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

"Computer Aided Translation: Design and Implementation".

Mr. O.K. Ewell,

Michael Downs Associates

It is indeed an honour to be a part of this International Conference on Computer Aided Translation. As I am not a linguist, but an Information Scientist, my address presumes that we accept the established machine translation theories and practices as true and correct, and that we agree, in principle with the concept, practicality and cost effectiveness of its application. My presentation will focus upon the more practical aspects of the design and implementation of computer systems that perform automatic translation, specifically from the systems that perform automatic translation, specifically from the English to the Arabic language.

While I have not been involved in language translation for the number of years as my colleagues, I have been involved in the design of computer hardware and software for the English to Arabic translation market for more than four years.

As a direct result of this involvement, which includes years of research into the ultimate end users requirements, exhausting evaluation of translation systems, hardware components, and most importantly the man-machine interface, Michael Downs Associates has conceived a design for an optimum English-to-Arabic Translation System, specifically for use in the Kingdom of Saudi Arabia. The Components of this system are on display in the outer hall. Briefly, the Hardware components are as follows:

The "Host" Computer

The "Host" computer is completely self contained in a sealed unit that fits beneath or beside a desk, and doesn't require any special environmental or electrical treatment. The architecture is based upon Digital Equipment Corporations KDJII processor which features a 15 MZ Clock and the versatile Q-Bus backplane. The System features a 80 Megabyte or 160 Megabyte sealed "Winchester" Disk sub-systems; Up to 4 Megabytes of "main" memory manufactured to withstand high temperatures; a 60 megabyte cartridge tape subsystem; and will allow connection of up to 128 either local or remotely connected *Translator Workstations*™, printers and other devices.

The Translator Workstation^{tm}

The Translator Workstation^{tm} is a self contained human-engineered workstation with multi-lingual capabilities. The workstation can display forty-two foreign character sets, and provides the facilities of a full-featured wordprocessing package, as well as the capability for the translator-user to transfer his reference cards to an electronic reference card system, fully integrated with the multi-lingual wordprocessor. The workstation will also run software that is available for the MS-DOS and CPM/86 operating systems such as LOTUS 1-2-3 and other personnel productivity and information center software.

The HOST software consists of:

A Relational Database Management System with a 4GL Language, which will allow the user to develop applications in any of the languages supported by the workstation from six to ten times faster than a conventional programming language. This System is the Information Management Processing and Reporting System, or IMPRS for short.

Extensive Network and Communications software which fully integrated the multi-processor network, and allows communications with foreign devices such as typesetting machines, Optical Character Recognition devices, Telex, modem and other communications devices, and other computer systems, such as the IBM computers under SNA, 3270, or 2780 communications protocols.

And finally the automatic language translation system. For our purposes, we have selected the Weidner Communications MACROCAT system, which automatically translates from the English to the Arabic Language as well as from the English to the French, German, Spanish, Portuguese and Italian languages; and from Spanish, German, French, and Japanese to the English language.

Each of the components of our system is loosely coupled for flexible integration at the time of implementation. Our system design is based upon the principles of distributed processing, and effectively integrated mainframe, minicomputer, and microcomputers into powerful management information networks with the ability to translate the information and data into the language desired by the workstation user, as it passes from the workstation through the host and back to the workstation, or to another workstation or computer in the network.

As this conference is concerned mainly with automatic language translation, I will devote the remainder of the time describing the Weidner language translation process.

The Weidner System is a dictionary based system. Several dictionaries can be on-line simultaneously, including a core-dictionary that is supplied with the system; industry/discipline specific terminology dictionaries; and user-specified terminology. The document to he translated may use any one or more of these dictionaries. The system is menu driven to allow the user to update the dictionary, to perform vocabulary searches of a document in order to discover any words in the document that may *not* be in the dictionary, and to send the document through the machine translation process. Post-Editing, under our system is done at the specially designed workstation with full-featured wordprocessing.

Basically, the computer steps through four main processes to accomplish its raw translation:

Pre-Analysis: is the process of breaking the source text into the smallest units which are meaningful to the computer. These might be phrases, words, or one word, depending greatly upon the complexity of the source sentence structure. Performed in two steps, the computer first scans the source document with the pre-processor routine which isolates sentence units, inserts formatting markers and identifies typesetting commands. Then the dictionary module morphologically analyzes source words to remove inflection and also locates the definition of the stem forms of target words in the system dictionary. Inflection information is preserved for later use.

Analysis: is the next phase. During this phase, the computer reconstructs the small individual units into the next higher level of meaning. Linguistically, this is called "parsing". The compaction routine isolates word groups that act as a single unit such as verb strings and idiomatic phrases. These word groups are then compacted into a single unit for the computer to deal with and are further flagged with syntactic and semantic information for later analysis. Homographs are examined within their context to determine which meaning should be selected. Finally, the phrase structure analysis routine performs a left to right parsing of the source text, identifying higher and higher levels of phrase structure. Phrase and clause structure information generated by this analysis is preserved for subsequent reorganization into their corresponding target language structure.

