Mapping Interlingua Representations to Feature Structures of Arabic Sentences

Khaled Shaalan[12] Azza Abdel Monem[3], Ahmed Rafea[4], Hoda Baraka[5]
[1]The Institute of Informatics, The British University in Dubai, P O Box 502216, Dubai, UAE
[2]Honorary Fellow, School of Informatics, University of Edinburgh, UK
[3]Central Lab. For Agricultural Expert Systems (CLAES), Argiculture Research Center, P O Box: 100 Dokki, Giza, Egypt
[4]Faculty of Computers and Information, Cairo University
5 Ahmed Zewel St., Orman, Giza, Egypt
[5]Faculty of Engineering, Cairo University, Dokki, Giza, Egypt
khaled.shaalan@buid.ac.ae, azza@mail.claes.sci.eg, rafea@mail.claes.sci.eg, hbaraka@mcit.gov.eg

The interlingua approach to Machine Translation (MT) aims to achieve the translation task in two independent steps. First, the meanings of source language sentences are represented in an intermediate (interlingua) representation. Then, sentences of the target language are generated from those meaning representations. In the generation of the target sentence, determining sentence structures becomes more difficult, especially when the interlingua does not contain any syntactic information. Hence, the sentence structures cannot be transferred exactly from the interlingua representations. In this paper, we present a mapping approach for task-oriented interlingua-based spoken dialogue that transforms an interlingua representation, so-called Interchange Format (IF), into a feature structure (FS) that reflects the syntactic structure of the target Arabic sentence. This approach addresses the handling of the problem of Arabic syntactic structure determination in the interlingua approach. A mapper is developed primarily within the framework of the NESPOLE! (NEgotiating through SPOken Language in E-commerce) multilingual speech-to-speech MT project. The IF-to-Arabic FS mapper is implemented in SICStus Prolog. Examples of Arabic syntactic mapping, using the output from the English analyzer provided by Carnegie Mellon University (CMU), will illustrate how the system works.

Keywords: Machine Translation, Interlingua Approach, Natural Language Generation, Syntactic Structure Representation, Arabic Natural Language Processing.

1. INTRODUCTION

Arabic is the fourth most widely spoken language in the world (Nwesri et al., 2005). It is a morphologically and syntactically rich language. Arabic morphological and syntactic analyses have gained the focus of Arabic natural language processing research for a long time in order to achieve the automated understanding of Arabic (Sughaiyer et al., 2004). On the other hand, Arabic generation has received little attention although the generation problems are as complex as those of the analysis (Habash, 2004; Shaalan et al., 2006).

With the recent technological advances in multilingual machine translations, Arabic natural language generation has received attentions in order to automate Arabic

translations. For machine translation systems that support a large number of languages, interlingua approach is particularly attractive (Levin et al., 2003): 1) it requires fewer components in order to relate each source language to each target language, 2) it takes fewer components to add a new language, 3) It supports paraphrase of the input in the original language, and 4) both the analyzers and generators can be written by mono-lingual system developers.

Researches on translations of Arabic using the interlingua approach are beginning to emerge. Cavalli-Sforza et al. (2000) and Soudi et al. (2002) developed a template-based Arabic realizer which is based on KANT (Mitamura et al., 1992). The Arabic generator is implemented using MORPHE (Leavitt, 1994) and Genkit (Tomita et al., 1988) tools that compile the morphological and grammatical rules into morphological and sentence generator programs, respectively. The problems with these tools are that they are not easily adaptable to Arabic script and syntax. That is why the generator has dealt with restricted forms of Arabic verbs and nouns. Abul seoud (2005) developed a prototype for transferring an Arabic parse tree, which is obtained from a DCG parser, into KANT-like Interlingua (Mitamura et al., 1992). A set of structural transformation and mapping rules has been described.

In the generation of the target sentence, determining sentence structures becomes more difficult, especially when the interlingua does not contain any syntactic information. Hence, the sentence structures cannot be transferred exactly from the interlingua representations. In this paper, we present a mapping approach for task-oriented interlingua-based spoken dialogue that transforms an interlingua representation, so-called Interchange Format (IF), into a FS that reflects the syntactic structure of the target Arabic sentence. This approach addresses the handling of the problem of Arabic syntactic structure determination in the interlingua approach. A mapper is developed primarily within the framework of the NESPOLE! (NEgotiating through SPOken Language in E-commerce) multilingual speech-to-speech MT project (for the interlingua specification of the NESPOLE project, see http://www.is.cs.cmu.edu/nespole/db/specification.html).

