[From: *Studies in machine translation*, international workshop…Riyadh, March 1985]

The Treatment of Grammatical Categories and Word Order in Machine Translation

Jonathan Slocum and Anthony Aristar

Microelectronics and Computer Technology Corporation (MCC)

Recent years have witnessed a mushrooming growth of interest in using computers to assist in the translation process. Interest in this topic is not by any means new; indeed, translation was perhaps the first non-numeric application proposed for computers, coming right after World War II. But interest has languished, following an early period of euphoria, until quite recently.

Machine Translation (MT) systems are now in active use around the world. This paper investigates the applicability of current and foreseeable MT technology to translation between one or more of modern Western European languages and Arabic. After an introduction in which we briefly sketch the history and current status of MT and comment on the situation vis-a-vis Arabic translation, we present some general design constraints for state-of-the-art MT systems before proceeding to consider problems posed by the Arabic language in particular. We then outline some approaches to the solutions of such problems, and indicate what special constraints these place on an MT system's design. We next present an architecture for a system that could handle Arabic, and draw some conclusions regarding the prospects for near-term application of such a system. We will not in this paper survey the field of MT; several such treatments are available elsewhere (e.g., [Slocum, 1984]).

Introduction

Machine Translation of natural human languages has had a long, colorful career. During its first decade in the 1950's, interest and support was fueled by visions of high-speed, high-quality translation of arbitrary texts (especially those of interest to the U.S. military and intelligence communities, who funded large MT projects). During its second decade in the 1960's, disillusionment crept in as the number and difficulty of the linguistic problems became increasingly obvious, and as it was realized that the translation problem was not nearly so amenable to automated solution as had been thought. The climax came with the delivery of the pessimistic National Academy of Sciences ALPAC report [1966]; as a result, most MT projects in the U.S. were cancelled. Optimism for MT, if not always R&D activity, was diminished worldwide, and its general disrepute resulted in a remarkable quiet third decade.

We are now into the fourth decade of MT, and there is not only a resurgence of interest throughout the world, but also a growing number of MT and MAT (Machine-aided Translation) systems in use by governments, business and industry. In part this interest is due to more realistic expectations of what is possible in MT, and realization that MT can be very useful though imperfect, but it is also true that the capabilities of the newer MT systems lie well beyond what was possible just one decade ago.

Concerning the Arabic language, one can note that some Arabic MT systems already exist; however, judging by the fact that this Workshop is being held, one must conclude that the results are not entirely satisfactory. In part this may be due to the fact that too little is known regarding contrastive issues between Arabic and other languages. We hope to show here that basic MT system technology seems capable of dealing with Arabic, and that producing a cost-effective Arabic MT system should be a matter of selecting the proper linguistic computational techniques upon which to base the system's development.

Building such a system will require a large-scale effort directed toward a formal linguistic description of Arabic, contrasting it with other languages of current translation interest (e.g., English and German). Such an effort, in conjunction with other work in progress to strengthen the theoretical underpinnings of MT technology, should pave the way for even better, more effective Arabic MT systems in the future. In addition, of course, one must assume that a substantial project will be mounted to develop Arabic technical terminology, and the associated translation pairings with other relevant languages.

General Design Constraints

During the course of many years' experience with software systems, applications of computers to Natural Language Processing (NLP) generally, and Machine Translation specifically, certain lessons have been learned that bear on the design of any modern MT system. It is now recognize that a successful MT system will adhere to a certain set of general design principles. We will briefly discuss several such principles before proceeding in the next section to present some design problems occasioned by the Arabic language in particular.

Separation of Program from Linguistic Theory

An MT system must clearly separate the program component from the linguistic component. Linguistic theory is not static. The linguistic theory on which an MT system is (initially) based must be able to undergo continual improvement with little or no impact on the programs implementing that theory. For example, the grammar rules which describe the languages covered by the system must be modifiable without concern for the programs that will utilize such rules for translation. Similarly, computational procedures may be found amenable to improvement with no undue restrictions imposed by the linguistic component.

Modularity

The requirement of modularity is the sine qua non of flexibility. In a parsing program, for example, one routine might be responsible for morphological analysis of words; another, for idiomatic analysis; another, for application of syntax rules; yet another, for application of transformations; and so forth. The observance of modularity is not to be confined to the programs alone, but applied to the linguistic component as well. With careful attention to separation of responsibility in this manner, a system will be easy to modify and extend in accordance with the dictates of experience. In an MT system especially, evolution must be provided for if the system is not to be rendered obsolete by its own design.

Linguistic Rule Base

The grammars and lexicons describing a language must be maintained in form optimized for use by linguists rather than MT programs. The issue of overall efficiency in research, development, and application precludes interest in machine efficiency alone. Machine Translation is an exceedingly difficult problem whose optimal solution is not yet well understood. Empirical results can and will dictate

that linguistic procedures be changed. For this to be effected by linguists, who are not computer scientists, the rule base must be expressed in a formalism with which they are familiar.

Minimal Constraints on Representation

The program component should impose minimal constraints on how the linguistic component represents interpretations of sentences. The standard representation formalism in modern linguistics, for example, is syntax trees; in related disciplines, other formalisms are preferred. In order to allow freedom of choice, special modules must be written for each desired representation, and the MT system must interface with these modules in a well-defined manner.

"Fail-soft" Mechanisms

One great drawback of traditional Natural Language Processing systems has been their fragility. When confronted by a sentence beyond the descriptions provided by the rule base, the parser usually terminates. Since such problems will arise with any system of fixed rules, some provision should be made to overcome them. In an MT system, a sentence which cannot be analyzed as a unit should be analyzed into the lowest possible number of phrases; these phrases should be translated individually. Each stage in translation should have a default result -- even if only its unchanged input -- which the following stage may make use of, if not recover from entirely.

