A Prototype English-to-Arabic Interlingua-based MT system

Abdelhadi Soudi*, Violetta Cavalli-Sforza†, Abderrahim Jamari#

* CLC, Ecole Nationale de L'Industrie Minérale,
Av. Hadj Ahmed Cherkaoui, B-P: 753 Agdal, Rabat, Morocco
asoudi@enim.ac.ma

† Department of Computer Science
San Francisco State Univ., 1600 Holloway Avenue, San Francisco, California, U.S.A.
vcs@sfsu.edu

Institut Universitaire de la Recherche Scientifique
Rabat, Morocco
iramaj5@hotmail.com

Abstract

This paper describes an ongoing research project on English-to-Arabic Interlingua-based machine translation. Section 1 gives a description of the system that generates Arabic sentences from Interlingua representations (IRs). In section 2, we show how basic sentential components are mapped. In this context, we address some of the differences between English and Arabic such as agreement in number which cannot be transferred exactly from the IR of an English sentence. Results and an example translation are provided in section 3. In this context, we address the issue of word order variation in Arabic.

1. The Architecture of the Arabic Generation System

An Interlingual approach to machine translation (MT) has a number of advantages over other approaches, such as the 'transfer' model. In an Interlingua-based architecture, source text analysis and target text generation are divided into separate components. A language-independent intermediate representation (or Interlingua) mediates between these two components. The decoupling of the analysis and generation phases allows the system to handle multiple-language output and avoids the reconfiguration of the system for each new language.

In the KANT Interlingua-based MT system (Nyberg, and Mitamura, 1992), each sentence is first conveyed into tokens. The KANT analyzer uses a lexicon, a morphological analyzer, source language grammar and semantic information in order to parse the tokenized sentence into a feature structure (FS), a list of feature-value pairs that reflects the syntactic structure of the source language (i.e., English). The interpreter then uses mapping rules to convert the FS into an IR. An IR is a tree-structured representation that abstracts away many of the syntactic details of both source and target language, while conveying the meaning of the source language. In section 3 below, we provide an example of a source language FS, the IR produced from this FS and the target language FS produced from the IR.

Generation of the target language sentence begins with the IR. The system which generates Arabic sentences from IRs consists of 4 subsystems: the mapping system, the sentence generation system, the sentence/morphology generation interface and the morphological generation system, as shown in Figure 1 below.

First, the generation mapping rules convert the IR into an FS that reflects the syntactic structure of the target language. The FS is a list of feature-value pairs that reflects the syntactic structure of the target language. Target language lexicon entries are FSs. They are retrieved during mapping and added to the sentence FS under construction. The Genkit grammar analyzer and generator (Tomita and Nyberg, 1988) processes the input FS and generates a preliminary target sentence string, calling MORPHE when it encounters lexical symbols in the generation grammar.[1] This string is optionally run through the CODA post-processing system to produce the final target sentence.

1.1. The Mapping System

The mapping system produces FSs for Arabic from IRs, using a set of mapping rules and a mapping lexicon. The mapper recursively traverses the Interlingua, stopping at each level to examine slots and their fillers (features, concepts and nested Interlinguas). Testing a hierarchy of rule declarations, the mapper performs a structure-building operation called mapping. The goal and result of mapping is a target-language FS whose contents reflect the contents of the Interlingua, expressed in terms of the syntactic and lexical properties of the target language. The mapping process involves three main stages:

- Selecting lexical items for each Interlingua concept;
- Mapping the semantic roles for each Interlingua concept (slots in the Interlingua frame) to grammatical functions (slots in the FS);
- Mapping semantic features for each Interlingua concept to the appropriate syntactic features in the FS.

The mapper's knowledge is represented as mapping rules that are stored in a mapping hierarchy. The use of a hierarchy allows one to write specific rules for specific concept/lexeme pairs and general ruleswhich are inherited.

[1] The morphology/generation interface consists of a lisp program that defines some functions that are used to call the morphological generator from the sentence generator.