The IF-to-Arabic FS mapper is implemented in SICStus Prolog. Examples of Arabic syntactic mapping, using the output from the English analyzer provided by Carnegie Mellon University (CMU), will illustrate how the system works (see http://www.is.cs.cmu.edu/nespole).

The rest of the paper is organized as follows: description of the interlingua representation is briefly summarized in section 2. This is followed by introducing the interlingua-to-Arabic FS mapper in Section 3. In Section 4, we present the computational model of the Arabic mapper. Section 5 presents examples of FSs of Arabic sentences that reflect the syntactic structure of the target Arabic sentences. Next, in Section 6, we describe the set of important issues that we encountered during the design and implementation of the mapper. Finally, we give concluding remarks in section 7.

2. THE DESCRIPTION OF INTERLINGUA

The NESPOLE! translation system (Metze et al., 2002) is designed to provide human-to-human speech-to-speech machine translation using an interlingua-based approach similar to that used in the JANUS system (Levin et al., 2000). The general goal of the system is to provide translation over the Internet to facilitate communication for E-commerce and e-service applications between common users in real-world settings.

The domain addressed in NESPOLE! is Travel & Tourism, a task-oriented domain. The NESPOLE machine translation project uses an interlingua representation

which is based on speaker intention rather than literal meaning. The IF is a task-based representation of the semantics of a unit of speech. Since the system translates spoken dialogue, these units are called Spoken Dialogue Units (SDUs), and they range in length from a single word ("hello") up to a full sentence ("I'd like to reserve a room"). An IF is based on a set of domain actions (DA) with parametric arguments. In general, each DA has a speaker tag and at least one speech act optionally followed by a string of concepts and optionally, a string of arguments. DAs can be roughly characterized as follows (Levin et al., 2003):

Speaker: speech act +concept* arguments*

The plus sign separates speech acts from the concepts and concepts from each other. Arguments are represented as an argument name followed by the "=" symbol followed by a value and/or subargument(s). Several examples of utterances, with corresponding IF representations, are shown below:

1. a:Good morning
   a:greeting (greeting=good-morning)
   صباح الخير
2. c: I'm planning a vacation this summer in Egypt
   c:give-information+plan+trip (who=i,visit-
   spec=(identifiability=no, vacation,
   time=(season=(identifiability=non-distant,
   summer)),location=name-egypt))
   أنا أخطط لأجازة في مصر هذا الصيف
3. c:How much does a double room with full board accommodation
   cost?
   c:request-information+price+accommodation (price=question, room-
   spec=(double-room, identifiability=no),include=(accommodation-
   board=full_board))
   كم سعر أقامة كاملة في غرفة مزدوجة؟
4. c:Tell me about sightseeing and transportations
   c:request-action+inform+object (object-spec=(operator=conjunct,
   [(sightseeing, identifiability=,yes), (transportation,
   quantity=,plural, identifiability=yes)]))
   أخبرني عن المعالم السياحية والانتقالات

3. THE INTERLINGUA-TO-ARABIC FEATURE STRUCTURE MAPPER

In interlingua-based machine translation, the second half of the translation process is generation. This section describes a proposed rule-based syntactic structure determination approach for Arabic from the interlingua representation used in NESPOLE! The basic architecture of the proposed Arabic mapping system is shown in Figure 1. The goal and result of mapping is a target-language FS, a list of feature-value pairs, whose contents reflect the content of the interlingua, expressed in terms of syntactic and lexical properties of the target language. The mapper uses domain ontology, mapping rules and a mapping lexicon to convert the IF into FS. The FS represents a target sentence structure. The mapper recursively traverses the interlingua (IF) hierarchy and applies the mapping rules in order to produce a FS. Since the mapper is written in Prolog, it follows a top-down, depth-first strategy for applying rules during mapping. Below, we present the components of the interlingua-to-Arabic FS Mapper in more details.