Extension to New Languages

An MT system must admit extension through the addition of new languages; the work should involve little more than writing new grammars and lexicons. This means that the framework in which the linguistic theory is formulated and expressed must be able to accent for languages other than the ones to which it is originally applied. Historically, attempts to take a system designed for translation from/into one language and modify it for translation from/into another have not been notably successful. The reason for this is in part due to the typical lack of extensibility built into the fabric of MT systems.

Multilingual Translation

For many applications it is desirable to translate a text into not just one but several languages. Typical the amount of time spent analyzing a source text greatly exceeds that spent synthesizing its Target Language (TL) equivalent, so that a system able to translate into several languages following a single analysis has a decided practical advantage. It may even have a theoretical advantage

insofar as such practice counteracts a tendency to produce a grammar that analyzes a Source Language (SL) only to the extent required to translate into a particular TL. Some systems have been constructed ab initio using a single-target strategy; the usual result is a complete inability to translate into any other language without a total revision of the system.

A Framework for Application

If a new theory of MT is proposed, and is claimed to advance the state-of-the-art, it cannot indulge in the luxury of confining its attention to isolated problems in small texts. It is frequently true that attempts at large-scale application or testing reveal striking deficiencies in methods that work well in small-scale experiments. Some proposals, while perhaps workable in theory, require access to encyclopedic knowledge which may not be available in appropriate form for another century. To some extent this can be regarded as indicative of the difficulty of translation. Nevertheless, any proposed advance in MT today must address the problems encountered in industrial translation; in doing so, the theory will benefit considerably through refinement in a real-world environment. Among other things, this implies a serious concern for efficiency in the underlying programs. It also implies a means for resolving the text-processing problems confronting any MT system.

A Framework for Research

No system today or in the near future will constitute an optimum solution to the MT problem. Instead, it will at best constitute an implementation of the most highly developed linguistic and computational theories of translation. Both kinds of theory will continue to evolve, and both would benefit considerably through large-scale testing of new ideas. Since an advanced MT system would be a prime vehicle for such testing, it seems only reasonable to require it to support a research function. Both research and application stand to benefit from this arrangement.

Some Problems Occasioned by Arabic

Every language presents its own special problems, though some languages may share aspects of behavior (i.e., problems) and thus might share computational treatment (i.e., solutions to those problems). The major Western European languages exemplify the sharing of behavioral aspects. In this section, we illustrate some problems presented by Arabic which may have little or no correspondence to those encountered in the major Western European languages - languages which have most often been the context for applications of Natural Language Processing.

Morphology

With respect to morphology, Arabic is one of the most difficult of human languages -- at least, as far as computer processing is concerned. Generally speaking, the problem of morphology per se has been ignored because most Natural Language Processing systems have been developed for English — a language in which morphology does not seem to exhibit especially interesting behavior (i.e., hard problems to be solved). Only recently has there been a flurry of interest in morphology per se, and one of the proponents of a leading technique has admitted that Arabic morphology constitutes a particularly hard problem which cannot be handled by that system [Kay, 1985].

As if this were not bad enough, the practice of not marking vowels in written Arabic text (with the exception of the Koran and children's literature) makes the problems of Arabic morphological analysis even more difficult. Vowels are omitted, unless absolutely required for disambiguation, under the correct assumption that the (human) reader can determine the intended meaning of a word in its context. One native Arabic informant goes so far as to claim that he ignores any vowel signs that are present, on his first reading of a sentence, and then re-examines the sentence for vowel marks in order to confirm his tentative interpretation [Prof. Peter Abboud, personal communication].

Needless to say, correct Arabic morphological analysis is not a problem that computers can effectively deal with using techniques that assume an almost unambiguous analysis of words, and little in the way of knowledge of the real world that could guide disambiguation in semantic context. Even the simpler problem of deriving all possible analyses is rendered difficult by the fact that the available analysis techniques are geared toward a situation where all the letters of a word are present; typographical error/spelling correction (which, for example, can overcome the accidental omission of a single letter from a word) has been rare in NLP systems. Thus, while the techniques required to deal with Arabic morphology might be argued to exist, they have seldom, if ever, been tested.

Topicalization

With respect to word order, Arabic is nominally a VSO language. That is, in sentences the Verb will come first, followed by the Subject and then the Object. (By contrast, English is nominally SVO, and Japanese, SOV.) In Arabic, however, topicalization undermines this simplification. Under many circumstances the topic, rather than the verb, will occupy the first position in a sentence. Although leftward movement of the topic may not occur for stylistic or syntactic reasons, it is much more likely to occur in Arabic than in European languages -- to such an extent that, unless there is a clear reason for not preposing a topic, non-preposed

topics are felt to be unacceptable, even sometimes ungrammatical. Since the unmarked Arabic word order is VSOX (where X stands for extra constituents), this has far-reaching results, because the subject is typically highly topical, and there is a strong impetus towards SVOX syntax. There is a lesser but nevertheless strong tendency towards OVSX and even XVSO syntax as well, since either O or X may also be thematic (i.e., topical), and would then tend to come first. To add to the complexity, topicalization often leaves a pronoun in the position vacated by the topic NP.

The reverse constraint -- that very non-topical NP's may not appear first in a sentence, or even in the post-verbal slot usually filled by the subject -- is even stronger. A sentence such as

*ma:ta waladun fi l-yawmi l-ta:liy
(a child died the following day)

is not simply bad style, it is ungrammatical, for an indefinite nominal, "waladun" ("child"), clearly non-topical, precedes the NP "yawmi" ("day"), which is fairly topical, being definite. The sentence

*waladun ma:ta fi l-yawmil-ta:liy

is even worse, if that is possible, since "a child" now fills the sentence-initial position, which only a true topic may occupy. Thus this sentence can only appear as

ma:ta fi l-yawmi l-ta:liy waladun
died on the following day a child

or as

al-yawmu l-ta:liy ma:ta fiyhi waladun
the day the following died on-it a child

With "the day" occupying the overt topic position.

