SYSTRAN

The following presentation excerpts and paraphrases the highlights of the oral presentation given at the FBIS Seminar on Machine Translation, Monday, March 8, 1976, at Rosslyn, Virginia.

The major claim made for SYSTRAN is that it works — reliably, economically, and to the satisfaction of its users. It has continued to satisfy old and new users because it cannot become obsolete. It is in no way a black box. SYSTRAN has a very strong and flexible software framework enabling

1) immediate glossary expansion;

2) immediate implementation and testing of new or additional lexicographic, semantic and syntactic rules; and

3) universality in natural language translation.

The SYSTRAN system is "universal" in that it allows incorporation of additional translation capabilities (translation between new language pairs) without requiring modification of the existing software. Moreover, the addition of new translation capabilities requires only the implementation of additional source language analysis or target language synthesis programs. Everything else — all the parts that make the system work — remains the same. Thus, for example, since the system was already capable of translating from Russian to English, when the pilot Chinese to English capability was developed, only the development of a set of rules for analyzing Chinese as a source language was

necessary. Everything else, from the dictionary lookup and update programs to the English synthesis (generation) module, remained unchanged.

The SYSTRAN linguistic macro language is a great aid to the efficient development of these source language analysis and target language synthesis modules. These macros were developed to allow linguists to program their own rules. The formulation of the macros reflects types of operations (questions or tests, etc.) conceptualized by linguistic researchers as opposed to straight data processing-type programmers. The existence of these macros allows our linguists to modify existing programs quickly and with minimum effort and, of course, to write and check out new programs or even parsing or synthesis modules within relatively short periods of time.

The SYSTRAN translation system can run on either a 360 or a 370 with a minimum of 450K core storage available for application programs and dictionaries. Additional random access space is required for intermediate and sort work files. Input Russian text is accepted on 9-track tape or random access from either an ATS print file or MT/ST converted file. An alternative input file is accepted on punched cards which is normally used for system test. Output English translation can be printed on-line, via the SYSOUT printer, or offline, utilizing magnetic tape.

The system is programmed in direct assembler language and in SYSTRAN macros.

The computer processes batches of text at a rate of 300,000 words per CPU (Central Processing Unit) hour during an elapsed

time of 3 to 5 hours. Processor time per sentence is 1.2 seconds; for 1,000 words 18 seconds is average. Since the majority of refinements are additions of dictionary items and codes, rather than major additions to the programs, this speed will not lessen. It will increase, however, as the next generation of computers will further decrease cycles on the nanosecond level.

While SYSTRAN requires no human intervention in performing its translation tasks (other than the initial mounting of a disk pack or system tapes), it allows a maximum amount of interaction with its human components. First of all, because its linguists are its programmers, they know the system inside and out. On top of that, it produces hexadecimal displays with each sentence translated at the option of the user. Our linguists evaluate these records of the computer memory to identify translation problems and to identify precisely what program or routine is at fault. Having identified the problem, they then request SYSTRAN to produce concordance listings of a sufficient number of sentences containing exactly the same problems. After the linguist analyzes the resultant corpus, he designs, programs, implements and tests the necessary modifications. Modifications to the system do not always require such extensive research. Sometimes they are self-evident and require only a change in a single line of coding. The SYSTRAN macro instructions used by the linguist are automatically converted to assembler language during processing.

Since the Government has sponsored the refinement of this system, LATSEC, Inc. feels that any Government agency has the

right to have the system installed at minimal cost. (Expenses incurred when staff members train the user's staff to run the system should be covered.)

Maintenance costs, i.e., those costs involved in simply running the system to achieve raw output, can be directly calculated by any potential user by just finding out the per-hour cost of machine time at his installation. Any cost for improvement after installation depends on the user's requirements.

Our average keypunch or MT/ST input rate is about 1,500 words per hour. You can use this figure, along with how much your agency pays its keyers, to determine input costs. Of course, these costs would be virtually done away with if we could use optical character recognition devices. There is no pre-editing. Post-editing varies according to the user. Costs will vary according to the type of post-editing desired. According to FTD representatives, they are increasingly favoring the use of either un-edited, raw output or minimally edited output. (At a Bidder's Conference last September, Mr. Robert Wallace, the FTD SYSTRAN system monitor, said that nearly half of the 15 million words of text translated were distributed without post-editing.) NASA routinely used raw output of translations of working papers for the Apollo-Soyuz project. Yet, even when post-editing was performed, NASA found it both cheaper and faster to use machine translation rather than human translation.

At present, the system translates from Russian to English, from English to Russian, from English to French, and it has

lesser abilities in German to English and Chinese to English. Each capability is achieved by source language analysis and target language generation modules which fit interchangeably in the basic SYSTRAN frame.

As a final note. SYSTRAN works; it has proved itself useful as an operational system for the past six years. At this point, we are not interested in theoretical models of syntax; we are interested in making SYSTRAN the best possible machine translation system. It incorporates many aspects of modern linguistic thought. In doing so, it has transformed hypotheses about language into actual rules or descriptions of the behavior of language.

— END —

PETER TOMA

President and Chairman of the Board

Latsec, Inc.

La Jolla, California

Dr. Toma studied at the Universities of Budapest, Basel, Geneva, and Bonn, and at the Graduate Institute of International Studies in Geneva. He holds a Ph.D. in Communications Sciences, Slavistics, and Computer Sciences. He first developed machine translation algorithms in 1956 and joined the Georgetown (GAT) project in early 1956. As head of programming, he demonstrated that system at the Pentagon 6 June 1959. It was this system which was eventually converted for use at Oakridge and Euratom. (See The Serna System, Peter Toma, Georgetown Press, 1959.)

As a guest lecturer, Dr. Toma taught about machine translation at the Universities of Frankfurt, Bonn, and Cologne, the Institute of Technology in Darmstadt, and at the European Atomic Energy Commission (EURATOM) in 1960 and 1961.

In order to achieve, as early as possible, an operational system which would prove economical and reliable for the Government, Dr. Toma spent several years working in a private environment. The results were, first, Autotran and then Technotran.

In 1964, while the ALPAC hearings were in progress, Dr. Toma, working abroad, had a new system on the drawing board: a fully automatic, universal machine translation system. This system was SYSTRAN. Unde contract with the German Science Foundation, he implemented the system. Later, in July 1967, Air Force sponsorship supported further SYSTRAN development. In 1968, LATSEC, Inc. was formed. LATSEC, Inc.'s staff expanded SYSTRAN's translation capabilities to include English-to-Russian, English-to-French, German-to-English, and Chinese-to-English. In 1973, the formation of World Translation Center, Inc. furthered the development of the English-to-French system which has received significant recognition from the Canadian government. It was recently installed for the Commission of the European Communities and will be the first machine translation system to be used by the Common Market.