine translator system for Arabic Based on Role and Reference Grammar. In *Proceedings of the 31st Annual Meeting of the Linguistics Association of Germany (DGfS 2009)*.

Schulz. Eckehard 2005. *A Student Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Trujillo, A., 1999. *Translation Engines: Techniques for Machine Translation*. Springer.

Turian, J. P., Shen, L., and Melamed, I. D., 2003. Evaluation of machine translation and its evaluation. In *Proceedings of the MT Summit IX, New Orleans, USA*, pages 386–393.

Van Valin, R., 1993. *Advances in Role and Reference Grammar*. John Benjamins.

Van Valin, Robert D. 2005. *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press.

Van Valin, R. and LaPolla, R. 1997. *Syntax: Structure, Meaning, and Function*.

Cambridge: Cambridge University Press.

Versteegh, K., 2001. *The Arabic Language*. Edinburgh University Press; New Edition.

Appendix-1: Evaluation test results

Evaluation Test-1: Copula with present progressive

In Table 1, the output of the Google translator is faulty in respect of the tense and the marking of the *V-ing* form in English, and the non-use of the copula verb *to be*. Microsoft's MT failed to translate most of the sentence with respect to tense, copula vs. matrix verb and word order. UniArab successfully translates the sentence entirely. Figure 13 shows this sentence output in the UniArab system.

Evaluation Test-2: Verb noun - one argument in different tenses

Arabic	شرب عمر اللبن šrb ᶜmr āllbn
Human	Omar drank the milk
Google	Omar drinking milk
Microsoft	drinking milk Omar
UniArab	Omar drank the milk.

UniArab System 09

Here is your translation

Omar drank the milk.

Enter an Arabic Sentence

شرب عمر اللبن

Clear / امسح   Enter / أدخل

<TNS:PAST[do'(Omar,[drink'(Omar,milk])]>

إذا اردت ان تضيف كلمات جديدء باللغه العربية: اختار القسم المناسب ثم املأ جميع الحقول
If you need to add new Arabic words in the database: click on the appropriate tab

Add Arabic Demonstratives / أضف أسم اشارة جديد   Add Arabic Adverb / أضف ظرف جديد   Add other Arabic Word / اضف اي كلمه اخرى

Add Arabic Verb / أضف فعل جديد   Add Arabic Noun / أضف اسم جديد   Add Arabic Adjective / أضف صفه جديدة   Add Arabic Proper nouns / اضف اسم علم جديد

Add Arabic Verb / اضف الفعل   English translate / اضف الترجمة

Logical structures / الهياكل المنطقية   Add number / العدد   Add Person / اضف نوع الضمائر

Add tense / الزمن   Add gender / التأنيت والتذكير   Enter / ادخل   Clear / امسح

262

Evaluation Test-3: Copula Verb 'to be'

Arabic	هو مهندس *hw mhnds*
human-translated	He is an engineer.
Google	Is the architect of
Microsoft	is the engineer
UniArab	he is an engineer.

Evaluation Test-4: Free word order (V N N - first possibility)

Arabic	يحب قيس ليلى *yḥb qys lylā*
human-translated	Qays loves Laila
Google	Qais likes of Laila
Microsoft	Love Qais laili
UniArab	Qays loves Laila

Evaluation Test-5: Free word order (V N N - second possibility)

Arabic	يحب ليلى قيس *yḥb lylā qys*
human-translated	Qays loves Laila
Google	Leila loves measured
Microsoft	Love laili Qais
UniArab	Qays loves Laila

Evaluation Test-6: Free word order (V N N - third possibility)

Arabic	قيس يحب ليلى *qys yḥb lylā*
human-translated	Qays loves Laila
Google	Qais likes of Laila
Microsoft	Qais love laili
UniArab	Qays loves Laila

266

Evaluation Test-7: Pro-Drop

Arabic	فاتتني الطائرة *fāttny ālṭāᵃyrh*
human-translated	I missed the plane.
Google	Missed the plane
Microsoft	فاتتني *fāttny* plane
UniArab	I missed the plane.

Evaluation Test-8: Intransitive sentences

Arabic	صهيب يقرأ *shyb yqr'a*
human-translated	Suhaib reads.
Google	Souhaib read
Microsoft	suhaib reads
UniArab	Suhaib reads.

Evaluation Test-9: DTV word order

Arabic	عمر أعطى خالد الكتاب *ʕmr ʕaṭā ḫāld ālktāb*
human-translated	Omar gave Khalid the book.
Google	Omar Khalid gave the book
Microsoft	Omar gave Khaled book
UniArab	Omar gave Khalid the book.

Evaluation Test-10: DTV word order (with prepositional phrase)

Arabic	عمر أعطى لخالد كتاب *mr ʿaṭā lḥāld ktāb*
human-translated	Omar gave a book to Khalid.
Google	Omar Khaled gave a book
Microsoft	Omar gave Khalid book
UniArab	Omar gave a book to Khalid.