Institute of Technology Blanchardstown
Institiúid Teicneolaíochta Baile Bhlainséir

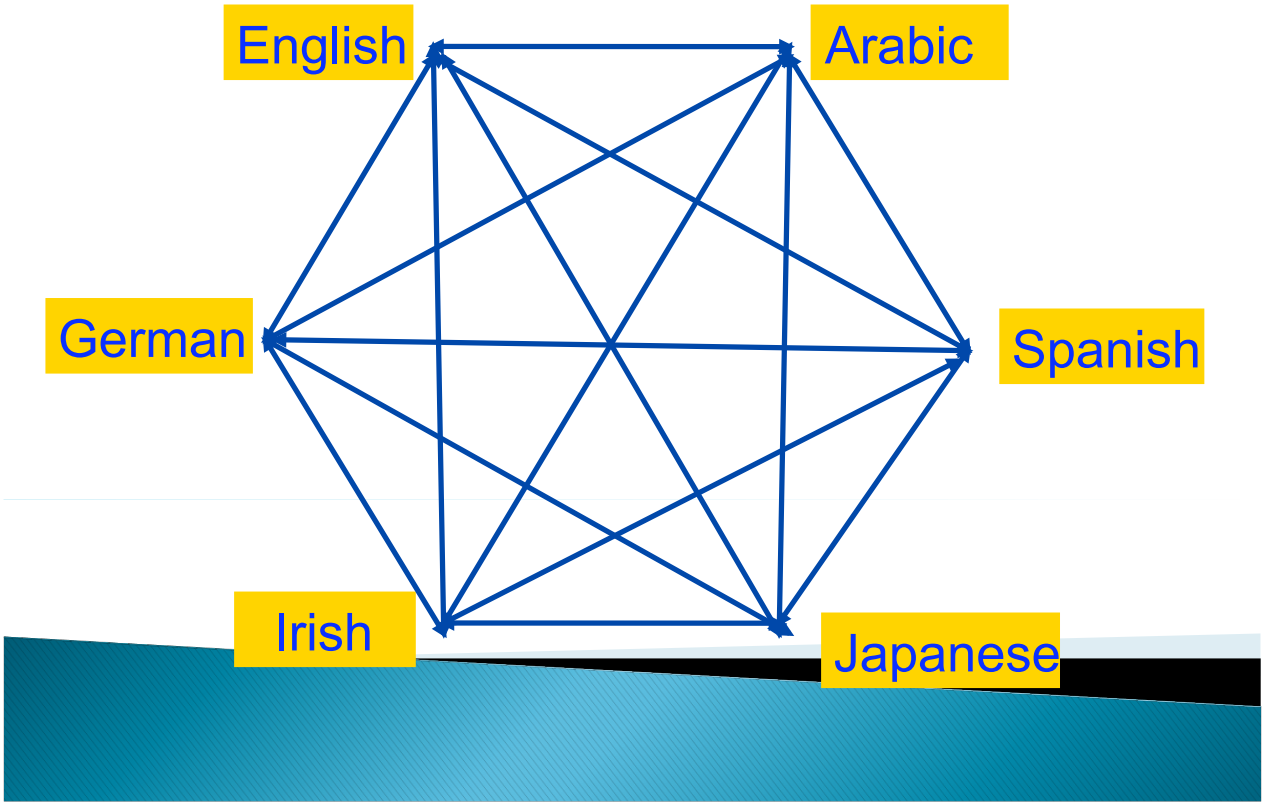
UniArab: An RRG Arabic-to-English Machine Translation Software

Dr. Brian Nolan
Yasser Salem

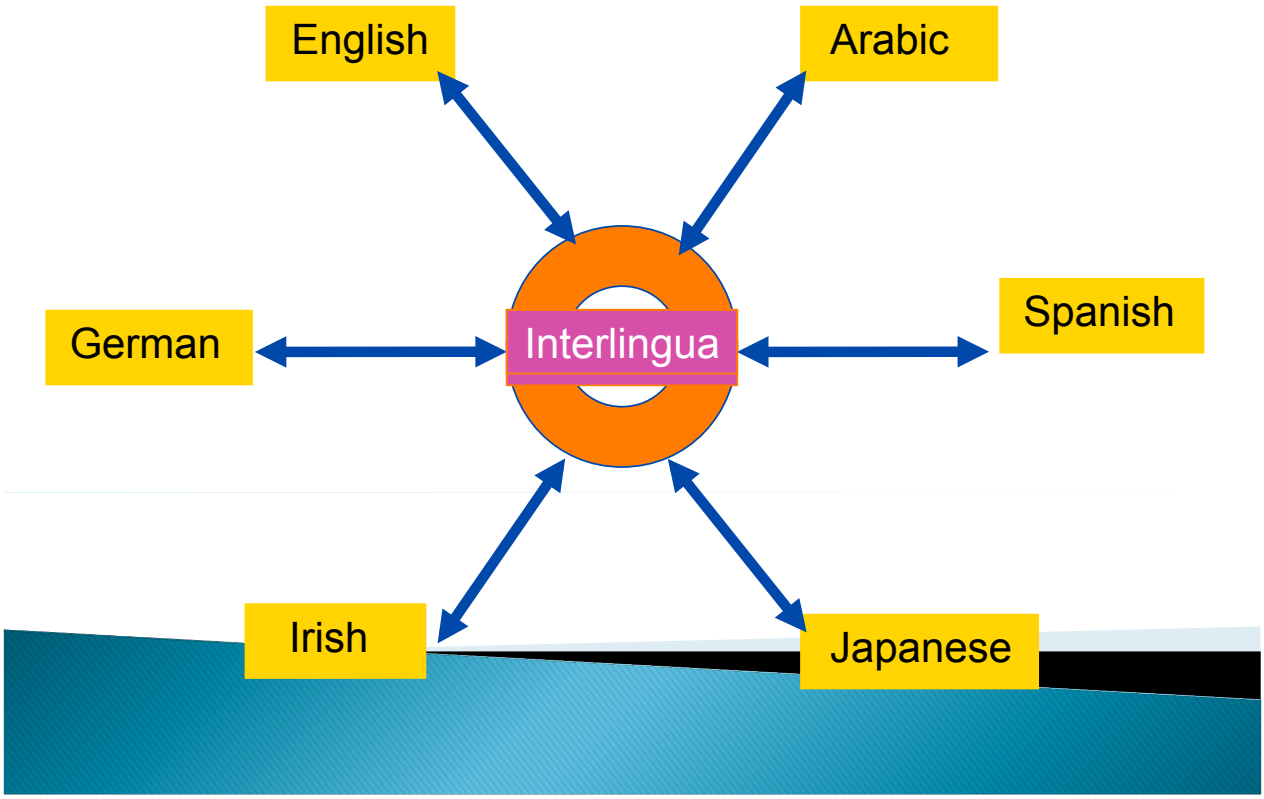
Introduction

- The RRG UniArab MT project
  1. An Interlingua model of Arabic MT
  2. Challenges of Arabic MT
  3. The Role and Reference Grammar (RRG)
  4. The UniArab System
- Summary and future Work

