Conference Papers                                                    Computational Functional Linguistics

2008-10-23

# Implementing Arabic-to-English Machine Translation Using the Role and Reference Grammar Linguistic Model

Yasser Salem
*Technological University Dublin*, yasser.salem@ucd.ie

Arnold Hensman
*Technological University Dublin*, arnold.hensman@tudublin.ie

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

Follow this and additional works at: https://arrow.tudublin.ie/itbinfocon

Part of the Other Computer Engineering Commons

## Recommended Citation

# Implementing Arabic-to-English Machine Translation using the Role and Reference Grammar Linguistic Model

Yasser Salem, Arnold Hensman and Brian Nolan

School of Informatics and Engineering
Institute of Technology Blanchardstown, Dublin, Ireland
E-mails: {firstname.surname}@itb.ie

## Abstract

This paper presents work-in-progress investigating the development of a rule-based lexical framework for Arabic language processing using the Role and Reference Grammar (RRG) linguistic model. A system, called UniArab is introduced in this research to support the framework. The paper outlines the conceptual structure of UniArab System, which utilizes the framework and translates the Arabic language into another natural language. Also, this paper explores how the characteristics of the Arabic language will effect the development of a Machine Translation (MT) tool from Arabic to English. Several distinguishing features of Arabic pertinent to MT will be explored in detail with reference to some potential difficulties that they might present.

Keywords: UniArab, Machine Translation, Role and Reference Grammar, Arabic

## 1   Introduction

Arabic is a Semitic language originating in the area presently known as the Arabian Peninsula. It has been spoken in its current form since the 2nd millennium BCE. The Arabic language is one of six major world languages, and one of the six official languages of the United Nations. The motivation for this research is to provide a proof of concept software application; to develop an automated translator sufficient in translating from Arabic into English. This research will investigate a rule-based lexical framework for processing the Arabic language using the Role and Reference Grammar linguistic model [VanValin and LaPolla, 1997]. As a language, Arabic has few irregularities and it is rich in morphological structure. Arabic is also rare in that it is a derivational language rather than concatenative. Words like 'went , go' - ذهب ، يذهب _dhb , y_dhb [1] can easily be seen as being part of a hierarchy of inheritance from a specific root (in this case ذهب _dhb). In English and in many other languages this is not always the case. The Arabic language is written from right to left. It has 28 letters, many language specific grammar rules and it is a free word order language. Each Arabic letter represents a specific sound so the spelling of words can easily be done phonetically. There is no use of silent letters as in English. Similarly, there is no need to combine letters in Arabic to achieve a certain sound that might be familiar to an English speaker. For example, the 'th' sound in English as in the word '**Th**inking' is reduced in Arabic to the character ث _t [Salem et al., 2008].

In addition to the standard challenges involved in developing an efficient translation tool from Arabic to English, the free word order nature of Arabic creates an obstacle unique to the language. The number of possible clause combinations in basic phrasal structures far exceeds that of most languages. There is no copula verb 'to be' in Arabic, resulting in a unique usage of the subject 'I'. The absence of the indefinite article, while not unique to Arabic still poses many difficulties within the context of the language structure. These and other issues are discussed in later sections. These distinguishing features pose a major challenge in processing Arabic text. The framework is to be evaluated using a machine translation system that translates an Arabic sentence as source language into an English sentence as target language. This

---

[1]Arabic examples are written here by using Buckwalter Arabic Transliteration which is converted in latex into the DIN 31635 standard of Arabic transliteration

paper presents an work-in-progress. The paper discusses a system based on RRG, called the UniArab system. The UniArab system translates Arabic sentences into a logical structure based on RRG. Based on this logical structure, the equivalent sentences in another natural language is generated. The paper outlines the conceptual structure of the UniArab system, which utilizes the framework. The remaining sections are organized as follows: Section 2 Machine translation. Section 3 The Role and Reference Grammar (RRG) linguistic model. Section 4 presents the UniArab system. Section 5 outlines the conceptual structure of the UniArab System. Section 6 outlines a series of tests to evaluate the UniArab system. Section 7 summarizes the paper and highlights the future work.

