# Towards Semantic Composition of ARABIC: A λ-DRT Based Approach

**Bassam Haddad[†] and Mustafa Yaseen[†]**

Petra University[†]
Amman, Jordan
bhaddad@go.com.jo
myaseen@ammanu.edu.jo

## Abstract

This paper addresses issues related to employing logic-based semantic composition as a meaning representation for Arabic within a unification-based syntax-semantics interface. Since semantic representation has to be compositional on the level of semantic processing λ-calculus based on Discourse Representation Theory can be utilized as a helpful and practical technique for the semantic construction of ARABIC in Arabic understanding systems. As ARABIC computational linguistics is also short of feature-based compositional syntax-semantics interfaces we hope that this approach might be a further motivation to redirect research to modern semantic construction techniques for developing an adequate model of semantic processing for Arabic and even no existing formal theory is capable to provide a complete and consistent account of all phenomena involved in Arabic semantic processing.

Keywords: λ-DRT, Discourse Structures, Compositionality, Unification-based Syntax-Semantics, ARABIC NLP, HDPSG.

## 1 Introduction & Motivation

For the last two decades concentration on Arabic processing has focused on Arabic from the morphological and syntactical point of view. In this field, some success has been achieved (Beesely 2001), (Ouersighni, 2001), (Ditters, 2001), (Al-Fedaghi and Al-Anzi, 1989) and many others.

Despite the importance of semantic processing for achieving the understanding capability, there were little works reported on semantic representation and semantic analysis of Arabic (Haddad and Yaseen, 2001), (Al-Johar and McGregor, 1997), (El-Dessouki et al., 1988) and (Al-Muhtaseb and Mellish, 1997). Therefore, we believe that there is an elemental need to make more effort to develop an adequate model for semantic processing for Arabic and even no existing formal theory is capable to provide a complete and consistent account of all phenomena involved in Arabic semantic processing.

One of the most important levels of Semantic processing is the construction and composition of meaning representation formalisms for Arabic sentences. This semantic level plays a decisive role in the whole semantic processing progression. In this paper important issues related to this level are directly addressedto this level are directly addressed.

Syntax-Semantics interfaces using unification-based or feature-based formalisms are increasingly common in the existing computational linguistics literature. HDPSG (Pollard and Sag, 1994) related system development is ongoing in numerous university and industrial settings for different languages. Unfortunately there are very limited HDPSG deep analyses for Arabic. HDPSG is based on GPSG and shares the other related grammatical frameworks such as LFG and Categorial Grammar with their most important criteria (Uszkoreit et al., 2000). In such a grammar the lexicon plays a pivotal role, where *semantics* and *syntax* can be integrated in the same grammar for expressing deep linguistic analysis.

We propose that the simulating of the λ-conversion process using logical feature structures within a unification-based grammar such as HDPSG enables us to achieve a practical technique for a compositional unification-based semantic framework for Arabic.

Inspired from the work of (Bos et. al., 1994), (Nerbonne, 1993), (Fischer, 1993) we propose relying on a λ-DRT implementation of ARABIC (Haddad, 2002) to integrate the semantic construction model presented in (Haddad and Yaseen, 2001) in a unification-based grammar.

## 2 Logical Representation for ARABIC

Assuring the modularity constraint in a natural language understanding system requires a compositional semantic formalism on the level of meaning representation. Despite the fact that standard predicate logic represents well-studied formal representation formalism, it does not provide any compositional facilities. λ-calculus offers an important framework for achieving such a goal but merely for the meaning construction of Arabic sentences (Haddad and Yaseen, 2001), (Montague, 1974).

In this context we have achieved some success in developing a model for the construction of meaning representation forms for Arabic sentences. Based on some *compositional rules* expressing the meaning of *syntactical categories of Arabic*, our approach employs a λ-conversion *process* to construct logical forms representing the meaning of Arabic sentences (Haddad and Yaseen, 2001).

In this model *determiners* play a central role in constructing semantic constituents. For example, the Arabic determiners such as (/ال$_n$/ "the$_n$ ", $n \in N$), (/كل /, "all"), (/بعض / "some") and others, could be considered as kind of *Arabic generalized quantifiers Language AGQL*[1].