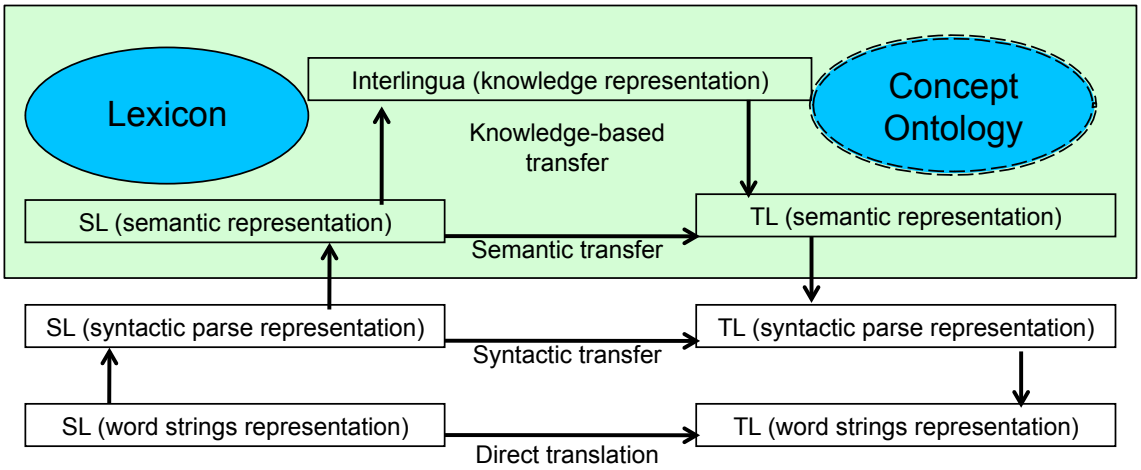
MT – Transfer approach



MT - Interlingua approach



Interlingua

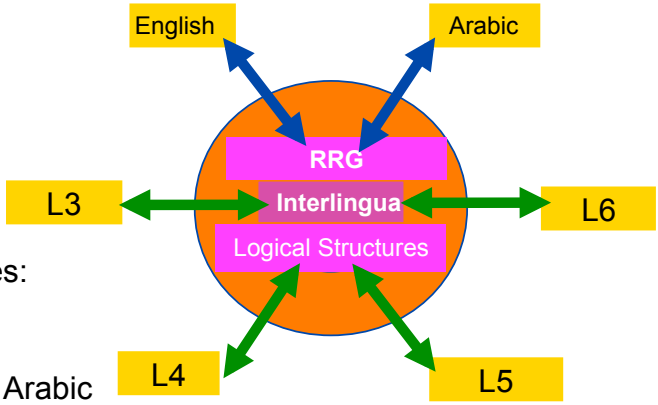


SL: source language
TL: Target language

Interlingua - considerations

- **Universal lexicon**
  - How do we construct a lexicon?
  - Must we include all distinctions made by any language?
  - How to differentiate similar items? (similar verbs in a class: 'shake vs vibrate)
- **Universal knowledge format**
  - How do we encode 'knowledge'?
  - What should be include? (pragmatic information?)
- **Preservation of ambiguity**

An Interlingua Model of Arabic MT



- The Interlingua model incorporates:
  - source language analysis
  - universal logical structure for Arabic
  - Interlingua kernel based on the **Role and Reference Grammar** model
- This model is more flexible and scalable for multiple languages generation.

Challenges of Arabic MT

- Large set of morphological features.

  These features are normally in the form of prefixes or suffixes that can completely change the meaning of the word.

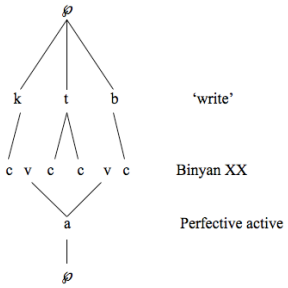
- In Arabic there are some words that hold the meaning of a full sentence e.g., سنسافر, would translate to “we will travel” in English.

The free word order makes determining *who did what* non-trivial to a computer system.

Challenges of Arabic MT

- **Root**: a relatively invariable discontinuous bound morpheme, typically three consonants in a certain order, which interlocks with a pattern to form a stem and which has **lexical** meaning.
- **Pattern**: a bound (maybe discontinuous) morpheme consisting of one or more vowels and slots for root phonemes, which either alone or in combination with one to three derivational affixes, interlocks with a root to form a stem, and which generally has **grammatical meaning**.
- **Patterns** signify grammatical or language-internal information, distinguishing word types and classes.
- These patterns can differentiate between nouns, verbs and adjectives.