# 2 Machine Translation

Machine translation of natural languages, commonly known as MT 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. While semi-automated tools have been applauded in the recent past as the most realistic path to follow, it is no longer the case. The current consensus is that fully automated, efficient translation tools should remain the primary goal. The nature of users of such systems and the type of text involved leave little room for continued dependence on human aids. The motivation for an Arabic-English translation tool is obvious when one considers that Arabic is the lingua franca of the Middle-Eastern world. Presently, 21 countries with a combined population of 450 million consider Standard Arabic as their national language. A simple test case during a study at Abu Dhabi University over three popular Arabic translation tools (Google, Sakhr's Tarjim and Systran) revealed little success in generating the correct meaning [Izwaini, 2006].

## 2.1 General MT Obstacles

For the purposes of this study, any proposed solution to and Arabic-English translator will be based upon the interlingua model of machine translation. A transfer model that directly maps from source language to target language will remove the benefit of similarities between an Arabic translator and others. Arabic is unique in many ways but is not immune to the standard challenges faced in prior developments of MT tools for other languages such as multiple meanings of words, non-verbalisation and insufficient lexicons.
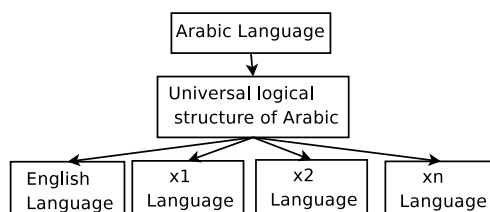


Figure 1: Interlingua model of Arabic MT

An Interlingua model that incorporates source language analysis, thereby creating a so called universal logical structure (in this case Arabic), will facilitate multiple language generation in a more flexible way. An Interlingua model is presented in Figure 1. For the elements of *subject*(S), *verb*(V) and *object*(O), Arabic's free word order allows the combinations of SVO,VSO, VOS, and OVS. The only combinations that do not occur in Arabic are OSV and SOV. Free word order is discussed later in this paper. Our research attempts to develop a rule-based and lexicon framework for the processing of Arabic using the Role and Reference Grammar (RRG) linguistic model. The framework is to be evaluated using a machine translation system that translates an Arabic text as source language into an English text as target language.

## 2.2 Challenges of Arabic to English Machine Translation

Arabic has a 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. Also, in Arabic there are some words that

hold the meaning of a full sentence for example, سنسافر *snsāfr* , in English would translate to; we will travel. This means any MT system should have a strong analysis to obtain the root or to realise in one word that there is fact a full sentence in the English equivalent. Arabic has free word order, this poses a significant challenge to MT due to the vast possibilities to express the same sentence in Arabic.

For example, consider the following word order. *(1) Verb Noun Noun    (2) Noun Verb Noun*

This means that we have a challenge to identify exactly which is the subject and the object. Table 1(a) and Table 1(b) shown this challenge. Please note that the sentences in Table 1(a) and Table 1(b) should be read from right to left.

Table 1:

(a) Verb Noun Noun example.

| ليلى قيس يحب | | *yḥb qys lylā* |
|---|---|---|
| Qays loves Laila | | |
| ليلى *lylā* | قيس *qys* | يحب *yḥb* |
| Laila | **Qays** | loves |
| noun | noun | verb |

(b) Verb Noun Noun example.

| قيس ليلى يحب | | *yḥb lylā qys* |
|---|---|---|
| Qays loves Laila | | |
| قيس *qys* | ليلى *lylā* | يحب *yḥb* |
| **Qays** | Laila | loves |
| noun | noun | verb |

The difference in Table 1(a) and Table 1(b) is the position of the actor. In Table 1(a), the actor is first argument of the verb. In Table 1(b), the actor is second argument of the verb. Both sentences in fact have the same meaning.

## 2.3   Other Approaches to MT

The current trend in Arabic MT is to use Knowledge-based and Empirical Methods [Soudi et al., 2007]. These systems are generally poor in terms of output results, part of the reason is that; they attempt to extract structure from the words themselves, which can be a complex and ambiguous. In the RRG we analyse the sentence as whole depending on logical structure of the verb in the first and remaining words. This will cover in section 3.