Pronominalization

A pronoun refers to a previously mentioned entity, almost never to a following one. Thus, sentences such as

as soon as he saw his mother, John knew her

where "he/his" refers to "John", are ungrammatical in Arabic. In the context of Arabic grammar, this has far-reaching results. For example, since object suffixes

are suffixes on the verb, and most Arabic sentences are verb-initial, the following should be the correct translation of

the man's sister saw him

where "him" refers to "man":

ra?at-hu ? uxtu l-rajuli.

But this would result in a pronoun preceding its referent, and thus the sentence has to be reorganized as

ra?at (3f.sg) il-rajula ?uxtu-hu
saw the man (obj) his sister (subj).

In short, the constraint on pronominalization here actually produces VOS order.

There are certain exceptions to this rule, such as indefinite third person masculine plurals in clauses such as

They say that...
yaquwluwna ?inna ...

where "they" refers to people in general. Pronouns are typically used to refer to some fact or idea which has not yet been mentioned, in a similar way to that used in English sentences such as

it is clear that John is tired

where "it" refers to "that John is tired". Thus the following type of sentence is common in Arabic:

qad ittadaha liy ?anna-hu yumkinukum ul-littifa: qa ma'a ha:'ula:i
l-na:si biduwni wasa: tatiy
PF has-become-clear to-me that-it is-possible-to-you the-agreement with
these the-people without my-mediation
(It has become clear to me that an agreement with these people is possible for
you without my mediation)

In this sentence, "-hu" ("it") refers to the whole clause following "the agreement." Certain conjunctions, e.g., "lamma:" ("when"), require a following verb, yet heavily modified subjects almost always follow less-modified objects or obliques. Thus when "lamma:" is used, and a heavy subject appears, infringements of the pronominal reference rule appear, as for example:

lamma: ya?tiyna: bi-thamarihi l-rajula yud'a: 'aliyyan ...
when comes-to-us with his-fruit the man called 'ali...

Genitive Constructions

The so-called "constructs" are very constrained in their ordering, co-occurrences, and interpretation, and are much more analogous to Indo-European compounds in these aspects than IE genitives. (Indeed, they are used both where IE languages would use genitives and where they would use compounds.) The head noun takes a special form (the "construct") and must directly precede the genitival nominal. Further, any definiteness value which the genitival nominal has must also inhere to the genitival nominal. A head noun cannot disagree in definiteness with its genitive.

Thus, the following phrases are grammatical:

mifta :ḥu ba:bi
key (const-NOM) the-door (GEN)
(key of the door)

mifta:ḥu ba:bi l-bayti
key (const-NOM) door (const-GEN) the-house (GEN)
(the key of the door of the house)

mifta:ḥu ba:bi baytin
key (const-NOM) door (const-GEN) house (GEN-IND)
(a key of a door of a house)

But "mifta: ḥu l-ba:bi" may never be interpreted as "a key of the house", nor may "mifta: ḥu ba:bin" be interpreted as "the key of a house." If a head noun disagrees in definiteness with its genitive, the two must be split into a nominal with a modifying prepositional phrase:

mifta:ḥun li-l-ba:bi
key (abs-NOM) to-the door
(a key of the door)

al-mifta:ḥu li-ba:bin
the-key to-door (abs-GEN)
(the key of some door)

The rigidity of the construction does not even allow adjectives which refer to a head noun to follow that noun directly, the normal position for an Arabic adjective, since that would result in an item intervening between head and genitive. Thus,

the big man's house

appears as

baytu 1-rajuli 1-kabiyri
house (const-NOM) the-man (GEN) the-big (M.GEN)

and

the man's big house

appears as

baytu 1-rajuli 1-kabiyri
house (const-NOM) the-man (GEN) the-big (M.NOM)

where only the case ending tells which noun the adjective agrees with. Where a number of adjectives appear with a construct, all adjectives follow the genitival phrase, and each one has to be linked with one of the nominals in the phrase; for example,

fiy busta:ni bayti ?abiy zaydin il-kabiyri 1-jamiyli
in garden (const) house (const) Abuw Zayd (GEN) the-big (M.GEN) the-beautiful (M.GEN)

is ambiguous as to which particular nominal the adjectives refer to, since all nominals are masculine, and all are in the genitive case. One of the possible meanings is

in the beautiful, big garden of Abu Zayd's house.

Another possible meaning is

in the garden of Abu Zayd's big, beautiful house

or even

in the beautiful garden of Abu Zayd's big house.

In "good" Arabic, embedded genitival phrases are acceptable, but it is not possible to coordinate head nouns to the same genitive. Thus,

busta: nu bayti ?abiy zaydin
the garden of the house of Abu Zayd

is possible, but the following is bad style:

busta:nu wa baytu ?abiy zaydin
the garden and house of Abu Zayd.

This should rather be phrased as

busta:nu ?abiy zaydin wa baytuhu
the garden of Abu Zayd and his house.

Passivization

The passive, which is formed by internal vowel change, is a totally different entity in Arabic from what it is in European languages. In fact, to call it a passive is something of a misnomer: "non-specific" or "indefinite" would be a better term for the form, since it is used almost exclusively to indicate that the agent is unspecified or unknown, and not, as in European languages and especially English, to move a non-agent into subject (usually-topic) position. This latter function is carried out by topic movement, which is very pervasive in Arabic. For example,

the house was built by the Arabs

would appear as

al-baytu (NOM) bana:-hu l-'arabu (NOM)
the house, built-it the Arabs.

As a result, clauses where a "passive" occurs in Arabic with a mentioned agent are rare, and indeed felt to be ungrammatical by most Arabs. While it is possible to translate many English passives directly into Arabic, e.g.,

hal turjima l-kita:bu ?ila 1-luğati l-'arabiyyati
(QUES) translated (PASS. 3rd m.sg) the-book to the-language the-Arabic
(Has the book been translated into Arabic?)

It is often necessary to translate sentences with an overt indefinite subject in English into Arabic passives as well. For example, compare

radiya 'an-hu
he-was-satisfied (ACT) with-it

with

rudiya 'an-hu
was-satisfied (3rd m.s.g.PASS) with-it
("Somebody/one was satisfied with it", or
"There was satisfaction about it").