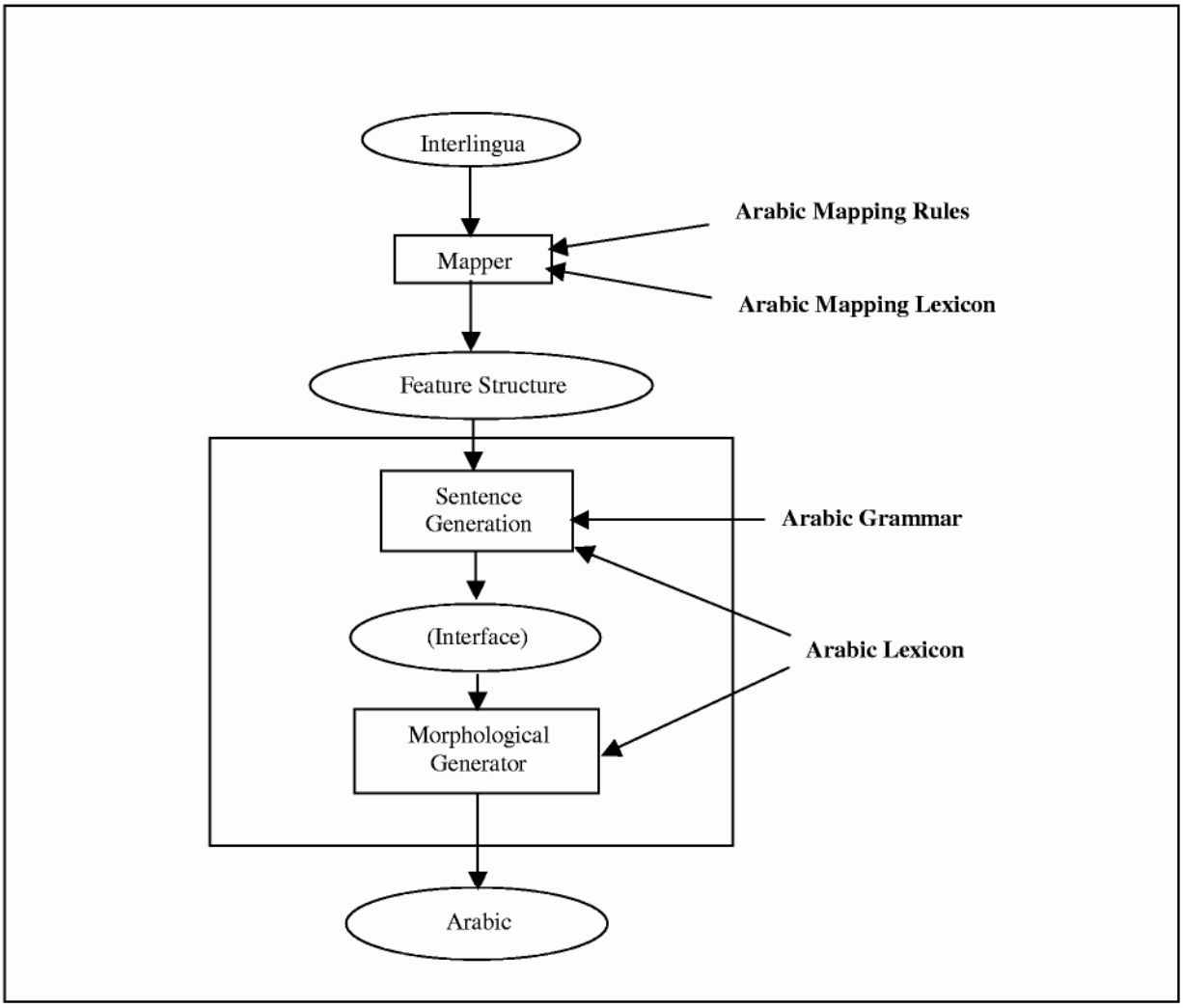

Figure 1. The Architecture of the Arabic Generation System

1.1.1. Concept Encoding Information

Each node in the mapping hierarchy has a name, a list of concepts, and a list of mapping rules to be executed. In addition, it has links connecting to one or more parent nodes. The examples in (1) below show how the concepts *shine* and *house* are encoded:

(1)
a. (node ?A-shine
:parents (VERB)
:encodes (*A-shine)
:rules ((:lex "ta?allaq")))

b. (node ?O-house
:parents (NOUN)
:encodes (*O-house)
:rules ((:lex "manzil")))

The node names *A-shine and *O-house are arbitrary symbols used to distinguish the nodes. They denote lexical interlingua concepts that would be associated with the lexical entries for the verb 'to shine' and the noun 'house' in the English lexicon. The :parents field specifies the part of speech that these nodes inherit from in the mapping hierarchy. The :encodes field and the :rules field specify which Interlingua concept this node will realize and the mapping rules associated with this node, respectively. ?A-shine and ?O-house denote the names of the lexical nodes used to determine the corresponding Arabic translation.