Arabic	Example	POS
كَتَبَ *kataba*	he wrote	verb
كاتَبَ *kātaba*	he corresponded	verb
كُتِبَ *kutiba*	it was written	verb
كِتاب *ktiāb*	book	noun
كُتُب *kutub*	books	noun
كاتِب *kātib*	writer; (adj) writing	noun
كُتّاب *kutāb*	writers	noun
مَكْتَب *maktab*	desk; office	noun
مَكاتِب *makātib*	desks; offices	noun
مَكْتَبَة *maktabah*	library	noun

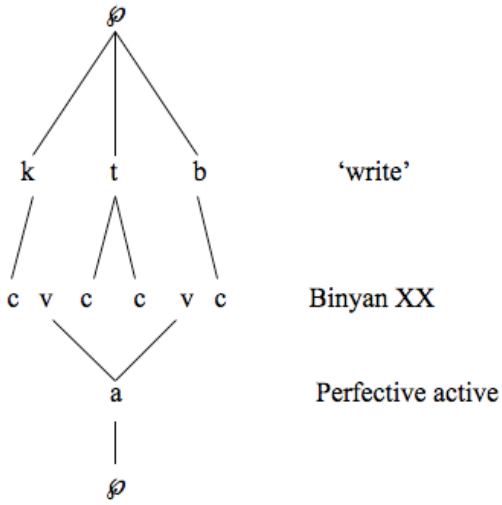


kattab: 'caused to write'

root: .k.t.b. + pattern: a.a.a ➔*kataba*
.c.c.c. v.v.v

Challenges of Arabic MT

- The **root and pattern word formation** patterns in the Semitic languages have been treated by segmenting morphemes on different auto-segmental 'tiers' or 'planes' (cf. McCarthy 1982).
- Tri-consonantal **roots** which bear the basic lexical meaning of the verb occupy one tier while vocalic melodies (the **pattern**) occupy a separate tier.
- The two tiers are organised by association with a 'template' or 'skeleton' consisting of syllables of prosodic structure within words.



kattab: 'caused to write'

Challenges of Arabic MT

- Arabic has **free word order**:
  - **Verb** Noun Noun
  - Noun **Verb** Noun

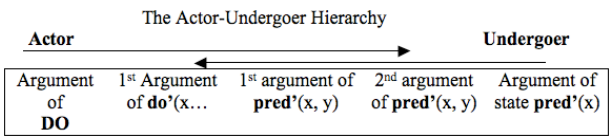
Brian rides the bus.	براين يركب الحافلة *brāyn yrkb ālḥāflh*
Brian rides the bus.	يركب براين الحافلة *yrkb brāyn ālḥāflh*

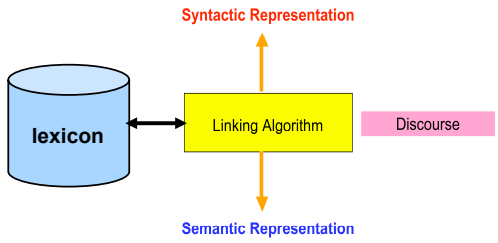
- Arabic has SVO, VSO, VOS and OVS orders in sentences.
- No copula verb in Arabic; 'to be' or 'to have'

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

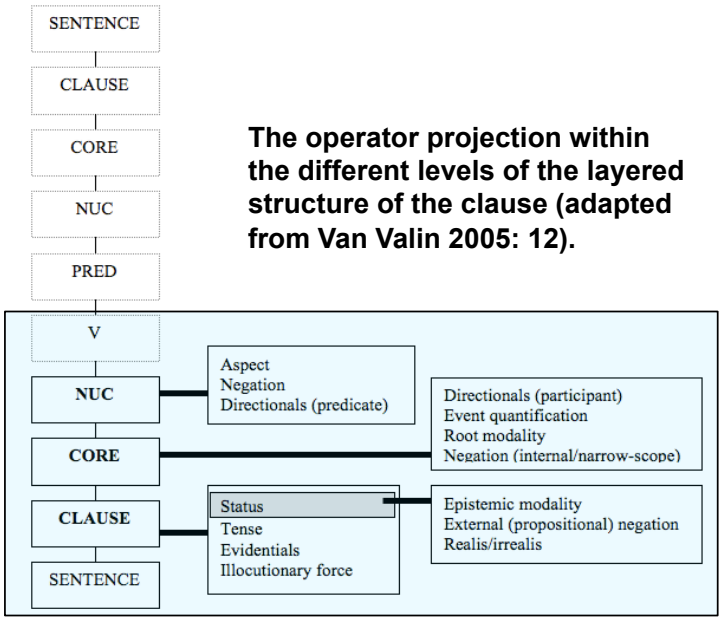
Role and Reference Grammar

- RRG describes a sentence of a specific language through its:
  - Logical structure and the actor-undergoer hierarchy
  - Application of the (bidirectional) linking system and argument structure
  - Layered structure of the clause
  - Layered structure of the noun phrase
  - Layered structure of the word
- We adopt RRG as the base for a multi-language interlingua translator system.





Role and Reference Grammar



The operator projection within the different levels of the layered structure of the clause (adapted from Van Valin 2005: 12).

(1) Verbal operators (Van Valin 2005)

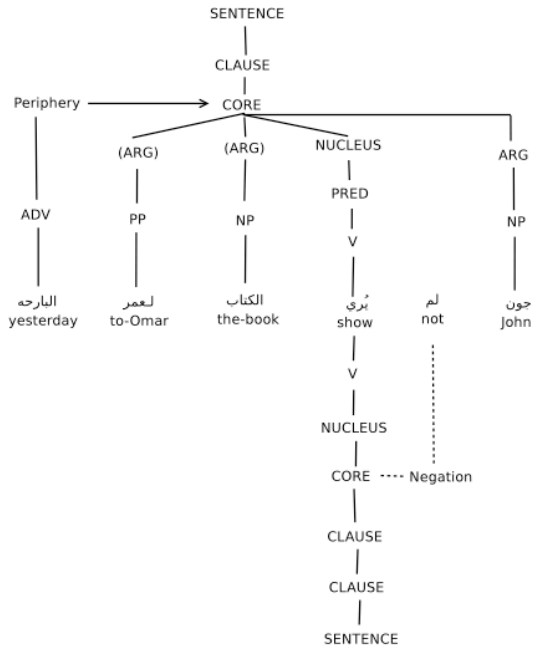
Nuclear$_V$ operator
- Aspect
- Negation
- Directionals (predicate)

Core$_V$ operators
- Directionals (participant)
- Event quantification
- Root modality
- Negation (internal/narrow-scope)