Generally the meaning of a quantifier, ||Quant||, can be expressed as follows:

$$\|\text{Quant}\| \Rightarrow \lambda R \lambda S(Quantifier(R, S)) \quad (1)$$

The Arabic definite determiner referring to one object (/ال$_1$ /, "the$_1$") combines basically two things together: a restriction *R* and a scope *S:*

||ال$_1$|| ⇒ $\lambda R \lambda S$(ال$_1$($x$, $R \wedge S$)) (2)
||The$_1$|| ⇒ $\lambda R \lambda S(\text{The}_1\ (x, R \wedge S))$

The following example might illustrate the basic concept of this approach. The function of the definite determiner (/ال$_1$ /, The$_1$) in the sentence (/يتعلمُ) (/الولدُ العربية, *"studies the boy the Arabic")* can be formulated as follows:

$$\|\text{VS}\| \xrightarrow{sem} \|\text{Subj}\| (\|\text{Obj}\| (\|\text{Verb}\|)) \quad (3)$$

[1]More details about these different types of Arabic Generalized Quantifiers are found in AGQL "ARABIC Generalized Quantifiers Languages" (Haddad, 2002) and in (Haddad and Yaseen, 2001)

Applying of (3) to (||ال$_1$|| "||The$_1$||") yields the following logical representation:

$\lambda R \lambda S$(ال$_1$($x$, $R \wedge S$)) (||ولدُ||)(||يتعلمُ ال عربية||) (4)
$\lambda R \lambda S(\text{The}_1(x, R \wedge S))$ (||boy||)(||studies the Arabic||)

$\lambda S$(ال$_1$ ($x$, ولد ($x$) $\wedge$ $S$)) (||يتعلمُ ال عربية||) (5)
$\lambda S(\text{The}_1(x, \text{boy}(x) \wedge S))$ (||studies the||)

ال$_1$($x$, ولد ($x$) $\wedge$ ال$_1$($y$, عربية($y$) $\wedge$ يتعلمُ ($x$,$y$))) (6)
$The_1(x, boy(x) \wedge The_1(y, Arabic(y) \wedge study(x,y)))$

In this example there are two generalized quantifiers of type (/ال$_1$ /, The$_1$) represented in a nested generalized quantifier.

## 3 DRT-Based Semantics for Arabic

In spite of the importance of logic-based *compositional models* for achieving Arabic understanding, such methods are rather constructed to deal with *Arabic sentence semantics* and in general they are inappropriate for treating text semantics (Haddad and Yaseen, 2001), (Al-Johar and McGregor, 1997)

The Discourse Representation Theory (DRT*)* is capable of capturing problems involved in representing anaphoric aspects and text semantics (Kamp, 1981), (Kamp and Reyle, 1993), Bende-Farkas and Kamp, 2001). In this approach the semantic function of sentences consists in constructing of Discourse Representation Structures (DRS's) by applying dynamically certain *DRS construction rules* within the context of the *referents* in the sentences.

For instance, the function of a definite article seems in the view of DRT, not in interpreting it as a unique quantifier. It has rather to be understood as a referent to a certain object in a nominal expression. Moreover, the interpretation of the indefinite articles appear in the first place not to be interpreted as existential quantifiers. An indefinite article or *indefinite Arabic article indications* introduce rather new referents to the context.

In addition, one of the most important aspects of DRT is its interesting interpretation of pronouns and in particular the *incorporated Arabic pronouns*. The interpretation of a pronoun is not a variable, which has to be locally bound, but much more as a definite label making a reference to a previously introduced discourse referent.

Therefore, a DRT-based semantic construction formalism of Arabic has to be in the first place not in constructing the logical meaning in an isolated mode but much more in a dynamic and modifiable one considering the characteristics of Arabic.