Thus when translating from English into Arabic, for example, one has to deal with a trichotomy -- active versus passive versus indefinites -- which has to be translated into a partially overlapping but quite distinct dichotomy of active versus indefinite, a dichotomy upon which topicalization processes are overlaid, and which produces a very different word-order under very different constraints from that which obtains in English.

Verb Agreement

The most common Arabic word order is VSOX. With this order, if nothing intervenes between subject and verb, the verb agrees with a third person nominal subject in gender but not number (though since all non-human plurals are grammatically feminine, even if in the singular they are masculine, this can cause gender-marking to accidentally mark plurality). Thus, we get

saraqa (m.sg) l-rija:lu thiya:bahum
(the men stole their clothes)

and

saraqat (f.sg) il-bana:tu thiya: bahum
(the girls stole their clothes).

But if word-order is SVO, then we get

al-rija:lu saraquw (m.pl) thiya:bahum

for the first sentence, and

al-bana:tu saraqna (f.pl) thiya:bahum

for the second.

But if anything intervenes between verb and subject, then the verb may optionally not even agree in gender with the subject, and may appear in the masculine singular form:

saraqa (m.sg) thiya-bahum binta:ni min al-qaryati
(two girls from the village stole their clothes).

Thus any verb agreement rule needs to be sensitive to syntax, and also to the grammatical category of the subject, since none of this applies to pronouns. With a pronoun subject, the verb always agrees with the subject no matter what order the subject and verb appear in.

Grammatical Category and Word Order

We have sketched a few problems of word order and sentence structure, contrasting Arabic with Indo-European languages generally (but with English particularly) . Such problems of correspondence as we have seen have been describable by means of rules having direct correspondents in formal linguistic theory. In particular, these rules have employed the syntactic (and, in this case, arguably interlingual) categories Subject, Verb, and Object, in order to indicate word order, and the semantic or pragmatic categories "topic" and "heavy" in order to indicate more about the underlying sentence structure. There is reason to suppose that more elaborate sets of rules based on the principles of formal linguistics can be employed to advantage in computational applications such as the translation of Arabic; the next section expands on this theme.

Approaches of Solutions

If solutions were in hand to such problems as those outlined above, there would be little or no need for this Workshop on Computer Assisted Translation. However, such solutions are not yet known to exist -- especially with regard to knowledge of contrastive linguistics. What remains, therefore, includes at least: (1) an exhaustive analysis of Arabic morphology; (2) the proposal of grammatical categories and linguistic rules describing the behavior of Arabic; and (3) contrastive studies of Arabic with respect to selected languages of greatest interest. Other areas should be investigated as well -- for example, semantic and pragmatic models of the world -- but many of them lack near -- or intermediate-term payoff. Instead, they should be considered as part of a long-term solution.

Arabic Morphology

Not only must Arabic morphology be exhaustively analyzed, but the result must incorporate (or admit the production of) algorithms expressing the solution

of the morphological analysis problem. Such a solution must allow for substantial lexical ambiguity, due to missing vowel signs. Such lexical ambiguity should be resolved via general techniques (not specific to Arabic), and should be naturally integrated with the rest of the MT system.

Formal description of Arabic morphology are just beginning to appear in the literature (e.g.) [McCarthy, 1981]), but it is not yet clear whether any of these is extensible to the entire language. Morphological analysis techniques of sufficient power to deal with the degree of variation present in Arabic exist in theory, tools providing the ability to actually perform such analysis can be argued to exist (based on such representation techniques as letter trees containing [in the case of Arabic] consonants only), or at least can be proposed, but they apparently have not been applied to large-scale Arabic text analysis. Nevertheless, there is no reason to suppose that these approaches are inadequate in principles, whatever practical problems may remain.

One of the more vexing problems — that of analyzing vowel-less texts -- could be avoided for a while by the simple expedient of translating into Arabic, not out of it. This is surely the direction of greatest near-term interest to this audience. The "mirror image" problem of deciding when to omit vowels from the synthesized Arabic text could be side-stepped completely under the assumption that the Arabic readership will not object to the presence of vowel marks. (If the presence of vowel marks were found objectionable, however, the problem of deciding which few vowels must be marked to eliminate undecidable ambiguity would be just as difficult as analyzing texts without vowels: both would require a storing model of human comprehension).

Grammatical Categories and Linguistic Rules

A coherent system of grammatical categories and associated linguistic rules for describing Arabic must be developed. These must be useful for expressing, not only the behavior of Arabic itself, but also the similarities and contrasts between Arabic and the other languages of social and commercial interest. A project devoted to this effort must adopt/adapt/develop linguistic representation tools for encoding cross-linguistic transformations and word-order variations, and for synthesizing Arabic text.

We are not aware of any large-scale efforts to formally describe the Arabic language in these terms; however, such projects have been mounted for other languages (especially English, French, and German). No such grammars are by any means complete — nor will they be so in the near future -- but they have already proven useful and cost-effective in computational applications, especially MT. There is no reason to suppose that Arabic cannot in principle be

described using these same tools for linguistic representation. Since these tools are based upon phrase-structure (or equivalent) descriptions of language, it would seem productive to launch an effort to describe Arabic using such rules.

Constrastive Analysis of Arabic

Constrastive studies of Arabic with respect to selected languages of greatest interest (e.g., English and German) must be undertaken. Of the three problem areas whose solutions are outlined in this section, this is the least developed. The contrasts must be drawn, not via off-hand sketches of similarities and differences, but via sets of formal rules describing those similarities and differences precisely. A standard means of representing such rules is via transformations conditioned upon syntactic categories and structures, syntactic and semantic sub-categorization features, and sometimes individual lexical items.

Again, we are not aware of any significant efforts in this area with respect to Arabic, but such efforts have been mounted for other language pairs, and the results are being put to effective use, especially in MT systems. As before, we are aware of no restrictions on such techniques that would render them ineffective with respect to Arabic, and it is reasonable to assume, in the absence of any evidence to the contrary, that they would suffice. What is missing is the basic language data on which to base rules founded upon such techniques.