1.1.2. The Syntactic Lexicon

The syntactic lexicon consists of two parts: templates and entries. The templates specify the default contents of various types of lexical FSs. (2) below illustrates an Arabic syntactic template:

(2) (soft-template conj ((cat conj)))

The entries associate each lexeme with a template class and specify the unique features for that particular lexeme, as is illustrated by the following example:

(3) (conj "wa" ((ROOT "wa")))

1.1.3. The Mapping Rules

A mapping rule is a set of slots and values that specify operations involved in building an FS from an Interlingua. The lexical nodes in (1a-b) above illustrate a :lex mapping rule, which retrieves a translation from the target language lexicon. Mapping rules may also contain other directives

(e.g. such :map, :test, :add, :force-add, :consume, etc.) for performing other operations on the IR and FS.

For the sake of concreteness, consider the following mapping rule from (Soudi, 1999, pg. 13):

```
(4)  (:test (:sem (number plural)
            :syn (:not (human +)))
     :force-add ((agr ((gender f) (number sg)))))
```

The mapping rule above consists of a set of slots and values associated with the noun mapping hierarchy node. The :test slot specifies a set of conditions that must be passed for the rule to be applied. The :syn subslot specifies a negated condition on the FS, namely the feature (:not (human +), that must be met. The :sem subslot specifies a condition on the IR, namely the FVP (number plural). The slot :force-add indicates that the FS under construction should have feminine as its gender value and singular as its number value. This slot actually overrides information in the IR: the value of the number feature in the IR, namely plural, is overridden here by the singular. The mapping rule above applies to the sound plural feminine in Arabic (i.e., the *-At* class). By way of example, in the IR for the French noun *les animaux* 'animals', we would have, inter alia, the feature-value pairs (number plural) and (gender masculine). This information should be overridden for the corresponding Arabic noun 'Hayawanaat' – which is (human -) – by the feature-value pairs (number singular) and (gender feminine). Note that the information specified by the :force-add slot in the example above relates to subject-verb agreement. Thus, the sound plural noun *Hayawanaat* is plural but has 'singulative' agreement with verbs.

1.2. The Generation Grammar

To generate Arabic sentences, we have used Genkit (Generation Kit) (Tomita and Nyberg, 1988), a system that compiles a grammar written in a formalism called Pseudo-Unification Grammar into a sentence generation program. The generator follows a top-down, depth-first strategy for applying rules during generation.

The following example shows a unification-based grammar rule for generating sentences. The rule consists of a context-free phrase structure rule and a list of pseudo equations.

```
(5)  (<S> ==> (<NP> <VP>)
          (((x1 agr) = (x2 agr))
           (x1 == (x0 subj))
           (x1 case) = nom)
           (x2 = x0)))
```

The non-terminals in the phrase structure part of the rule are referenced in the constraint equations as $x_0 \ldots x_n$, where x_0 is the non-terminal in the left-hand side (here, <S>) and x_n is the *n-th* non-terminal in the right hand side. In these equations, x_1 represents <NP> and x_2 represents <*VP*>. The rule in (5) is for sentences with an <*NP*> and a <*VP*> that agree in number, person and gender. The equation ((x1 agr) = (xs agr)) indicates that the <*NP*>'s agr feature has a value that unifies with the value of the <*VP*>'s agr feature.