The following example might illustrate some of these observations:

**Example**: (/يدرسُ ايمن لغة يحبُها, **"**Ayman studies a language, he-likes-it")

| |
|---|
| |
| Empty DRS |

| x y e n | |
|---|---|
| ايمن*(x)* | *"Ayman(x)"* |
| لغة*(y)* | *"language(y)"* |
| *e:* يدرس*(x,y)* | *"study(x,y)"* |
| $e \subseteq n$ | |
| DRS for يدرسُ ايمن لغة | |
| DRS for "Ayman studies a language | |

| x y e n z w s | |
|---|---|
| يدرس*(x)* | *"study(x)"* |
| لغة*(y)* | *"language(y)* |
| *e:* يدرس*(x,y)* | *" e:study(x,y)"* |
| $e \subseteq n$ | |
| *s:* يحب*(z,w)* | *"s: like(z,w)"* |
| $s \subseteq n$ | |
| $z = x$ | |
| $e \subseteq s$ | |
| $w = y$ | |
| DRS for "يدرسُ ايمن لغة يحبُها" | |
| DRS for "studies Ayman a language he-likes-it" | |
| e: an event, n: time of speech | |

**Figure 1**. A facet of the dynamic DRS construction process for (/يدرسُ ايمن لغة يحبُها/, "Ayman studies a language, he-likes-it"). The pronoun and pronoun indication are incorporated in the verb (/يحبُها/, he-likes-it). Quantifiers are not involved in this example.

The interpretation of this discourse starts with an empty DRS. After interpretation of the first part of the sentence (/يدرسُ ايمن لغة/, "Ayman studies a language"), the DRS is expanded by adding the next referents and conditions. The referent *e* represents an event of studying (/يدرسُ/,"study"). The referent *n* is used to denote the time of speech (see figure. 1).

In the final stage of representation the resulted Discourse Representation Structures are interpreted model theoretically in logical forms.

It is obvious that DRT-based semantic construction proceeds from another point of view than the *montague-style* in the construction process and it is therefore not compositional. Furthermore, the semantic construction is given in top-down manner and is not declarative, that means the processing order effects the binding possibilities.

## 4 λ-DRT as a Compositional Semantics for ARABIC.

The combination of *lambda* conversion process in DRT extends DRS's to be compositional without losing the important feature of representing text anaphoric. λ-DRT as a compositional DRS-based Representation formalism has been used in several NLP systems (Fischer, 1993), (Bos et al., 1994), (Konard et al.,1996) (Haddad, 2002).

Based on (Bos et al. 1994) the semantic function of sentences consists in constructing of Discourse Representation Structures by applying some DRS construction rules within the context of the referents. The $DRS_n$, consists for instance of a pair "$<DR_n, COND_n>$", where $DR_n$ represents a universe of discourse, i.e. a set of Discourse Referents and $COND_n$ represents a set of conditions about the $DR_n$.

As an additional feature of the language of λ-DRT, we adopted the merge operation ⊗, which combines *two DRS's* by taking the union of the sets of discourses and conditions separately.

$$\langle DR_1, COND_1 \rangle \otimes \langle DR_2, COND_2 \rangle \equiv \langle DR_1 \cup DR_2, COND_1 \cup COND_2 \rangle \quad (7)$$

For example the meaning representation for constructing the *DRS* for the sentence (/كلُ طالب يجتهدُ/ ,"each student studies-hard") could be represented in terms of λ-DRT as follows (see also (1)):

||كل|| ⇒ λRλS <{x},{x: Any }> ⊗ R(x) $\xrightarrow{\text{كل}}$ S(x) (8)

||each|| ⇒ λRλS <{x},{x: Any }> ⊗ R(x) $\xrightarrow{\text{each}}$ S(x)

||طالب|| ⇒ λ y <{},{y: Individual, طالب (y)}> (9)

||student|| ⇒ λ y <{},{y: Individual, student(y)}>

||يجتهد || ⇒ λ z <{},{e: Event, z: Individual, يجتهد (e, $z_{<agent>}$)} (10)

||studies-hard|| ⇒ λ z <{},{e: Event, z: Individual, studies-hard(e, $z_{<agent>}$)}

The DRS in (10) means, that there is an event of hard-studying (/يجتهدُ /,"study-hard") which takes an individual as an argument and plays the role of an agent.