Effects on System Architecture

There are certain ways in which the attempt to deal with Arabic will influence the design of an MT system. A stronger influence, however, would be exerted by a decision to develop an Arabic MT system in the near future. We will discuss these issues in this section.

The most obvious design constraint imposed by Arabic would be on the morphological component. Analysis would have to be non-deterministic, to be sure; but, more importantly, an Arabic analyzer would mandate powerful tools for comparing/contrasting/evaluating interpretations, so as to choose the right one in context. For near-term applications, it is unlikely that sufficiently powerful methods could be developed. Thus efforts to produce an Arabic MT system would most profitably concentrate on translation into Arabic, and not out of it, with an attendant decision to generate (or omit?) all vowels rather than be concerned about which ones to delete and which to retain.

Given this decision, there would remain the problem of describing the morphology per se (including vowels). With a language like Arabic that exhibits a

large degree of internal inflection, the use of techniques geared toward affixational or even synthetic languages (e.g., English, French, and German) is doomed to failure. Rather, one must develop a new set of computational tools for this purpose. Luckily, pattern-matching techniques have been well-explored in Artificial Intelligence, and methods exist which can be adapted to the description of Arabic morphology. A morphological analysis tool based on such techniques could actually replace one geared toward less complex languages, since it is more powerful and can subsume the necessary functionality. This decision would probably be made on efficiency grounds.

In order to describe any human language in the near future, one needs strong syntactic tools, including subcategorization features and structural transformations. Strong semantic models do not yet exist for any language; indeed, it is not yet clear that the right solutions have even been proposed for this problem. Syntax rules of one form or another are currently the most powerful mechanism for describing human language in large-scale applications, and an MT system to be developed in the near term will consequently make substantial use of syntax rules.

In order to adequately deal with semantic/pragmatic categories such as "topic", and the notion "heavy", it will be necessary to have strong semantic models of the world, as well as of human speech interaction patterns. Such models await further research, however, because the necessary linguistic theories do not appear to exist. In order to integrate future semantic/pragmatic models with the largely syntactic models feasible now, it will be necessary to pay very careful attention to the details of system design, especially the modularity of the linguistic components. Pending strong proposals for semantic models, system developers will have to provide weaker semantic subcategorization features — a technique well understood, and already in widespread application -- as a stopgap measure. In the case of Arabic, where topicalization is so important, this will not be entirely satisfactory, but some beneficial effects will derive nonetheless.

Architectural Example

In order to provide a concrete example for discussion, we describe in this section the core of an existing MT system whose design, we believe, allows (if it does not already incorporate) the functionality necessary for translating into (and, eventually, out of) a language like Arabic. We describe METAL [Lehmann et al., 1981] -- the actual translation component of the larger LRC MT system developed at the Linguistics Research Center of the University of Texas, under the joint sponsorship of Siemens AG and Computer Gesellschaft Konstanz, both of West Germany. This system was recently delivered to the sponsors for market testing.

The top-level control structure in METAL is quite simple: the function TRANSLATE is invoked with a sentence in the Source Language (currently, German) and returns as its value an equivalent sentence in the Target Language (currently, English). TRANSLATE invokes four functions in succession: PARSE (for sentence analysis), INTEGRATE (for, e.g., intra- and extra-sentential anaphora resolution), TRANSFER (for structural translation), and GENERATE (for sentence synthesis). After sketching the format and content of dictionary entries, we will briefly discuss how the linguistic rules (lexicons and grammars) govern analysis, transfer, and synthesis, illustrating this three-step process using example German and English sentences. (The INTEGRATE step is performed by LISP code, rather than formal linguistic rules, and will not be described here; however, it is this component that would, e.g., identify the discourse topic so important in Arabic).

Dictionary Entries

METAL lexicons are divided into two types: monolingual, and bilingual (called "transfer"). A monolingual lexicon must be created for each of the languages involved in the translation process. Transfer lexicons link the source -- and target-language monolingual lexicons. Monolingual lexicons consist of entries for each lexical item. Each entry begins with a left parenthesis followed immediately by the canonical or "dictionary" form of the entry, then a series of feature labels, each with a sequence of zero or more values enclosed within parentheses. The entry is terminated by a right parenthesis. The entries for the German noun stem (NST) "Ausgabe" and the corresponding English NST "output" will serve as examples (see Figure 1).

Space constraints do not allow a full analysis of the entries. Simply stated, each monolingual entry provides METAL with the information necessary for analysis and synthesis of the lexical items. In addition to entries for distinct word stems, the METAL monolingual lexicons contain separate lexical entries for such morphemes as prefixes, infixes, suffixes, and punctuation.

Transfer lexicons consist essentially of canonical word pairs which indicate the many-many correspondence between the SL and TL word stems. Each pair may be augmented by an arbitrary collection of context restrictions that must be met in order for the indicated translation to take place. A sample transfer entry for the pair "Ausgabe - output" is included in Figure 1. There are no restrictions (conditions) placed on this transfer [indicating the translation of "Ausgabe" into "output", or vice versa], other than the Subject Area tag [DP = Data Processing].

(Ausgabe	**CAT(NST)**
ALO	(Ausgabe)
PLC	(WI)
SNS	(1)
TAG	(DP)
CL	(P-NS-O)
DR	(NPRD)
FC	(PP)
GD	(F)
SX	(N)
TY	(ABSDUR)

English monolingual entry:

(output	CAT(NST)
ALO	(output)
PLC	(WI)
SNS	(1)
TAG	(DP)
CL	(P-SS-01)
DR	(NPRD)
FC	(PP)
ON	(VO)
SX	(N)

German-English Transfer entry:

```
(Ausgabe (NST DP) 0      ! output (NST DP) 0)
                         +
```

Figure 1
German monolingual,
English monolingual, and
Transfer entries for
Ausgabe = output

As an example of transfer restriction, it is possible to specify that a given German preposition corresponds to any of several English prepositions depending on the semantic type of its object noun. Four entries for the German preposition "vor", shown in Figure 2, will illustrate. In these entries the appropriate English translation is defined by a restriction on semantic type (TY) and sometimes Grammatical Case (GC). These transfer entries are valid for ALL subject areas,

but must be tried in a particular order (as evidenced by numeric "preference factors" in the entries). Thus the presence in context of an object noun of semantic Type other than Abstract, Durative, or Punctual results in the English translation "in front of"; else the presence of a Dative object noun of type Abstract or Punctual will result in the English translation "before", else the presence of a Dative object noun of type Durative will result in the English translation "ago" [which will later be postponed]; otherwise, the translation "in front of" is chosen.