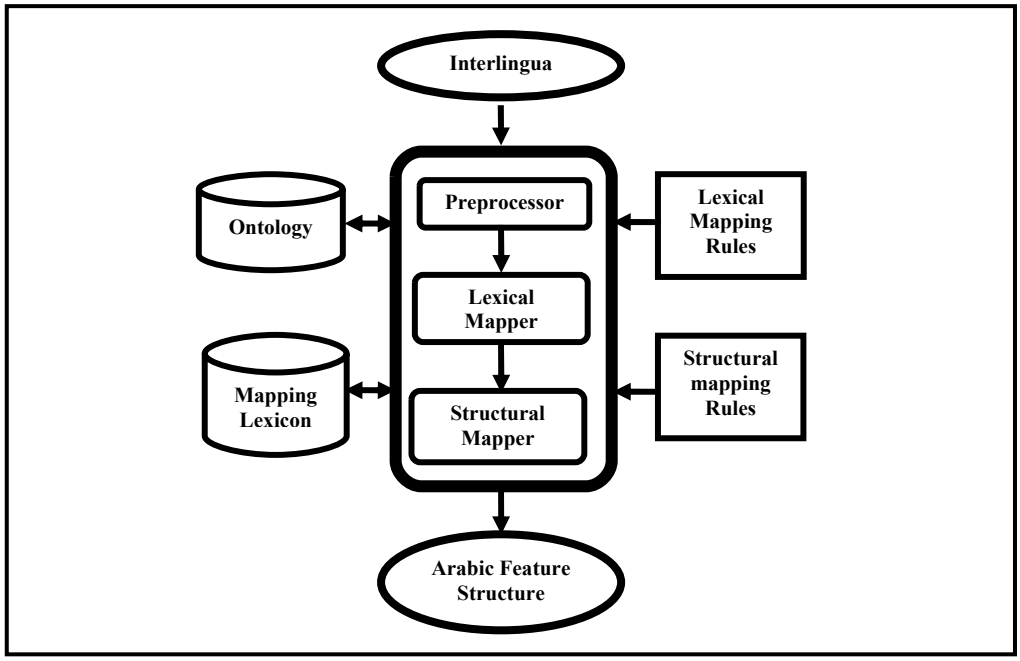

FIGURE 1: Architecture of the Interlingua-to-Arabic Feature Structure Mapper

3.1 Interlingua and Ontology

The domain ontology contains a formal definition of what the legal IF representations are. The definition is based on the IF specification language, which has been agreed upon by the NESPOLE consortium. The domain ontology consists of the definitions of the legal concepts and speech actions, legal arguments, legal values, and their relationships in an abstract way. Domain actions (speech acts and concepts) in the IF are essentially flat classes whose purpose is to categorize each SDU based on the intention of the speaker. Figure 2 contains examples of ontology representations in Prolog.

```
% speech_act/1 defines legal speech acts
speech_act('give-information').

% concept/1 defines legal concepts
concept('+trip').

% da_arg/2 defines legal arguments for each speech
as or concept
da_arg('+trip',[ 'who=','visit-spec=',…]).

% arg/2 defines legal subarguments for each argument
arg('visit-spec=',['identifiability=','time=',
   'location=',…]).


% arg_val/2 defines legal values of an argument
arg_val('who=', [i,you,we,they],name-_,…]).
```

FIGURE 2: Examples of ontology in Prolog

3.2 Mapping Lexicon

The mapping lexicon defines the relation between interlingua, which represents the meaning conveyed in the source text, and the target lexical items. We differentiate between general and specific lexical mapping entries. Both entries relate an element of semantics with an Arabic lexeme and its part of speech (POS). A general lexical mapping entry relates an argument value with an Arabic word. Figure 3 contains examples of general lexical mapping entries in Prolog.

```
% map_vlaue/3: given a word and
% its category, it maps it into Arabic word

% mapping nouns
map_value(i,pronoun,'أنـا').
map_value(vacation,noun,'أجـازة').
map_value(summer,noun,'صيف').
map_value(name-egypt,noun,'مصر').
% mapping verbs
map_value(interest,verb,'أهتم').
% mapping particles
map_value(conjunct,particle,'و').
```

FIGURE 3: Examples of general lexical mapping entries in Prolog

3.3 Mapping Rules

Mapping rules are language specific knowledge about the relationship between the meaning patterns in IF representation and the syntactic structure of the target language. There are two types of mapping rules: lexical mapping rules and structural mapping rules. Lexical mapping rules use the mapping lexicon to transform lexical values in the IF into the corresponding Arabic words. Structural mapping rules are used to determine a syntactic structure of the Arabic sentence.