Simulating the basic aspects of the λ-conversion process presented in (Haddad and Yaseen, 2001) and applying it to the DRS's established above would lead in a simplified form to the following semantic representation:

*<{x},{x: Individual, طالب ($x_{<agent>}$)}>* $\xrightarrow{\text{كل}}$ *<{}, {e: Event, يجتهد (e, $x_{<agent>}$)}* (11)

*<{x},{x: Individual, student($x_{<agent>}$)}>* $\xrightarrow{\text{each}}$ *<{}, {e: Event,study-hard (e, $x_{<agent>}$)}*

Additionally we have incorporated some basic rules to resolve temporal anaphora which have been neglected in the original approach.

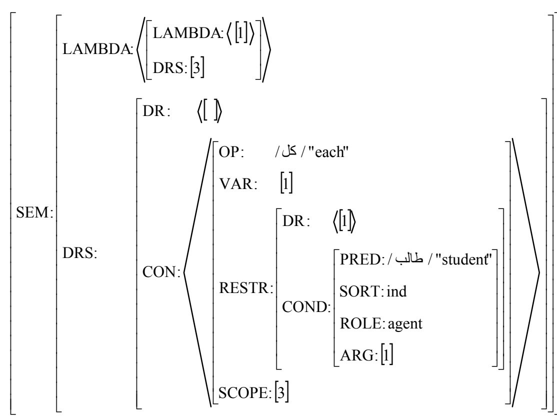
### 4.1 Semantic construction within a Unification-based Formalism for Arabic

A λ-Expression representing the meaning of an Arabic constituent could be formulated in terms of *feature structures*. Such structures should be integrated within a unification-based representation such as HDPSG. Syntactical feature structures involved in such a representation have been in this paper ignored for simplicity reasons.

Semantic feature structures might be represented by a LAMBDA and a DRS feature structure. A LAMBDA feature structure specifies a list of the appropriate arguments, which are involved in the expression, while a DRS feature structure represents the body of the λ-expression. Furthermore, additional pragmatic notations could be also embedded in the DRS feature structures. Compositional rules expressing the meaning of syntactical constituents are also integrated in the lexical entries of a DRS.

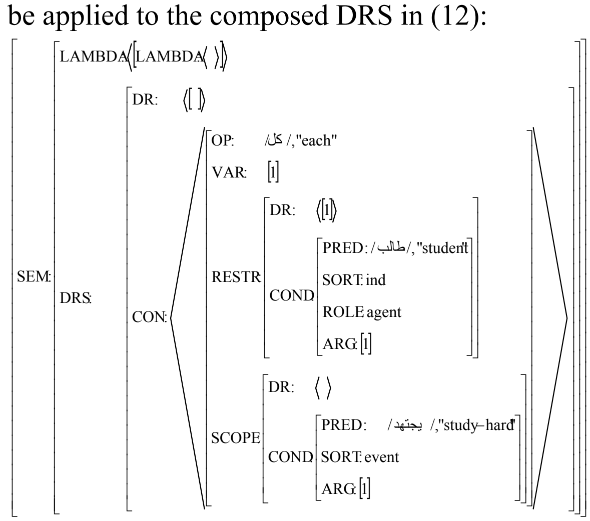
A unification-based semantic construction can be achieved by unifying the values of a LAMBDA feature structure with the representations of the feature structures involved in the arguments. And then storing the results of the unification in the DRS feature structure of the processed syntactical constituent. This process corresponds to λ-conversion proposed in (Bos et al. 1994).

Constructing the meaning of (/كلُ طالب/, "each student") in the sentence (/كلُ طالب يجتهد/, "each student studies-hard") requires the application of the feature structures involved in (8) to the feature structures in (9) (see also "figure 1" and (3), (4), (5), (6)):




(12)

To construct the meaning of the whole sentence (/كلُ طالب يجتهد/, "each student"), "DRS: [3]" has to be applied to the composed DRS in (12):




(13)

It is obvious that (13) corresponds to the logical form in (11).