Morphological Analysis

Entries in METAL monolingual dictionaries are indexed by both "canonical form" (the usual spelling one finds in a printed dictionary) and "allomorph" (the stem, without productive affixes). The affixes themselves are separate dictionary entries; although their semantics is necessarily different in kind from content morphemes, they are treated identically by the system software. If a particular stem exhibits internal inflection (e.g., German nouns that umlaut in the plural), or varies for other reasons, then multiple entries are stored, one for each stem variation [allomorph]. At first this may seem wasteful, but the majority of such cases in our dictionaries are German strong verbs - which sometimes behave differently, depending on inflection, and thus would require separate entries anyway. For languages like Arabic, which exhibits substantial internal inflection, this approach is of course impractical; a replacement module would have to be implemented.

The actual analysis component is a separate module that accepts as input a word from the text, and produces as output a formally structured set of (all possible) interpretations. The current processing technique is based on a "letter tree" of the characters in each allomorph, and the analyzer may use this to perform spelling correction, among other functions. (For Arabic, this tree would presumably lack vowels, since such may not be present in the text. This would entail a second step in which any vowels that happened to be marked in the word would be used to select a subset of the interpretations proposed in the first step).

Sentential Analysis

For human-engineering reasons, one of the most convenient forms for expressing a grammar is via context-free (CF) phrase-structure (PS) rules. Context-free rules alone may or may not fully describe all human languages (see [Gazdar, 1983] for arguments that CF grammars are indeed sufficient), but, in any case, more general phrase-structure rules preclude efficient computational treatment, and CF rule-based systems seem to function as well as or better than any other technique, in practice. It has become traditional to augment the context-free

rules by associating with them procedures in some formal language (sometimes a programming language, such as LISP) in order to provide more generative power while maintaining computational tractability. In METAL, these formal "rule-body procedures" are invoked as soon as the parser finds a phrase matching their constituent phrase structure.

```
(vor (PREP ALL) 30                  ! in-front-of (PREP ALL) 0)
   OPTTY * ABS DUR PNT

vor (PREP ALL) 20                   ! before (PREP ALL) 0
   GCD
   TY ABS PNT

vor (PREP ALL) 10                   :ago (PREP ALL) 0
   GCD
   TY DUR

vor (PREP ALL) 0                    in-front-of (PREP ALL) 0
```

Figure 2
German-English Transfer entries
for vor = in front of, before, or ago

```
NN              NST              N-FLEX
 0               1                 2

(LVL0)          (REQWI)          (REQWF)

TEST            (INT1CL2CL)

CONSTR          (CPX1ALOCL)
                (CPY2NUCA)
                (CPY1WI

ENGLISH         (XFR1)
                (ADE1 ON)
                (CPY1MCDR)

SPANISH         …
```

Figure 3
A German Context-free PS rule for building a Noun Stem + an inflectional ending into a Noun

The traditional purpose of such procedures is to restrict the application of a rule by tests on syntax (e.g. number agreement between noun arid verb) and/or semantics (e.g., whether the proposed syntactic subject can be interpreted as an agent). If such tests fail, the syntactic phrase is not built. In METAL, these procedures not only accept or reject rule application, but they also construct an interpretation of the phrase. Traditional papers automatically build a "parse tree" and may add the output of such procedures as semantic information; in METAL, the parser (i.e., the LISP program) makes no commitment to a syntactic structure, but instead, linguistic procedures construct the interpretation (phrase) and compute its weight, or plausibility measure. The weight of a phrase is used when comparing it with any others that span the same sequence of words in order to identify the most likely reading.

A rule-body procedure in our system has several components: a constituent test part that checks the sons to ensure their utility in the current rule; an agreement TEST part to enforce syntactic and/or semantic correspondence among constituents; a phrase Constructor, which formulates the interpretation (phrase) defined by the current rule; and one or more Target-Language-specific transfer parts which operate during the second stage of translation (following complete sentence analysis). The inter-constituent test, the phrase constructor, and the transfer procedures may include calls to case frame procedures and/or structural transformations, as well as simpler routines to test and set syntactic and semantic features/values.

Case frames may apply semantic and syntactic agreement restrictions to the predicate (verb structure) and its arguments (noun and prepositional phrases) when constructing a clause. Each predicate's lexical entry specifies its possible "central arguments". For German, the case frame will identify the case role-players according to voice (e.g., active) and mood (e.g., indicative) of the clause, and information about each potential argument such as its semantic type, form (noun phrase or prepositional phrase), and grammatical case (e.g., accusative) or prepositional marker. The restrictions can be general, or specific to the individual verb, preposition, and/or noun. The frame will fail, causing application of the clause rule to be rejected, if any of the restrictions are not met. Otherwise, case roles are assigned to the central arguments and the "peripheral arguments" are then identified. (Arabic does not differ significantly from other languages with respect to the existence of "central" and "peripheral" arguments).