2. Arabic Noun and Verb Mappings

The generation of properly inflected Arabic verbs and nouns is a concern of both the mapper and the generator for a partial integration of the Arabic Morphology system into the KANT system). For example, the generation of correct agreement between nouns and their modifiers or other parts of the sentence may be performed either during mapping or during generation. Different cases must be considered:

(a) **Subject-Verb/Verb-Subject Agreement:** In Arabic, agreement in number between subject and verb depends on the nature of the subject of the sentence and word order. On a VS order, verbs do not agree in number with a plural subject. Agreement is always singular. Verbs, however, agree with their subjects in person and gender, as is illustrated by the following rule for generating a VS order sentence (from Soudi, 1999, pg. 16)):

```
(6)  (<s> ==> (<vp> <np>)
          (((x1 agr) = (x0 subj agr))
           ((x1 agr number) <= 'sg)
           (x2 == (x0 subj))
           ((x2 case) = nom)
           (x1 = x0)))
```

(b) **Intrinsic Number:** In most cases, the number feature for a noun is determined by the input sentence, reflected in the IR, and mapped directly from the IR into the FS by the mapper. Some nouns, however, may have agreement constraints already present in the lexicon. While lexical entries for nouns are usually assumed to be singular, certain nouns may be intrinsically plural in terms of agreement. For example, the noun *naAs* 'people', would contain the agreement information (number pl) in the lexicon, and the mapper should not override it with information that may be present in the Interlingua (for example, if the source language were Italian or Spanish, in which the word is a singular collective noun).

(c) **Number-Noun Agreement:** Number-noun agreement is governed by a set of complex rules. With the number 'one', agreement is as expected, but there may be a reversal of word order (e.g. *kitaabun waaHidun* 'one book' (nominative)). The number 'two' is expressed by the dual of the noun. Numbers 'three' through 'ten' require the noun to be plural and the gender of the number to be the opposite of the gender of the singular noun. For example: *xams* 'five' (masculine) *sanawaat* (plural of *sanat* 'year', feminine) but *xamsatu* 'five' (feminine) *kutub* (plural of *kitaab* 'book', masculine). Up to ten (plural of paucity), numbers and nouns agree in case, which is determined by the syntactic construction they appear in. Numbers above ten (plural of multiplicity) require a singular noun in the indefinite accusative. Agreement decisions can be made in the generator with the help of a callout function, but are most easily handled using the mapper.

3. An Example Translation and Results

To demonstrate the function of the components described in section 1, we will use the example sentence below:

(7) Jakarta and Bangkok are shining the most.

In the current system, the mapper takes as input the IR in (8), which is generated by the KANT analyzer and interpreter, and produces the FS for Arabic (9), using a set of mapping rules and a mapping lexicon (Soudi, 1999, pg. 20):

(8) **The Interlingua**

```
(*A-SHINE
  (FORM FINITE)
  (TENSE PRESENT)
  (MOOD DECLARATIVE)
  (PUNCTUATION PERIOD)
  (PROGRESSIVE +)
  (IMPERSONAL -)
  (ARGUMENT-CLASS AGENT)
  (MANNER
  (*M-THE-MOST
    (POSITION POSTVERBAL)
    (UNIT -)
    (DEGREE POSITIVE)))
    (AGENT
      (*G-COORDINATION
        (PERSON THIRD)
        (IMPLIED-REFERENCE +)
        (CONJUNCTION  (*CONJ-AND))
        (CONJUNCTS
          (:MULTIPLE
            (*PN-JAKARTA
              (UNIT -)
              (PERSON THIRD)
              (NUMBER SINGULAR)
              (REFERENCE NO-REFERENCE))
            (*PROP-BANGKOK
              (UNIT -)
              (PERSON THIRD)
              (NUMBER SINGULAR)
              (REFERENCE NO-REFERENCE)))))))
```

(9) **The FS**

```
((ADV ((CAT ADV) (ROOT "?ak#ar")))
 (form 4)
 (CAT V)
 (ROOT "ta?allaq")
 (VOICE ACT)
 (TENSE IMPERF)
 (MOOD INDIC)
 (SUBJ
   ((ELEMENT
      (*MULTIPLE*
          ((AGR ((GENDER F) (PERSON 3) (NUMBER SG)))
           (CAT N) (ROOT "jakarTaa"))
          ((AGR ((GENDER F) (PERSON 3) (NUMBER SG)))
           (CAT N) (ROOT "baankuuk"))))
     (CONJ ((CAT CONJ) (ROOT "wa")))))
 (PUNCTUATION ((ROOT PERIOD))))
```

Most of the linguistic features used in the KANT Interlingua and FS (e.g., punctuation, form, tense, argument class, number, person) should be self-evident. Some other features are artifacts of KANT's evolution as a technical text system. The IMPLIED-REFERENCE feature is used for nouns, such as the proper noun in the example above. G-COORDINATION contains all conjuncts that are coordinated and the conjunction that is used.[2]