## 5 Conclusion and Prospects

In this paper we attempted to present some results of our view of a compositional model for semantic construction of Arabic. We believe that the progress, that has been made in the last years, is also *applicable to Arabic with some modifications*. This model is based on the integration of λ-DRT in a unification-based grammatical framework such as HDPSG. This model has been successfully used in several NLP systems to achieve deep syntax-semantic Analysis. Unfortunately there are still little works reported from the Arabic computational linguistic community for semantic construction and HDPSG deep analysis for Arabic. Concentration on Arabic processing has focused on Arabic from the morphological and syntactical point of view. We hope that this approach might be a further motivation to redirect research to modern semantic construction technologies for developing an adequate model of semantic processing for Arabic and even no existing formal theory is capable to provide a complete and consistent account of all phenomena involved in Arabic semantic processing.

## 6 Bibliographical References

Al-Fedaghi and Al-Anzi. 1989 *A New Algorithm to Generate Arabic Root-Pattern Forms.* Proceedings of the 11th National Computer Conference, Saudi Arabia, 1989, pp. 391-400.

Badr AL-Johar and Jim McGregor. 1997. *A Logical Meaning Representation for Arabic (LMRA)*. Proceedings of the 15th National Computer Conference, Riyadh, Saudi Arabia, 1997, pp. 31-40.

Husni Al-Muhtaseb and Chris Mellish. 1997. *Towards an Arabic Upper Model: A proposal*. Proceeding of the 15th National Conference, Riyadh, Saudi Arabia 1997.

Beesley. 2001. *Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans 2001.* ACL/EACL01: Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects.

Bende-Farkas and Kamp. 2001. *Indefinites and Binding: From Specificity to Incorporation*, 13th European Summer School In Logic, Language and Information, ESSLLI, 2001.

Johan Bos, E. Mastenbroek, S. McGlashan, S. Millies and M. Pinkal. 1994. *A Compositional DRS-based Formalism for NLP Applications.* Report 59, VerbMobil, Universität des Saarlandes.

Everhard Ditters. 2001. *A Formal Grammar for the Description of Sentences Structures in Modern Standard Arabic.* ACL/EACL01: Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects.

El-Dessouki, Nazif, Ahmad. 1988 1 *An Expert System for Understanding Arabic Sentences.* Proceeding of the 10th National Computer Conference, Jeddah, Saudi Arabia, 1988, pp 745-759.

Haddad and Yaseen. 2001. *Towards Understanding Arabic: A Logical Approach for Semantic Representation*. ACL/EACL01: Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects.

Haddad. 2002. *AGQL: An Arabic generalized Quantifiers Language.* Internal Artificial Intelligence Projects at F.I.T, Amman University.

Hans Kamp. 1981. *A Theory of Truth and Semantics Representation*. In J. Groendijek, T. J. Stokhof, eds., Formal Methods in Study of Languages. Mathematish Centrum, Amsterdam.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic. Kluwer Academic Publishers.*

Karsten Konard, H. Maier, M. Pinkal. 1996*. CLEARS Ein Werkzeug für Ausbildung und Forschung in Computerlinguistik.* In Natural Language Processing and Speech Technology, 3rd KONVENS Conference, Bielefeld. Mouton de Gruyter

Ingrid Fischer. *1993 Die Kompositionale Bildung von Diskursrepräsentationsstrukturen über einer Chart*. Masters thesis, University of Erlangen, Germany

Richard Montague. 1973. *The Proper Treatment of Quantification in Ordinary English.* In: Philosophy, Language , and Artificial Intelligence, ed., J. Kulas, J. H. Fetzer and T. Rankin, Kluwer Academic Publishers, 1988.

Riadh Ouersighni. 2001. *A major offshoot of the Dinar-MBC project: AraParse, a morphosyntactic analyzer for unvowelled Arabic texts.* ACL/EACL01: Conference of the European Chapter, Workshop: Arabic Language Processing: Status and Prospects.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago.

John Nerbonne. 1993. *A feature-based Syntax/ semantics Interface.* In Annals of Mathematics and Artificial Intelligence 8(1993), J.C. Baltzer AG, Science Publishers.

Has Uszkoreit, Flickinger and Ivan Sag 2000. *Deep Linguistic Analysis with HDPSG.* In Verbmobile: Foundation of Speech-to-Speech Translation, Ed. W. Wahlster, Springer.