The structural mapping extracts the basic constituents of the Arabic sentence from the IF representations. These constituents are used to construct the Arabic syntactic structure that will be used to generate the Arabic sentence. Figure 4 shows an example of a structural mapping rule that extracts four constituents of the Arabic sentence: an optional coordination, subject, verb, and complement.

```
% syntactic structure consisting of <Coordination Subject Verb Complement>
get_sent_structure(statement,Speech_act,IF,FS):-
     get_coordination(SentenceType,IF,Coordination,[]),
     get_verb(SentenceType,IF,Verb0,[]),
     get_subject(SentenceType,IF,Subject,[]),
     get_complement(SentenceType,IF,Complement,[]),!,
     (has_negate_prefix(Speech_act) -> % e.g. negate-dialog-hear
            Verb=[negate|Verb0]
     ; Verb = Verb0
     ),
     (Coordination= [] ->
            FS=[subject:Subject,verb:Verb,complement:Complement]
       ;    FS=[coordination:Coordination,subject:Subject,verb:Verb,
                    complement:Complement]
       ).
```

FIGURE 4: An Example of structural mapping rules

4. THE COMPUTATIONAL MODEL OF THE ARABIC MAPPER

The mapping process involves three main stages:

- *Preprocessing.* The preprocessing is mainly based on the domain ontology. It performs three tasks: transforms an IF representation into a Prolog term, associates the arguments to their concepts, and checks the IF for correctness.
- *Lexical Mapping.* Performing lexical look-up for the lexical entries in order to associate lexemes with semantic IF concepts and values.
- *Structure Mapping.* Determining the syntactic structure. It performs two tasks:
    - Use the Speech Act to determine the sentence mode. There are four sentence modes that can be derived from the IF representation:

statement, command (imperative), interrogative (question), and fragment (word or phrase)

- Use the sentence mode and the rest of the IF to:
  - Group the words to construct the constituents. There are five constituents that: coordinates sentences, occurs as subject, occurs as verb, occurs as interrogative, and occurs as complementary of a sentence.
  - Order the constituents to form the syntactic structure of the Arabic sentence.

The structural mapping rules follow the transformation grammar formalism to order the recognized constituents and construct the Arabic FS that reflects the syntactic structure of the target Arabic sentence. They are processed in order and use the pattern shown in Table 1 to construct the Arabic FS. The sequence of the feature-value pairs in the FS corresponds to the syntactic structure that will be used to generate the Arabic sentence.

Sentence Mode	Syntactic structure	Example of Target Arabic sentence
statement	[Coordinate] S V C	أنا أرغب في حجز غرفة هذا البرنامج يشمل إقامة فى غرفة مزدوجة وأفطار
	[Coordinate] V C	يوجد فقط غرف مزدوجة فى الأسبوع الثانى عشر
	[Coordinate] S V1 V2 C	نحن نرغب فى أن نسبح فى الفندق
	[Coordinate] NP C	أجازتي من العاشر من يوليو إلي الثاني عشر من أغسطس
command	[Coordinate] V S C	أحجز لنا أربع غرف من فضلك
question	[Coordinate] Q V S C	هل تقبل شيكات سياحية ؟
	[Coordinate] Q V S	ماذا أكلت ؟
	[Coordinate] Q V C	هل أستطيع تأجيل حجز البرنامج ؟
	[Coordinate] Q C	أين الفندق ؟
fragment	[Coordinate] C	في الفندق

TABLE 1: Syntactic structure of the target Arabic sentence

5. EXAMPLES OF FEATURE STRUCTURES OF ARABIC SENTENCES

In the following, we describe examples of the FSs representing the syntactic structure of Arabic sentence that results from applying the structural mapping rules. Due to the limitation in space and to avoid redundancy, we will only present an example for each mode of sentences, i.e. statement, command, interrogative, and fragment. To explain how the mapper maps the input into an Arabic FS, we provide a complete example in Figure 5 for a sentence of mode statement. For the examples of the other mode of sentences, we will show only the input and output for the mapper.