# 3   The Role and Reference Grammar (RRG) Linguistic Model

The Role and Reference Grammar (RRG) is a model which presupposes a direct mapping between the semantic representation of a sentence and its syntactic representation; there are no intermediate levels of representation [Van Valin, 2007]. The general view of RRG is presented in Figure 2.

Syntactic Representation

↑

Linking
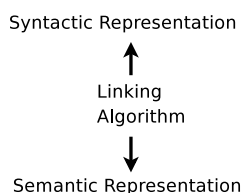Algorithm

↓

Semantic Representation

Figure 2: Layout of Role and Reference Grammar

The RRG creates a 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. Before developing the linking algorithms that govern these mappings, it is necessary to first introduce a general principle constraining these algorithms [VanValin and LaPolla, 1997]. 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, the linking rules must refer to the morphosyntactic features of the sentence. One question however remains; why should a grammar deal with linking from syntax to semantics at all. Simply specifying the possible realizations of a particular semantic representation should suffice. They refute this using the argument that theories of linguistic

structure should be directly relatable to testable theories of language production and comprehension, [Van Valin and LaPolla pp339-340]. One of our hypotheses it that RRG is very suitable for machine translation of Arabic via an interlingua bridge. It is a mono strata-theory, positing only one level of syntactic representation, the actual form of the sentence. Linking algorithm can work in the both directions from syntactic representation to semantic representation or vice versa. UniArab will fulfil this role. 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. We briefly illustrate the active voice linking in (a) and (b) where (a) is an SVO clause and (b) is the VSO equivalent. Arabic allows variation in clause word order. The active-voice linkings, those in the sentence in (a)-(b), are illustrated in Figure 3.

a. عمر رأي زيد *zyd ray mr*    Zaid saw Omar زيد *zyd* MsgNOM see.past عمر *mr* - MsgNOM

b. عمر زيد رأى *raā zyd mr*    Saw Zaid Omar see.past زيد *zyd* MsgNOM عمر *mr* MsgNOM
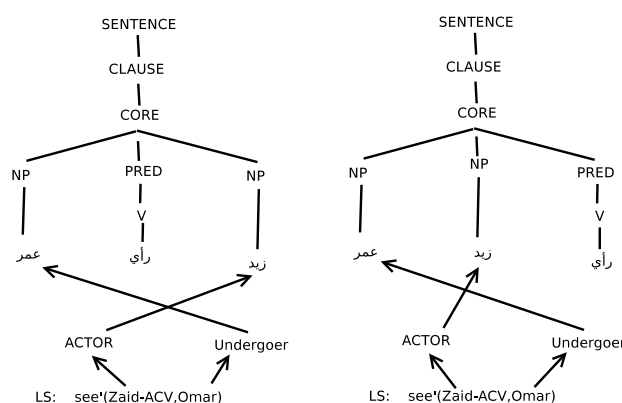


Figure 3: Arabic sentence types; verb subject object or subject verb object

The first (leftmost) argument of '*see*' in the logical structure is the actor, the second the undergoer, following the RRG Actor-Undergoer hierarchy. Since Arabic is an accusative language and رأى *raā* '*see*' is a regular verb, the actor will receive nominative case and the undergoer accusative case. On the other hand, in Arabic we can start a sentence with a verb as shown in (1b). The only changes in the clause are the form of the verb and the form of the actor NP; the arrangement of the arguments has not changed in the logical structure.

## 4  The UniArab System

This section presents an Arabic to English machine translator system, called UniArab. UniArab is an acronym for **Uni**versal machine translator system for **Arab**ic language. High quality machine translation systems can be developed only if a naturalistic way to treat meaning in natural languages is found. We have attempted to complement this meaning connection to syntax via the RRG linking system (syntax to semantics and vice versa) as indicated in Figure 2.
We are aiming to build a system which can translate a wide variety of simple sentence types. We aim to make this system as scalable as possible by allowing user addition to the lexicon and later, to include complex sentences.