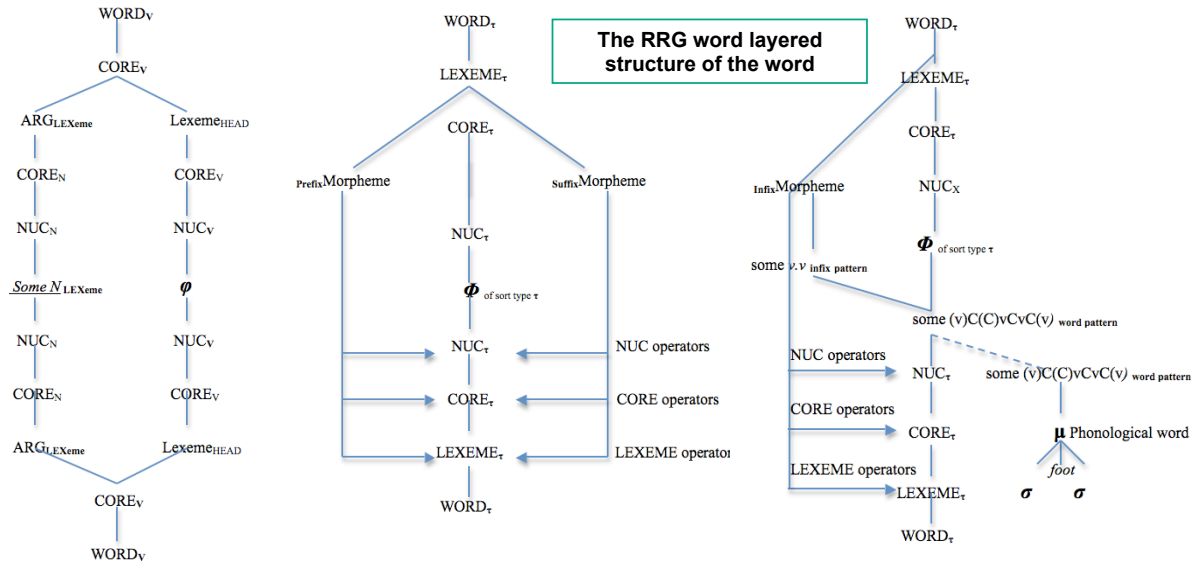
V operators
- Tense
- Evidentials
- Illocutionary force

Role and Reference Grammar

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



Role and Reference Grammar



The RRG word layered structure of the word in a derivation

The RRG word layered structure of the word in syntactic inflection

The generalised RRG word layered structure of the word for infixation

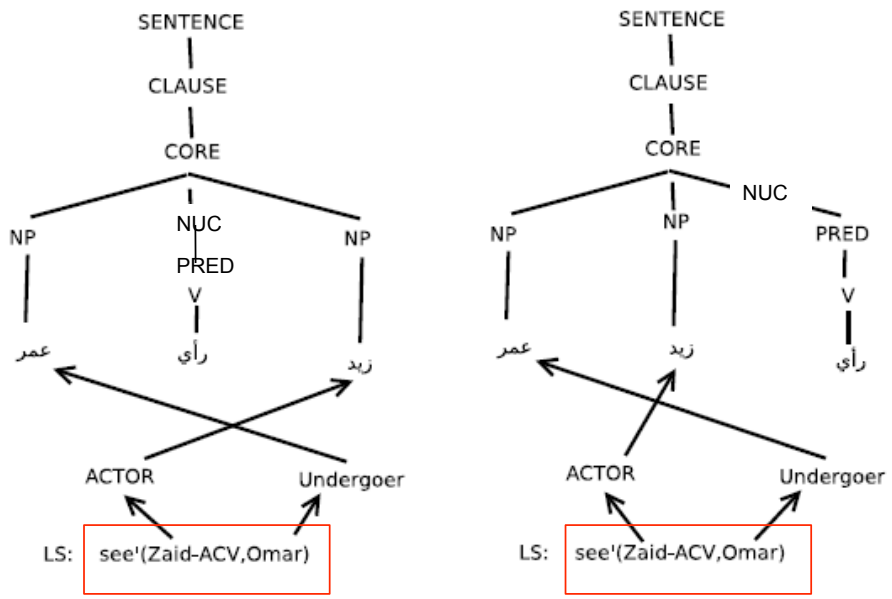
Role and Reference Grammar

- Arabic allows variation in clause word order.

	3	2	1			1	2	3
a.	عمر	رأي	زيد	*zyd ray ʿmr*		Zaid	saw	Omar
زيد	*zyd*	MsgNOM	see.past	عمر	*ʿmr* - MsgNOM			
b.	عمر	زيد	رأى	*raā zyd ʿmr*		Saw	Zaid	Omar
see.past	زيد	*zyd*	MsgNOM	عمر	*ʿmr* MsgNOM			

- (a) is a SVO sentence and
- (b) is the VSO equivalent.

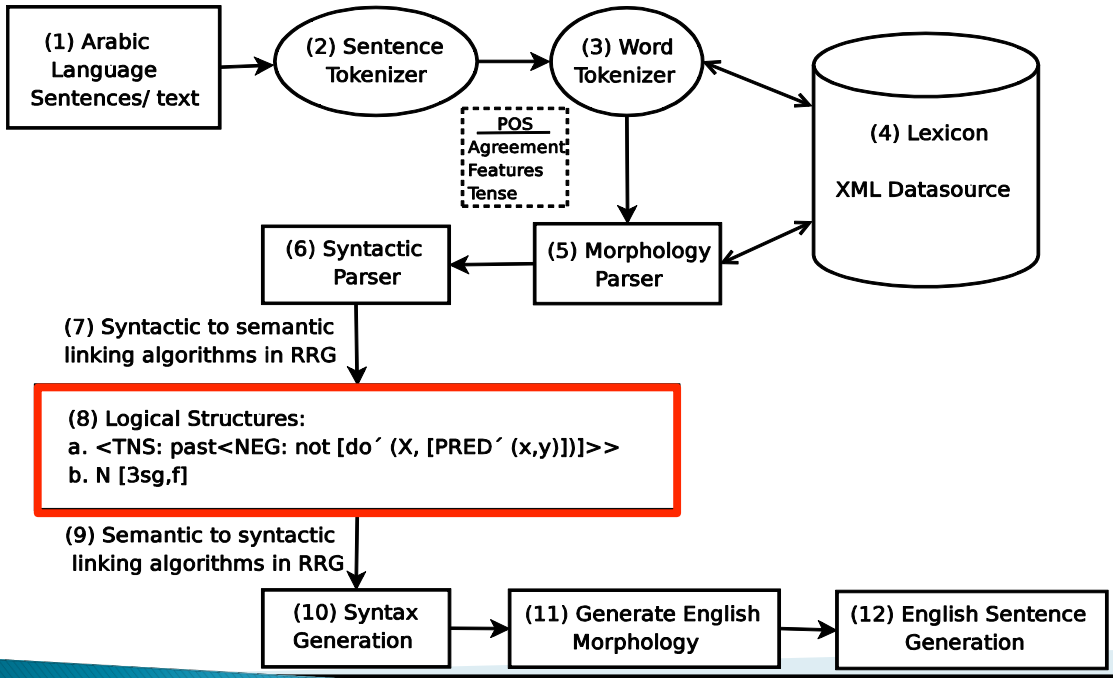
Role and Reference Grammar



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 **Uni**versal **Arab**ic machine translator system.
- **UniArab** is based on the linking algorithm of RRG (syntax to semantics and vice versa).