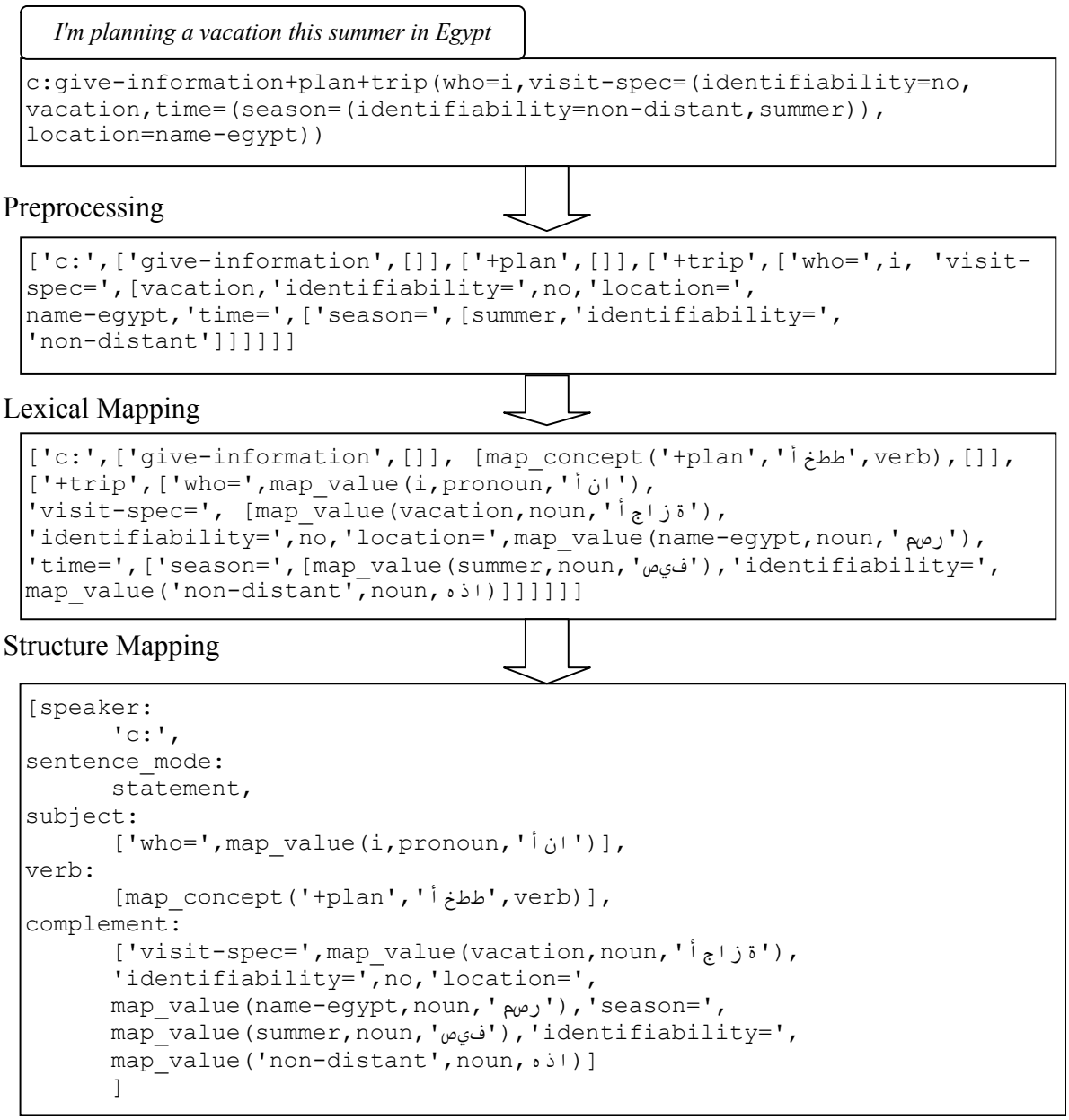

FIGURE 5: An illustrative example of mapping an IF to FS of a statement

As an example of mapping an IF to a sentence of mode command, consider the IF:

```
c: reserve four rooms for us
c:request-action+action+room(action=e-reserve-2, for-whom=we,
room-spec=(room,quantity=4))
```