### 4.1  The Conceptual Structure of UniArab System

The conceptual structure of the UniArab system is shown digram in Figure 4. 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. 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
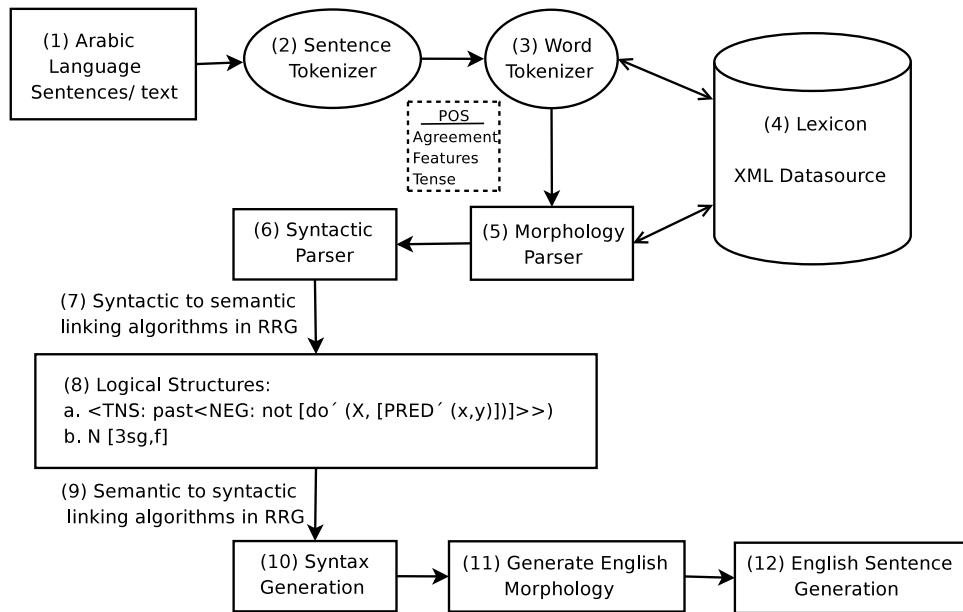
Figure 4: The conceptual architecture of the UniArab system

affairs' of the sentence. In the UniArab system we intend to have a strong analysis system that can unpack all information from the attributes.

## 4.2 The Technical Architecture of the UniArab System

The structure of the UniArab system in Figure 4 breaks down into the following phases:

**Phase (1) - 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. For example. قرأ خالد الكتاب. خالد تلميذ ذكي. *qra ḫāld ālktāb. ḫāld tlmyḏ ḏky.* will be two tokens; قرأ خالد الكتاب . *qra ḫāld ālktāb .* and خالد تلميذ ذكي. *ḫāld tlmyḏ ḏky.* the translation of these two sentences is *Khalid read the book. Khalid is a clever student*

**Phase (3) Word Tokenizer** There, sentences are split into tokens. قرأ خالد الكتاب *qra ḫāld ālktāb Khalid read the book*, the output of phase 3 is as follows;

```
<sentence>
<word>قرأ qra</word>
<word>خالد ḫāld</word>
<word>الكتاب ālktāb</word>
</sentence>
```

**Phase (4) Lexicon Datasource** A set of XML documents for each component category of Arabic.

**Phase (5) Morphology Parser** Directly works with both the Lexicon and Tokenizer to produce the word order. A connection is made to the datasource 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 system is in a position to apply a logical structure towards generating the English translation.

**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, we must begin with this approach.

**Phase (8) Logical Structure** Creation of logical structure (which at this stage of our research has been fully completed) is the most crucial phase. An accurate representation of the logical structure of an Arabic sentence is the primary strength of UniArab. Below is a sample output from the UniArab system. The Arabic equivalent of the past tense sentence 'Khalid read the book' قرأ خالد الكتاب *qra ḥāld ālktāb* is input as the source.

الكتاب *ālktāb* book:N        خالد *ḥāld* Khalid:MsgN        قرأ *qra* read:V

read [do'(x,[read'(x,(y)])] sg 3rd M PAST قرأ *qra*

The results of the parse can be seen here with LS as :

Verb read [do'(x,[read'(x,(y)])] sg 3rd M PAST قرأ *qra*>

where the Proper Noun is Khalid sg unspec M خالد *ḥāld*

and the Noun is, the book sg def M الكتاب *ālktāb*

Consider the following example; Omar is a student. be'(Omar,[student']). in Arabic عمر تلميذ *ʿmr tlmyḏ*. This is a challenge since there is no verb 'to be' in Arabic, but this must be inferred for correct translation. Instead of saying *'Omar is a student'*, the Arabic equivalent would be *'Omar student'*. We also face the challenge of inferring the indefinite article, which 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.