The UniArab Architecture



The UniArab Architecture

- **Phase (8)** – Universal (RRG) Logical Structure Creation is the most crucial phase of the UniArab system.

'Khalid read the book' الكتاب خالد قرأ *qrʾa ḫāld ālktāb*

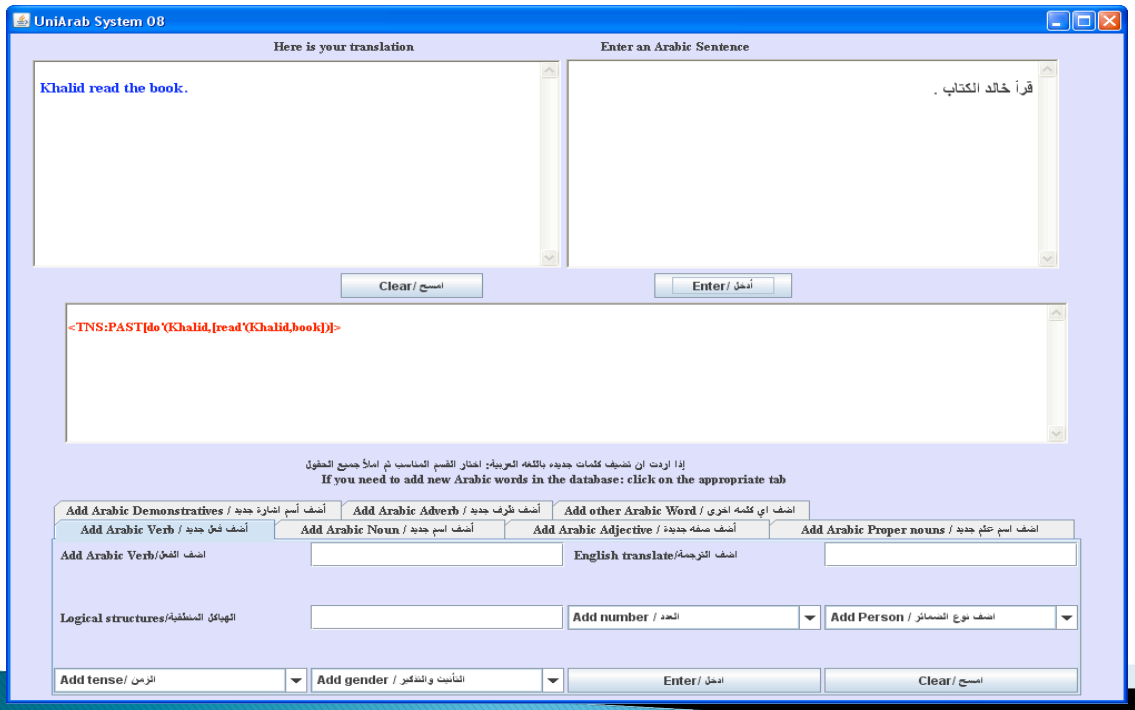
الكتاب *ālktāb* book:N خالد *ḫāld* Khalid:MsgN قرأ *qrʾa* read:V

<TNS:PAST[do'(x,[read'(x,(y)])]>

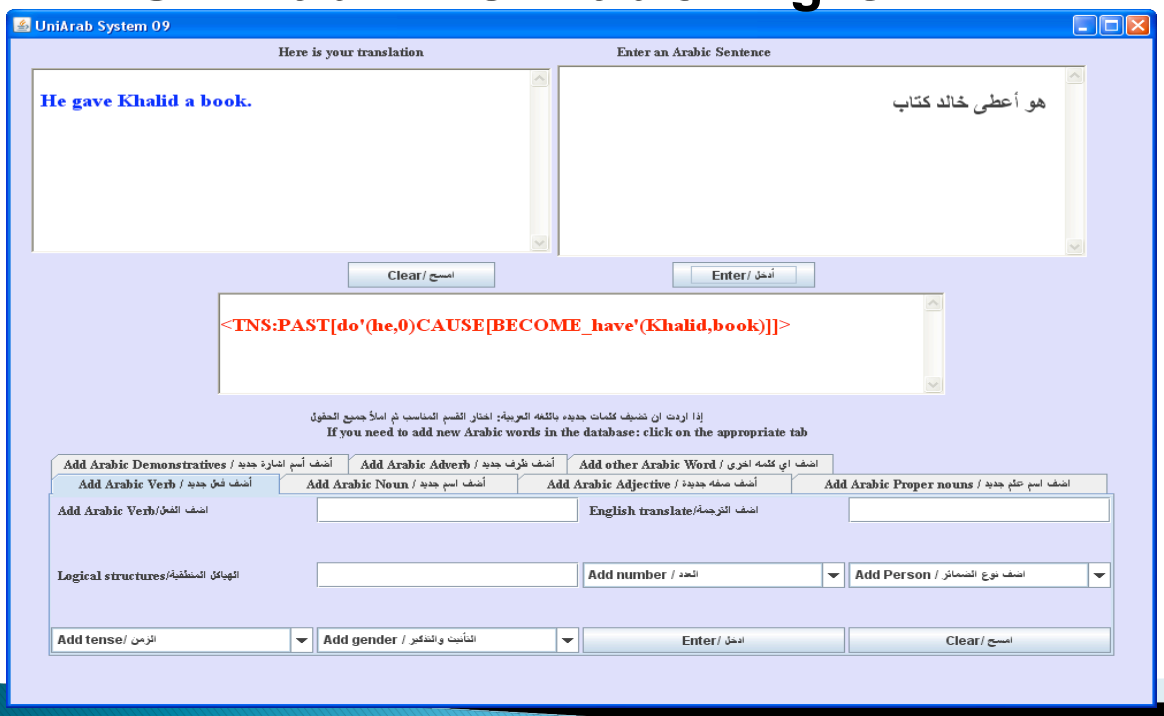
- The results of the parse can be seen here with LS output of our system:

<TNS:PAST[do'(Khalid,[read'(Khalid,book)])]>

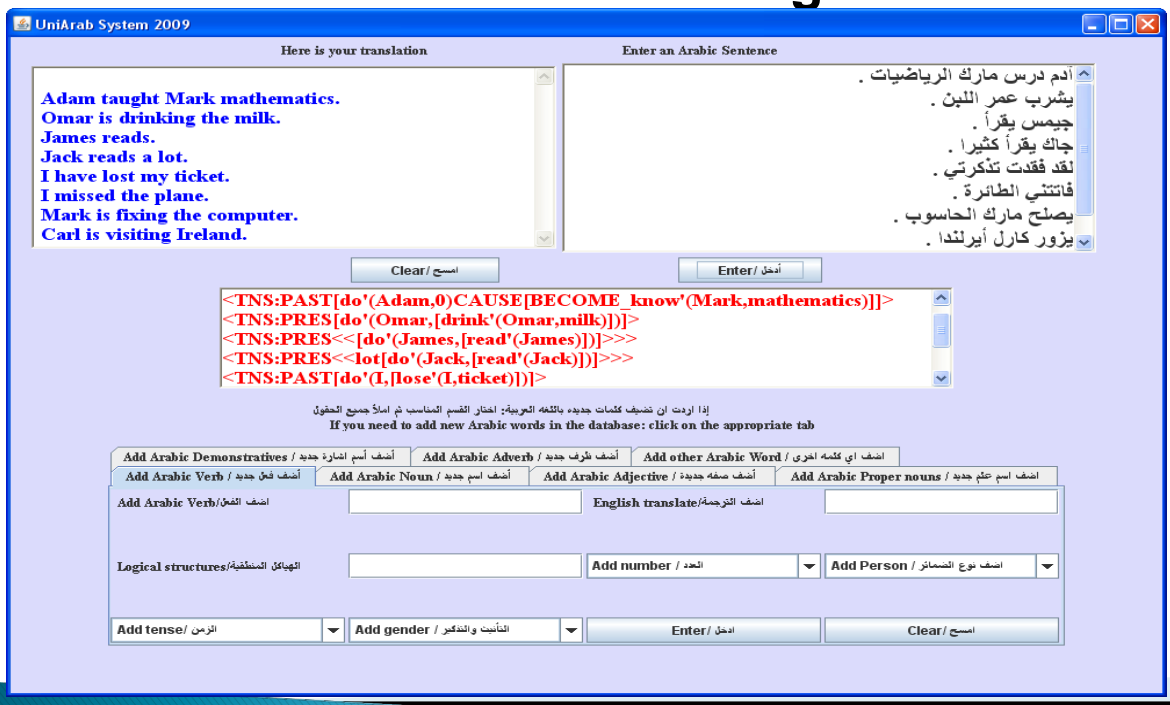
UniArab: RRG Arabic-English MT



UniArab: RRG Arabic-English MT



UniArab: RRG Arabic-English MT



UniArab: RRG Arabic-English MT

UniArab: RRG Arabic-English MT

Processing unrecognised Arabic words

In a case:

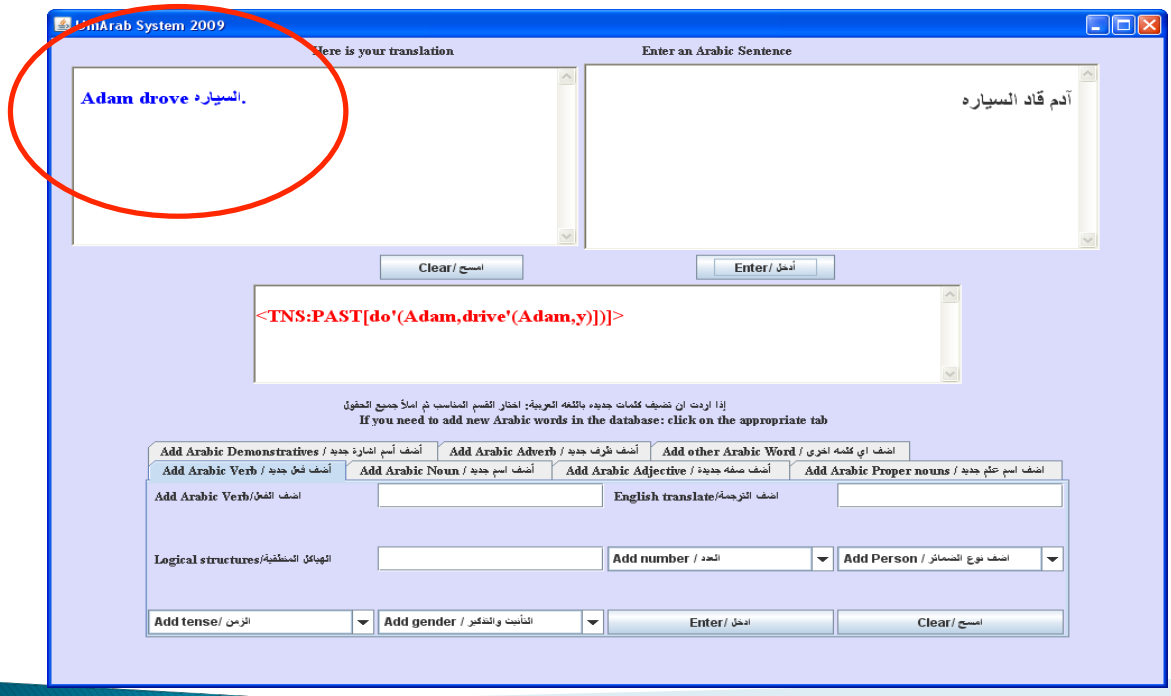
1. where a word is not available in the lexicon
2. but the logic structure is recognised

➔

UniArab will output a correctly structured translation, but with the unknown Arabic word in its position within the English sentence.