When this IF is fed into the mapper, the FS structure shown in Figure 6 is produced.

```
[speaker:
      'c:',
sentence_mode:
      command,
verb:
      [map_value('e-reserve-2',verb,'أحجز')],
complement:
      ['for-whom=',map_value(we,pronoun,نحن),
       'room-spec=',
map_value(room,noun,'غرفة'),'quantity=',4]
]
```

FIGURE 6: An Example of a FS of a command

As an example of mapping an IF to a sentence of mode interrogative, consider IF:

```
c: Is there a train from Cairo to Aswan?
c:request-information+existence+transportation)
transportation-spec=train,origin=name-cairo, destination=name-
aswan(
```

When this IF is fed into the mapper, the FS shown in Figure 7 is produced.

```
[speaker:
   'c:',
sentence_mode:
   question,
interrogative:
   [map_value(do,noun,هل)],
verb:
   [map_concept('+existence',noun,يـوجد,verb)],
complement:
   ['transportation-spec=',map_value(train,noun,'قطار'),
   'origin=',map_value(name-cairo,noun,'القـاهرة'),
   'destination=',map_value(name-aswan,noun,'أسوان')]
   ]
```

FIGURE 7: An Example of feature structure for an interrogative

As an example of mapping an IF to a fragment, consider the IF:

```
C: and the child
c:give-information+concept (conjunction=discourse,
person-spec=(child,identifiability=yes))
```

When this IF is fed into the mapper, the FS shown in Figure 8.

```
[speaker:
      'c:',
sentence_mode:
      fragment,
coordination:
      ['conjunction=',map_value(discourse,particle,'و')],
complement:
      ['person-spec=',map_value(child,noun,'طفل'),
      'identifiability=',yes]
]
```

FIGURE 8: An Example of feature structure for a fragment

6. ISSUES RELATED TO MAPPING INTERLINGUA INTO ARABIC FEATURE STRUCTURE

In this section, we will discuss issues related to the mapping of interlingua into Arabic FS, and present how we have handled them.

6.1 Issues Related to Lexical Mapping

Lexicalization ambiguity resolution. During the mapping of the interlingua into a FS, the lexicalization ambiguity arose. Lexicalization ambiguity occurred because a value produces more than one Arabic word with different meaning (e.g. guide could be mapped to '**مرشد سياحي**' or '**دليل**'). We resolved this ambiguity by applying the lexical mapping using a value-argument pair, (i.e. `person-spec=guide` and `info-object=guide`).

Implicit values. In general, the value of arguments in IF is mapped into Arabic lexemes. In particular, there are certain argument-subargument combinations that map to implicit values that need special handling during mapping. For example, in the following interlingua, the expression "`age=(quantity=8)`" refers to the values "age and 8". The value "age" is an implicit value. Thus, our system maps the values in the following interlingua to "**أبن عمر ثماني**" rather than "**أبن ثماني**". This is illustrated by the following example:

My son is eight

```
c:give-information+personal-data (experiencer=
(offspring, sex=male, whose=i), age=(quantity=8))
أبني عمره ثمانية
```

6.2 Issues Related to Structure Mapping

Implicit constituents. During the recognition of constitutes, we found that some constitutes could be implicitly defined in the IF. The following implicit constitutes are handled during the structure mapping:

- For the interrogative sentence where there is no indication in the IF of any interrogative particle, e.g. the reserved value `question`, the interrogative particle ('**هل**') is used. For example,

  ```
  Can someone pick us up to the apartment?
  request-information+feasibility+pick-up (feasibility=feasible,
  who=someone, to-whom=we,destination=(apartment,
  identifiability=yes))
  هل يستطيع أحد توصيلنا إلي الشقة ؟
  ```

- For a statement where there is no indication in the IF of a verb but there is an indication of a tense feature expressed by the argument `e-time=`, the verb '**أكون**' (to-be), is used. For example,

  ```
  your room will be available at two o'clock
  give-information+feature+room (e-time=following, room-
  spec=(room, whose=you), feature=(modifier=available),
  time=(start-time=clock=(hours=2))))
  غرفتك ستكون متاحة الساعة الثانية
  ```

Order of verb and noun phrases. Since Arabic syntactic structure is flexible in word order and there is no indication of explicit case markings in the IF representation, we assumed a default word order of noun phrases (NPs) and verb. The patterns that we follow in the generation of the Arabic syntactic structure include:

- <NP in nominative case>,

- <NP in nominative case>, <verb, trans>, <NP in accusative case>,
- <NP in nominative case>, <verb,intrans>, <Prep> <NP in genitive case>,
- <NP in nominative case>, <Prep>, <NP in genitive case>,
- <verb, trans>, <NP in accusative case>, and
- <verb, intrans>, <Prep>, <NP in accusative case>.