The resulting FS serves as input to the Arabic morphological and sentence generator, producing Arabic surface forms:

(10) baAnkuwk wa jakaroTaA tata^alGaqaAni ^ako#ar

A major problem with the current implementation of the system relates to the word order variation in Arabic. Arabic is basically a VSO language, in which constituents can change order according to the constraints of text flow or discourse. The grammatical roles of constituents are identified by explicit morphological case markings. However, the KANT analyzer does not mark constituents as topic or focus. That is, this information is not provided in the IR. For example, there is no information structure for the system to decide whether to generate a VS order (12a) or an SV order (12b) from an IR for the English sentence in (11):

(11) Zayd ate the apple.

(12)

a. ?akala zayd-un t-tuffaaHat-a.
   ate Zayd-nom the-apple-acc

b. zayd-un ?akala t-tuffaaHata
   Zayd-nom ate the-apple.

Currently, the system produces all sentences in the S(=topic)V order.

While there are challenges to be worked out where the source language and target language differ greatly in their morphology and syntax, an Interlingua approach allows for a flexible integration of software modules for languages that differ in their realization of the same unit of meaning. Indeed, most of the morphological and syntactic differences between the source language and the target language can be handled by either the mapper or the generation grammar.

The system is still under construction. It has been tested on 29 different structures and has produced good results.

4. Conclusion

In this paper, we have described an ongoing research project on English-to-Arabic Interlingua-based machine translation. After giving a description of the system that generates Arabic sentences from IRs, we have shown how basic sentential components are mapped. In this context, we have addressed some of the differences between English and Arabic, such as agreement in number which cannot be transferred exactly from the IR of an English sentence. We have also provided an example translation and results.

[2] To promote representational consistency, the same structure is (*G-COORDINATION) is used if there is no explicit conjunction. In this case, the feature CONJUNCTION will have the value NULL.

5. References

Aronoff, M., 1994. *Morphology by Itself: Stems and Inflectional Classes*. Cambridge, Mass: MIT Press.

Beard, R., 1995. *Lexeme-Morpheme Base Morphology: A General Theory of Inflection and Word Formation*. State University of New York Press.

Cavalli-Sforza, V., A. Soudi, A., and T. Mitamura, 2000. Arabic Morphology Generation Using a Concatenative Strategy". *Proceedings of the North American Association For Computational Linguistics (NAACL)*, 2000, Seattle, United States.

Fassi Fehri, A., 1993. *Issues in the Structure of Arabic Clauses and Words*. Dordrecht, Holland: Kluwer Academic Publishers.

Mitamura, T., E.H. Nyberg, and J. Carbonell, 1991. An Efficient Interlingua Translation System For Multilingual Document Production. *Proceedings of the 3rd Machine Translation Summit*.

Nyberg, E.H. and T. Mitamura, 1992. The KANT System: Fast, Accurate, High Quality Translation in Practical Domains. *Proceedings of COLING'92*.

Schramm, G., 1962. An Outline of Classical Arabic Verb Structure. *Language*, 38:360-75.

Soudi, A., 1999. Interfacing an Arabic Morphological Generator with an Interlingua-based Machine Translation System. ms. Carnegie Mellon University, USA.

Soudi, A., V. Cavalli-Sforza, and A. Jamari, 2001. A Computational Lexeme-based Treatment of Arabic Morphology. *Proceedings of The Arabic Processing Workshop, Association For Computational Linguistics*, Toulouse, France.

Soudi, A., V. Cavalli-Sforza, and A. Jamari, 2002. The Arabic Noun System Generation. *Proceedings of the International Conference on Arabic Processing*, University of Manouba, Tunisia.

Timothy, A.B., 1990. Lexicographic Notation of Arabic Noun Pattern Morphemes and their Inflectional Features. *Proceedings of the Second Cambridge Conference on Bilingual Computing in Arabic and English*. No pagination.

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

Wright, W, 1966. *Lectures on The Comparative Grammars of Semitic Languages*. Amsterdam:Philo Press.

Wright, W., 1988. *A Grammar of the Arabic Language*. Cambridge:Cambridge University Press, 3rd edition.