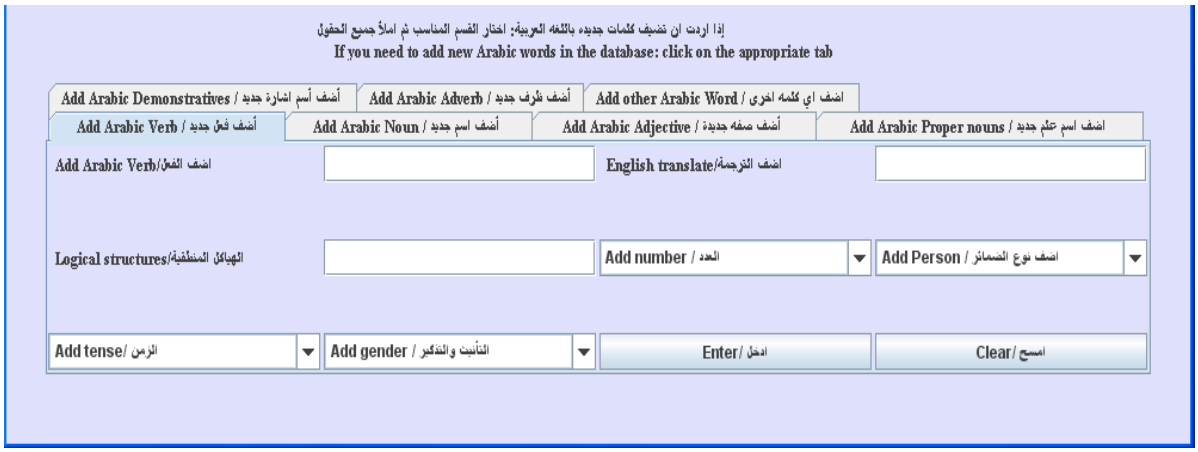
- This makes the system resilient to slight misspellings which can be recognised and corrected.

UniArab – processing unrecognised words



UniArab's Lexicon Interface

The graphical interface is quicker and easier when a user adds a new word in the lexical (XML data source).



UniArab – testing and evaluation

- Test strategy
  - verb-subject agreement
  - Demonstrative adjective-noun agreement
  - Gender-ambiguous proper nouns
  - 'to be' and 'to have'
  - Free word order (Verb Noun Noun)
  - Pro–drop

Arabic	human-translated
قرأ جاك الكتاب *qrʿa ǧāk ālktāb*	Jack read the book.
قرأت ماري الكتاب *qrʿat māry ālktāb*	Mary read the book.

Arabic	human-translated
يشرب عمر الحليب *yšrb ʿmr ālḥlyb*	Omar is drinking the milk.
شرب عمر الحليب *šrb ʿmr ālḥlyb*	Omar drank the milk.
مارك قرا الكتاب *mārk qrā ālktāb*	Mark read the book.
سيشرب مارك اللبن *syšrb mārk āllbn*	Mark will drink the milk

UniArab – testing and evaluation

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.

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

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

Arabic	أنا المهندس *ʿanā ālmhnds*
human-translated	I am the engineer.
Google	I Engineer
Microsoft	i am engineer
UniArab	I am the engineer.

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

Arabic	لقد قمت بالحجز *lqd qmt bālḥǧz*
human-translated	I have made a reservation.
Google	I have made a reservation
Microsoft	You have a booking
UniArab	I have made a reservation.

Arabic	يصلح مارك الحاسوب *yṣlḥ mārk ālḥāswb*
human-translated	Mark is fixing the computer.
Google	Mark works computer
Microsoft	Marc computer works
UniArab	Mark is fixing the computer.

Arabic	مارك يري آدم الرسالة *mārk yry ʿādm ālrsālh*
human-translated	Mark is showing Adam the letter.
Google	Mark Adam see the letter.
Microsoft	Marc finds Adam message.
UniArab	Mark is showing Adam the letter.

Arabic	هو أعطى خالد كتاب *hw ʿaṭā ḫāld ktāb*
human-translated	He gave Khalid a book.
Google	Khaled was given a book
Microsoft	is given Khaled book
UniArab	He gave Khalid a book.

Arabic	عمر أعطى لخالد كتاب *ʿmr ʿaṭā lḫāld 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.

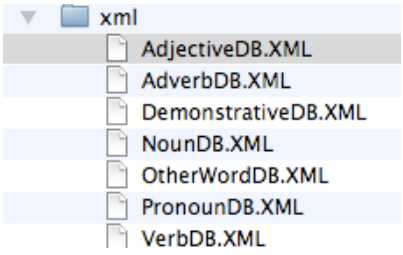
RRG Lexicon in XML

- Verb lexicon - fragment as XML

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<Verbs>
</"EnglishTranslate="read" LogicalStructures="&lt;TNS:PAST[do'(x,[read'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST قرأ>
</"EnglishTranslate="wrote" LogicalStructures="&lt;TNS:PAST[do'(x,[write'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="F" personVerb="3rd" tenseVerb="PAST كتبت>
</"EnglishTranslate="played" LogicalStructures="&lt;TNS:PAST[played'(x,y)]]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST لعب>
</"EnglishTranslate="wrote" LogicalStructures="&lt;TNS:PAST[do'(x,[write'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST كتب>
</"EnglishTranslate="loves" LogicalStructures="&lt;TNS:PRES[do'(x,[love'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PRES يحب>
</"EnglishTranslate="loves" LogicalStructures="&lt;TNS:PRES[do'(x,[love'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="F" personVerb="3rd" tenseVerb="PRES تحب>
</"EnglishTranslate="drank" LogicalStructures="&lt;TNS:PAST[do'(x,[drink'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST شرب>
```

- Noun lexicon - fragment as XML

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<Nouns>
</"EnglishTranslate="book" Number="sg" PartOfSpeach="Noun" gender="M" type="Indef كتاب>
</"EnglishTranslate="the book" Number="sg" PartOfSpeach="Noun" gender="M" type="def الكتاب>
</"EnglishTranslate="books" Number="pl" PartOfSpeach="Noun" gender="M" type="Indef كتب>
</"EnglishTranslate="the books" Number="pl" PartOfSpeach="Noun" gender="M" type="def الكتب>
</"EnglishTranslate="free" Number="sg" PartOfSpeach="Noun" gender="M" type="Indef حر>
</"EnglishTranslate="the food" Number="pl" PartOfSpeach="Noun" gender="M" type="def الطعام>
</"EnglishTranslate="his food" Number="sg" PartOfSpeach="Noun" gender="M" type="def طعامه>
</"EnglishTranslate="his hair" Number="sg" PartOfSpeach="Noun" gender="M" type="def شعره>
</"EnglishTranslate="his teeth" Number="pl" PartOfSpeach="Noun" gender="M" type="def أسنانه>
</"EnglishTranslate="his clothes" Number="pl" PartOfSpeach="Noun" gender="M" type="def ثيابه>
</"EnglishTranslate="his shoes" Number="sg" PartOfSpeach="Noun" gender="M" type="def حذائه>
</"EnglishTranslate="the television" Number="sg" PartOfSpeach="Noun" gender="M" type="def التلفاز>
```