7. CONCLUSION

Arabic syntactic structure determination is an important task to be handled in interlinguabased generation. This paper presents a computational model which is designed to achieve this task in interlingua approach. It takes individual sentences represented in a specific interlingua formalism and produces a FS that reflects the syntactic structure of the target Arabic sentence. It utilizes a rulebased approach to this intrlignua-to-FS task. This approach is based on solid linguistic knowledge. It takes all the information about the target language from four knowledge resources: ontology, mapping lexicon, lexical mapping rules, and structural mapping rules. We have discussed issues related to the lexical and structure mapping encountered in the mapping of interlingua representations used in NESPOLE! into FSs of Arabic sentences. For these problems we have described how we handled them. Our future work is to design and implement the Arabic generator that takes the Arabic FS and generates the target Arabic sentence. This will require developing both Arabic morphological generator and Arabic sentence generator.

REFERENCES

Abul seoud, R. (2005). Generating Interlingua from Arabic Parsing Tree, M. Sc., Faculty of Engineering, Cairo University, Egypt.

Cavalli-Sforza, V., Soudi, A., and Mitamura, T. (2000). Arabic Morphology Generation Using a Concatenative Strategy, In the Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000), Seattle, PP. 86-93.

Habash, N. (2004).Large scale lexeme based Arabic morphological generation. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04). Fez, Morocco.

Leavitt, J. (1994). MORPHE: A Morphological Rule Compiler. Technical Report, CMU-CMT-94-MEMO.

Levin, L., Lavie, A., Woszczyna, M., Gates, D., Gavalda, M., Koll, D., and Waibel, A. (2000). The Janus III Translation System: Speech-to-Speech Translation in Multiple Domains, Machine Translation, 15(1-2), PP. 3-25.

Levin, L., Gates, D., Wallace, D., Peterson, K., Pianta, E., and Mana, N. (2003). The NESPOLE! Interchange Format, Project Deliverable D13.

Metze, F., McDonough, J., Soltau, H., Lavie, A., Levin, L., Langley, C., Schultz, T., Waibel, A., Cattoni,R., Lazzari, G., Mana, N., Pianesi, F., and Pianta, E.. (2002). Enhancing the Usability and Performance of NESPOLE! - a Real-World Speech-to-Speech Translation System. In Proceedings of HLT-2002 Human Language Technology Conference, San Diego, CA.

Mitamura T., Nyberg, E. (1992). Hierarchical Lexical Structure and Interpretive Mapping in Machine Translation, In the Proceedings of COLING-92.

Nwesri, A., Tahaghoghi, S., and Scholer. F. (2005). String Processing and Information Retrieval: 12th International Conference, SPIRE 2005, Buenos Aires, Argentina, LNCS, Springer Berlin / Heidelberg, PP. 206 – 217.

Shaalan, K., Abdel Monem, A., and Rafea, A. (2006). Arabic Morphological Generation from Interlingua, 4th International Conference on Intelligent Information Processing (ICIIP), 20-23 September, Adelaide Australia.

Soudi,, A., Cavalli-Sforza, V., and Jamari, A. (2002). A Prototype English-to-Arabic Interlingua-based MT System, In the Proceedings of the Workshop on Arabic Language Re-sources and Evaluation - Status and Prospects, The 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain.

Sughaiyer I. and Al-Kharashi,I. (2004). Arabic Morphological Analysis Techniques: A Comprehensive Survey. Journal of The American Society for Information Science and Technology, 55(3):189-213.

Tomita, M., Nyberg, E. (1988). Generation Kit and Transformation Kit, Version 3.2, User's, Manual, Technical Report, Carnegie Mellon-Center for Machine Translation.