Transfer: is the third phase of machine translation and is where the structure of the source language is transformed into the target language structure. The insertion routine inserts articles, prepositions, etc., into the text as required by the target language. It also deletes words which remain from the source language where the words(s) is inappropriate. Words and phrases marked for reordering are moved to the appropriate positions in the target language sentence, and the text is reformatted accordingly. The expansion routine transforms the phrases that were previously "compacted" into their corresponding target language structure.

Synthesis: is the final phase of machine translation. Here the translated text receives its final touches. Inflection of verbs, adjectives, determiners, nouns, etc., is performed to provide "agreement" between sentence elements. Then the cosmetics routine makes the miscellaneous orthographic and other necessary, low-level adjustments that are required in the target language output. Finally, the post-processor puts the machine translation into the format of the original text, restores typesetting commands and performs special handling of untranslated words or phrases. The machine translation is then stored in an output file, awaiting editing and/or supervisor approval.

There are of course, considerable differences in linguistic structure between English and Arabic, especially when compared to language pairs like English-Spanish, English-French, and so forth. Accordingly, the analysis and transfer routines, although fundamentally similar to those used in other language pairs, are necessarily somewhat more complex for English-Arabic. I include these examples to illustrate this point; I will ask the Chairman to repeat the Arabic translation of these examples (as they came from the machine):

VERB PHRASE COMPACTION:

Even complex disjoined English verb phrases are identified, compacted, and later expanded into their corresponding Arabic Structures:

SOURCE TEXT:

The Boy *will* certainly *have finished* his work.
(The machine translates this as:)

سوف يكون الولد بالتأكيد قد أنهى عمله .

Will be the-boy *have finished* work-his with-the-certainty.

Idiom Compaction: Idiomatic expressions and other English phrases which do not translate "literally" into Arabic are likewise compacted and later expanded into equivalent Arabic structures:

SOURCE TEXT:

Your sister *looks* a lot *like* a woman I met in the store,
(the machine translates this as:)

تشبه أختك كثيرا إمرأة قابلتها في الدكان .

Resembles sister-your much (a) woman I-saw-her in the-store.

Homographs: English words having identical orthographic forms but different meanings and syntactic functions are analyzed and resolved to the appropriate corresponding Arabic forms:

SOURCE TEXT:

That man *that* you met knows *that* the system works,
(the machine translates this as:)

يعرف ذلك الرجل الذي قابلته أن الجهاز يشتغل .

That (DEMONSTR. ADJ) the-man *that* (REL. PRONOUN) you-met-him knows *that* (CONJUNCTION) the-system works.

SOURCE TEXT:

She plays the main role in *the plays.*
(the machine translates this as:)

تلعب الدور الرئيسي في المسرحيات .

She plays (VERB) the-role the-main in *the plays* (NOUN).

Reordering: of nouns and possessive pronouns, nouns and adjectives, and verbs (or auxiliary verbs) and subject noun phrases also takes place correctly:

SOURCE TEXT:

His little girl was playing in the street,
(the machine translates this as:)

كانت بنته الصغيرة تلعب في الشارع .

Was girl-his the-little playing in the-street.

Insertion: words not present in the English text but required in the Arabic translation are supplied:

SOURCE TEXT:

The student that I met was from Cairo.
(the machine translates this as:)

كان الطالب الذي قابلته من القاهرة.

Was the-student that I-met-him from Cairo.

DELETION: of words not required or permitted in the translation are deleted.

SOURCE TEXT:

I bought a new car.
(the machine translates this as:)

إشتريت سيارة جديدة.

I (INFLECTIONAL SUFFIX) - bought car new.

Agreement/Inflection: adjectives are inflected to agree in gender, number, case and definiteness with the nouns they modify. Verbs are inflected to agree with their subjects in person, number, gender, and so forth:

SOURCE TEXT:

The President and his aids were discussing the new proposals,
(the machine translates this as:)

كان الرئيس ومساعدوه يناقشون الإقتراحات الجديدة.

Were (3.m.sg.) The-President and aides-his discussing (3.m.pl.) the-proposals.

Unlike English, Arabic is a highly inflectional language. Nouns and adjectives are inflected by the Weidner system according to more that 250 different patterns, each consisting of 15 further possible inflections. Verbs are inflected according to more than 150 different basic patterns, each consisting of 106 different inflectional possibilities. Upon entering a word in the dictionary though, the translator need only supply two basic forms of the word in the case of nouns and adjectives (four, in the case of verbs). From that point, the system analyzes the elicited forms and automatically assigns the appropriate inflection rule number.

In conclusion, it has been pointed out, that there is no machine that can translate 100% correct (consistently). Post-translation editing is ultimately necessary. Through the use of the special workstations developed by Michael Downs Associates, together with the "host" computer equipment. Michael Downs Associates has effectively brought the cost of machine translation within reach of even the professional translator. The ability to "arabasize" applications programs, to develop custom Arabic language applications, and to communicate worldwide, in several languages and with virtually any computer system or device, makes our systems design far less expensive and far more flexible to implement. In the implementation of these systems, Michael Downs Associates utilizes the systems concept or "holistic" approach, bringing machine translation to the management information system, in a cost effective and efficient manner, while providing the additional benefit of literally integrating the information system into the organizations management structure.

* * *