The geometry of interpretations typically (though not always) parallels their original phrase structure. In other words, they are usually topologically equivalent to what the parser would produce if it were automatically constructing a tree. Some rules, however, incorporate transformations which may arbitrarily alter the phrase being constructed. The transformation module allows a linguist

to specify a structural descriptor to any depth, to perform syntactic and/or semantic tests as in rule body procedures, and to specify a new structure into which the old is transformed. The transformation program attempts to match the "old" pattern descriptor with the currently instantiated phrase. If the match is successful, and the specified conditions are met, a new phrase is constructed using the "new" pattern descriptor, with the (old) matched phrase usually providing (most of) the structural contents, and constructor operations may further annotate the phrase with new features and/or values. The transformation module can have no effect on the parsing algorithm, whatever the outcome of its application, unless the rule is written so that failure to complete a transformation causes the interpretation to be rejected; in such a case, only the fact of the rejection has an effect on the parser: it abandons that search path, just as it would if any other condition in the rule-body procedure were unsatisfied.

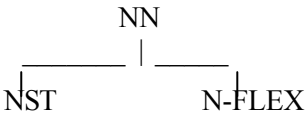
A grammar in METAL consists of a number of partially-ordered (Leveled), augmented phrase-structure syntax rule, plus a collection of indexed transformations. A relatively simple PS rule for building nouns will be used to illustrate the parts and format of METAL grammar rules (see Figure 3). Rules such as this could be written to describe Arabic, as well.

The first line consists of a left-hand element, the "father" node (here, NN), and one or more right-hand elements -- the "sons" (here, NST and N-FLEX). In the example rule, the left-hand element is the noun (NN) node and the right-hand elements are the noun stem (NST) and the nominal ending (N-FLEX) nodes. The second line enumerates the elements (from 0 to n) for reference in the rule-body procedure. Each constituent may have individual conditions, called "column tests", to restrict exactly what elements fit the rule. If any column test fails, the grammar rule will fail -- i.e., the parser will abandon its attempt to apply this rule. In this example, the column test for the first element (NST) requires it to be word-initial (WI) -- i.e., preceded by a blank space in the matrix sentence; the column test for the second element (N-FLEX) requires it to be word-final (WF) -- i.e., followed by a blank space.

In addition to the column tests, which apply only to single elements, each rule has a TEST part that states agreement restrictions between the right-hand elements. Failure of any agreement test will also result in failure of the entire rule. In the example rule, the single agreement test states that there must be an intersection (INT) of the inflectional class (CL) values for the two constituents; i.e., the values for the feature CL coded on the NST and the N-FLEX are compared to ensure that they have at least one value in common.

Only after all conditions have been satisfied is it possible for METAL to build the appropriate syntax tree. This is done in the CONSTR part of the rule, which

can also add or copy information in the form of features and values from the sons to the father. In the example rule, the Constructor (by not applying a transformation) would produce the tree represented below:



In the example rule-body procedure, the Constructor will copy all features with their associated values from the first element (i.e., the NST), except for the allomorph (ALO) and inflectional class (CL) features, using the operation CPX. CONSTR in this rule will also copy (CPY) the grammatical number (NU) and case (CA) features from the second constituent (the N-FLEX), and the word initial (WI) feature from the first constituent (the (NST).

Transformations may be applied in the TEST, CONSTR, and/or Transfer portions of grammar rules. These range from simple movement and deletion operations to highly complex transformations which add structure, perform tests, etc. The following exemplifies a simple movement transformation:

```
(XEM (& : 1 (& : 2 & : 3) )
        (& : 1 (& : 3 & : 2) ) )
```

This transformation simply exchanges the two sons (# 2 and # 3) of the current node (# 1): each ampersand represents one and only one constituent, or node. Transformations are used to move constituents around within a sentence -- as would be necessary for topic-fronting in Arabic.

Determining whether a sequence of words constitutes a clause is handled by a case frame, which is invoked in the TEST portion of clause-level rules. Simply stated, the case frame uses the argument information coded on the verb stem's lexical entry to identify its arguments, perform agreement tests, and label those arguments. In METAL, an argument may be a noun phrase, prepositional phrase, or adverbial phrase, depending on the verb. For a more detailed discussion of the grammar or lexicon, see [Bennett, 1982].

Transfer

The purpose of the TRANSFER module is to restructure the most plausible interpretation of the SL sentence into an interpretation of an equivalent sentence in the TL(s). Every non-terminal node (phrase) in every sentence interpretation has attached to it the "suspended" rule-body procedure that originally created it. This eliminates the need to search through a monolithic "transfer

grammar" for a matching pattern or routine -- and also eliminates the danger of inadvertently applying an inappropriate pattern or routine that happened to match (part of) the same structure. The suspended procedure associated with the root phrase in the most plausible interpretation is (re)invoked by TRANSFER. The appropriate Target-Language-specific Transfer part of a rule-body procedure can recursively transfer all or some of the node's sons (i.e., its non-terminal constituents) in any order, apply transformations, and/or lexically transfer a terminal son. Lexical transfer replaces a SL canonical form with a TL canonical form using the appropriate transfer lexicon. This process may be sensitive to sentential context. The TL stem is created and appropriate suffixes are added to create the proper TL word. Features in TL lexical entries may be used to help select the proper sense (i.e., word).

The final parts of a grammar rule are the Transfer sections [in Figure 3, ENGLISH and SPANISH]. In the multi-lingual METAL system, there is a separate transfer section for each Target Language; thus METAL can translate into multiple languages (e.g., German into English, Spanish, and/or Arabic). The appropriate Transfer section(s) are individually invoked only after a sentence [S] has been analyzed, at which point the system will perform the Transfer operations specified, generally moving down the tree to the terminal nodes where lexical substitution takes place. In our example rule (Figure 3), the first operation is (XFR 1) which causes the system to recursively invoke TRANSFER on the first son (i.e., the NST). Because the NST happens to be a terminal (lexical) node, it will be translated using the appropriate Transfer entry. The remaining two operations (ADF and CPY) are performed as the system ascends the tree. Thus, while analysis generally proceeds bottom-up, transfer proceeds top-down. At each node in the tree, all nodes below are accessible for reading (to determine context) and writing (to pass down information necessary for proper transfer).