RRG Lexicon in XML

<قرأ

EnglishTranslate="read"

LogicalStructures= "<TNS:PAST[do'(x,[read'(x,y)])]>"

NumberVerb="sg"

P.O.S="Verb"

genderVerb="M"

personVerb="3rd"

tenseVerb="PAST"

/>

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<Verbs>
</"EnglishTranslate="read" LogicalStructures="&lt;TNS:PAST[do'(x,[read'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST قرأ
</"EnglishTranslate="wrote" LogicalStructures="&lt;TNS:PAST[do'(x,[write'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="F" personVerb="3rd" tenseVerb="PAST كتبت
</"EnglishTranslate="played" LogicalStructures="&lt;TNS:PAST[played'(x,y)]]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST لعب
</"EnglishTranslate="wrote" LogicalStructures="&lt;TNS:PAST[do'(x,[write'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST كتب
</"EnglishTranslate="loves" LogicalStructures="&lt;TNS:PRES[do'(x,[love'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PRES يحب
</"EnglishTranslate="loves" LogicalStructures="&lt;TNS:PRES[do'(x,[love'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="F" personVerb="3rd" tenseVerb="PRES تحب
</"EnglishTranslate="drank" LogicalStructures="&lt;TNS:PAST[do'(x,[drink'(x,y)])]&gt;" NumberVerb="sg" P.O.S="Verb" genderVerb="M" personVerb="3rd" tenseVerb="PAST شرب
```


Java code Fragment

....

Builds LS based on verb

This fragment determines appropriate gender marking

```
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
```

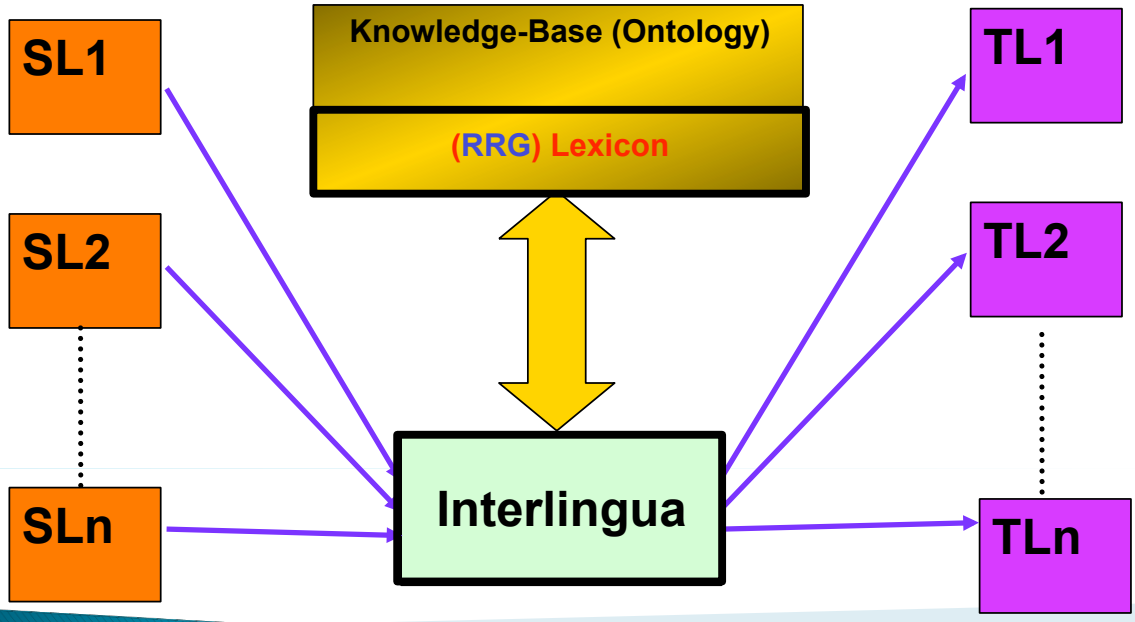
Summary and Future work

- We have discussed our **rule-based lexical framework** for Arabic language processing using RRG.
- The framework is supported by a system called UniArab.
- UniArab is an RRG motivated XML-based machine translator implemented using Java.
- Our framework and system demonstrates the capabilities of RRG as a basis for multi-language interlingua machine translation systems.

Summary and Future work

- **Our main future work:**
  - **Supporting** more complex Arabic sentences and word ambiguity.
  - **Comparing** UniArab against MT systems based on non-RRG techniques, such as systems provided by Google and Microsoft.
  - **Evaluating** the performance of UniArab with multi-languages processing.
  - **Adding** extra languages with interlingua bridges.
  - **Connecting** the lexicon with a rich concept **ontology** to enhance the granularity of the logical structures

Linguistics and KB Interlingua



Linguistics and KB Interlingua

Linguistics-Based Interlingua:

- **Linguistic model** (+ its semantic and linking theory) identifies meaning representations and constraints and hence provide a sufficient basis for an interlingua representation.

Knowledge-Based Interlingua:

- **Linguistic meaning** is dependent on non-linguistic knowledge.
- Use **real world knowledge** to augment meaning representations.

Institute of Technology Blanchardstown
Institiúid Teicneolaíochta Baile Bhlainséir

Thank you ……. !

Questions?

brian.nolan@itb.ie