*Phase (10) - Syntax Generation  Phase (11) - Generate English,Phase (12) - English Sentence Generation*. The development of the final phases is currently ongoing.

# 5 UniArab: An Arabic Language Processing System Based On RRG

The UniArab system is a natural language processing application based on Role and Reference Grammar (RRG) for translating the Arabic language into any other language, using an RRG based interlingua bridge. An interlingua based MT approach to translation is done via an intermediate-semantic representation of the source language [Hutchins, 2003]. The conceptual architecture of the UniArab system in Figure 4. To apply it to any other language, we need only change the phases 9, 10, 11 and 12.

## 5.1 The RRG-based Functions of the UniArab System

The UniArab system can generate a target language by classifying every Arabic word in the source text. There are six major parts of speech in Arabic. These are Verbs, Nouns, Adjectives, Proper nouns, Demonstratives, Adverbs and we create a seventh, so called 'other' category for Arabic words which do not fit into any of these six categories. The major part of speech in the Arabic language have their own attributes, and we use these attributes within the UniArab system. For example, the verbs in the Arabic language agree with the subject in gender. In Arabic, there is no neutral gender. In the UniArab system we record the gender associated with a verb in syntax for a particular subject NP. Adjectives and demonstratives also agree with subject in gender too. Arabic words are of two types with regards to gender: masculine and feminine. In Arabic, words come into three categories with regards to number:

They are (1) singular, indicating one, e.g. 'one man'. (2) dual, indicating two, e.g. 'two men' and (3) plural, indicating three or more. The UniArab system records these attributes of gender or number. It is important to understand that source language specific features may not be used or may be different in the target language. For example, Arabic number category of dual or plural. The UniArab system is based on RRG and uses logical structures for each verb in the lexicon.

## 5.2 The RRG-Based Logical Structure of the UniArab System

RRG employs the semantic logical structure from the lexicon to explain the structures of the sentence. In the uniArab system, we use the RRG model of the lexicon to facilitate a rich and accurate representation of the state-of-affairs of the underlying input source. Once the Arabic source is captured UniArab further captures the meta-data about the grammatical consequences. For example, tense information and agreement constraints between subject and verb. The hypothesis is that the use of the RRG linguistic model to motivate the software design will remove problems for translation, and target language generation caused by an incorrect linguistic analysis and description of the source Arabic language. The UniArab system therefore seeks by design to avoid the inaccurate analysis of Arabic source language [Izwaini, 2006]. For example, the UniArab system is used to extract the expressions in (c) and (d) by analysing all attributes of the words. The UniArab system has the ability to understand the word agreement features and tense.

c.    Mary did not read the book.    $<$TNS: past $<$NEG: not [do'(Mary,[PRED'(Mary,book)])]$>$ $>$
d.    Mary    N [3sg, f]

## 5.3 Technical Challenges

**Arabic letters in the GUI** We cannot write Arabic letters in UniArab's GUI. We use Unicode to represent them. *Our Unicode Converter System* allows us to enter Arabic text and click on a button to generate the equivalent Unicode.

**Arabic letters in Eclipse IDE for Java** We used the Eclipse IDE for Java development. We cannot write Arabic as a string in Eclipse. While Java does support Arabic, the problem lies in the operating system not supporting Arabic letter shapes within the IDE. We used Windows XP and Windows 2000 which both have the same problem. To fix this we changed to Ubuntu Linux. Under Linux we can write Arabic text as string in the Eclipse IDE.

**Arabic in datasource** We chose to create our datasource as XML files, for optimum support of different platforms. It was also easier as we used Arabic letters rather than Unicode inside the datasource. XML fully supported Arabic. We created our search engine using Java. We used HashMap to make the keyword in Arabic when we searched the datasource. We used $verbMap.containsKey(word)$ in order to check the presence of an Arabic word in the datasource.