Transfer in METAL is not a particularly simple process. Consider the following sentence pair:

German: die auszugebenden Resultate
Gloss: the to-be-output results
English: the results to be output

Here, the German participial verb form must be postponed in English. A transformation (conditioned on the form of the participial phrase) must be employed in cases like these. Prepositions present notorious problems; they must be translated and positioned with respect to their object NP's at least:

German: vor diesem Haus
English: in front of this house

German: vor dieser Woche
English: before this week

German: vor einer Woche
gloss: ago one week
English: one week ago

Clearly the relationship is complex: both the German noun (i.e., its semantic type) and its determiner (if any) influence the selection of a suitable English translation, as well as its position in the phrase.

A TL verb case frame, when applied during the transfer phase, will order the case role-fillers as required by the verb based on voice, mood, etc. The syntactic form of the central arguments is chosen and, if necessary, prepositions are introduced as specified in the Transfer verb entry. Consider the following examples:

German: aus Gold besteht die Tür
gloss: of gold consists the door
English: the door consists of gold

German: auf Gold besteht der Mann
gloss: on gold insists the man
English: the man insists on gold

Here, it is not only true that the complements must be re-ordered in English, but it is also necessary to translate the verb-preposition combination as a unit. This, in turn, may help disambiguate the semantic type of the matrix-subject, as the following examples illustrate:

German: aus Gold besteht er
gloss: of gold consists [it]
English: it consists of gold

German: auf Gold besteht er
gloss: on gold insists [he]
English: he insists on gold

Various of these factors can and do interact, as illustrated by the following example:

German: die aus Gold bestehende Tür
gloss: the of gold consisting door
English: the door consisting of Gold

In the METAL system, the Transfer procedures attached to analysis rules interact with complex Transfer lexical entries to determine the proper form and wording of the Target-Language structures. Generally speaking, each node appearing in an analysis tree is responsible for producing its appropriate translation, in context. (This is not always true, since a higher-level node can usurp the function of one or more of its sons — either performing transfer directly, or assigning a new transfer procedure to be executed in place of the original). We have found this combination of techniques (lexical transfer interacting with grammatical structural transfer procedures) to be a flexible and powerful tool that facilitates high-quality translation. It is highly efficient without requiring that analysis be performed once for each TL translation.

The top-level node (phrase) in the newly constructed TL tree is eventually returned by TRANSFER as its functional value, and this in turn is used for synthesis.

Synthesis

The GENERATE function synthesizes the translation by simply taking the TL tree produced by TRANSFER, and inflecting and appending together all of the lexical allomorphs (words and their inflections) located in its terminal nodes. The value of the function GENERATE is a sentence; it is returned to the function TRANSLATE, which returns that sentence as its functional value. For synthesis into multiple Target Languages, transfer and synthesis (but not analysis) may be invoked multiple times.

Summary

We have seen how the METAL uses syntactic categories and individual lexical entries - with syntactic, semantic, and pragmatic subcategorization features and transformations conditioned upon them - in order to effect translation from German into English in three stages. This system is now poised for commercial introduction. Early experiments with other language pairs (German into Spanish and Chinese, and English into German) indicate that these same techniques continue to be applicable, and there is no reason to suppose that METAL (or any MT system offering equivalent functionality) could not be extended to cover languages like Arabic as well.

Conclusions

Successful MT systems have been built and tested using techniques such as those employed in the METAL system, and there is no reason to suppose that Arabic will present any insurmountable problems to such approaches. None of these systems produce "perfect" translations: none ever will, if for no other reason than the fact that "perfect translation" cannot be defined, and is in any case considered to be impossible in principle (whether performed by humans or machines). But existing MT systems have proven that cost-effective machine translation is possible.

We know, therefore, that the basic hardware and software techniques for effective MT now exist. Furthermore, we know that effective linguistic techniques exist. Both kinds of techniques are evolving, and must continue to evolve if the quality and cost-effectiveness of MT is to be improved. It appears that some MT systems could be modified for application to Arabic; the problems seem to be related less to the lack of computational tools than to the lack of sufficient linguistic knowledge (to be fair, a much more difficult problem). In particular, our linguistic knowledge of Arabic seems deficient, but amenable to acquisition via the standard techniques used for other languages. The availability of a sufficient number of Arab-speaking linguists may be the major bottleneck.

The effective application of MT to Arabic will require hard work, but need not start from scratch. Some current systems are available for adaptation; this offers the opportunity of starting with substantial linguistic models of potential Source Language (e.g., English, French, German). What remains is the acquisition for formal linguistic theories of Arabic suitable for inclusion within such systems. Perhaps this Workshop marks the beginning of a serious effort in this direction.

Reference

ALPAC. Languages and Machines: Computers in Translation and Linguistics. A report by the Automatic Language Processing Advisory Committee [ALPAC], Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Publication 1416, Washington, D.C., 1966.

Bennett, W.S., "The Linguistic Component of METAL", Working Paper LRC-82-2, Linguistics Research Center, University of Texas, July 1982.

Gazdar, G., "Phrase Structure Grammars and Natural Languages", Proceedings of the English International Joint Conference on Artificial Intelligence, Karlsruhe, West Germany, 8-12 August 1983, vol. 1, pp. 556-565.

Kay, M. "The Nuts and Bolts of Lexical Access", presented at the Workshop on The Lexicon, Parsing, and Semantic Interpretation, City University of New York, 17-18 January 1985.

Lehmann, W.P., W.S. Bennett, J. Slocum, et al. "The METAL System", Final Technical Report RADC-TR-80-374, Rome Air Development Center, Griffiss AFB, New York, January 1981. Available as Report AO-97896, National Technical Information Service, U.S. Department of Commerce, Springfield, Va.

McCarthy, J., "A Prosodic Theory of Non-concatenation Morphology", Linguistic Inquiry 12,1981, pp. 373-417.

Slocum, J., "Machine Translation: Its History, Current Status, and Future Prospects", Keynote Address presented at the Tenth International Conference on Computational Linguistics [COLING 84], Stanford University, California, 2-6 July 1984. Also available as Working Paper LRC-84-3, Linguistics Research Center, University of Texas, May 1984.

* * *