# 6   Evaluation of MT

The evaluation of MT systems is a difficult task. This is not only because many different metrics are involved, but also because translation itself is difficult [Laoudi et al., 2004]. The most important criteria for a potential test is to determine the translational capability. Therefore, we need to draw up a complete overview of the translation process, in all its different aspects. To evaluate the quality of any translation is difficult, since it is not entirely clear what the focus of the evaluation should be. A good translation has to effectively capture the meaning. Current general function MT systems cannot translate all texts consistently. Output can have very poor quality. It should also be mentioned that the subsequent editing requirement increases with translation quality is poor [Turian et al., 2003].
The quality of a machine translator can be evaluated using a number of methods. These evaluation methods are based on comparison. IBM's BLEU was one of the first metrics to report correlation with human judgements of quality. The metric is currently one of the most popular in the field [Papineni et al., 2002]. The NIST metric is based on the BLEU metric [Doddington and George, 2002]. The METEOR metric addresses some of the weakness inherent in the BLEU metric. The metric is based on the weighted harmonic mean of *unigram precision* and *unigram recall* [Banerjee and Lavie, 2005]. We will create variants of Arabic sentences that represent all possible structures of sentences that UniArab can translate.

We will evaluate the result of output by comparing between human-translated and alternative machine-translated versions from Google and Microsoft.

# 7  Summary and future work

This paper has introduced an MT system called UniArab, which is based on the RRG model. It creates a logical structure from an Arabic sentence based on RRG. The UniArab system provides information captured for the main components of the Arabic sentence; verb, noun, pronoun, adjective, demonstrative and adverb. The conceptual structure of the UniArab system has been outlined. The RRG model is used as it presupposes a bi-directional mapping between the semantic representation of a sentence and its syntactic representation; there are no intermediate levels. Arabic is a free word order language and we can use the RRG logical structures technique to capture these clauses. UniArab system datasource has been implemented as a set of XML documents. The use of XML has the added advantage of portabilty. UniArab will effectively work the same regardless of the platform. Each of the seven categories selected from Arabic are held in one XML document. When the system begins its analysis of any word, it simply searches these documents. The attributes of every word are fed back to the system. Some words could be in more than one category, for example the word حر *ḥr* in Arabic can mean both 'heat' (noun) or 'free' (adjective). An Arabic to English MT system that uses the RRG model by creating a logical structure as its interlingua bridge offers a unique contribution to the field of MT.

Our main future work is to 1) Implement the target language generation phases of the UniArab system from the logical structure that is presently complete. 2) Evaluate the UniArab system in translating from Arabic to English. Based on this logical structure and further elements of the UniArab system, the equivalent sentence of the target natural language will be generated. English will be the initial target language.

# References

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan*.

[Doddington and George, 2002] Doddington and George (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *In Proceedings of the Second Conference on Human Language Technology (HLT-2002). San Diego, CA.*, pages 128–132.

[Hutchins, 2003] Hutchins, J. (2003). Machine translation: General overview. In *Oxford University Press, Oxford*.

[Izwaini, 2006] Izwaini, S. (23 Oct. 2006). Problems of arabic machine translation: evaluation of three systems. pages 118–148.

[Laoudi et al., 2004] Laoudi, J., Tate, C., and Voss, C. (2004). Towards an automated evaluation of an embedded mt system. In *roceedings of the European Association for Machine Translation Workshop, Malta*.

[Papineni et al., 2002] Papineni, Kishore, Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: amethod for automatic evaluation of machine translation. In *In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

[Salem et al., 2008] Salem, Y., Hensman, A., and Nolan, B. (May 2008). Towards arabic to english machine translation. In *ITB Journal Issue Number 17*.

[Soudi et al., 2007] Soudi, A., van den Bosch, A., and Neumann, G. (2007). *Knowledge-based and Empirical Methods*. Springer.

[Turian et al., 2003] 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, 2007] Van Valin, R. (2007). The role and reference grammar analysis of three place predicates. In *Contemporary Linguistics*, volume 63. Hrvatsko filoloko drutvo.

[VanValin and LaPolla, 1997] VanValin and LaPolla, R. (1997). *Syntax: Structure, Meaning, and Function*. Cambridge University Press.