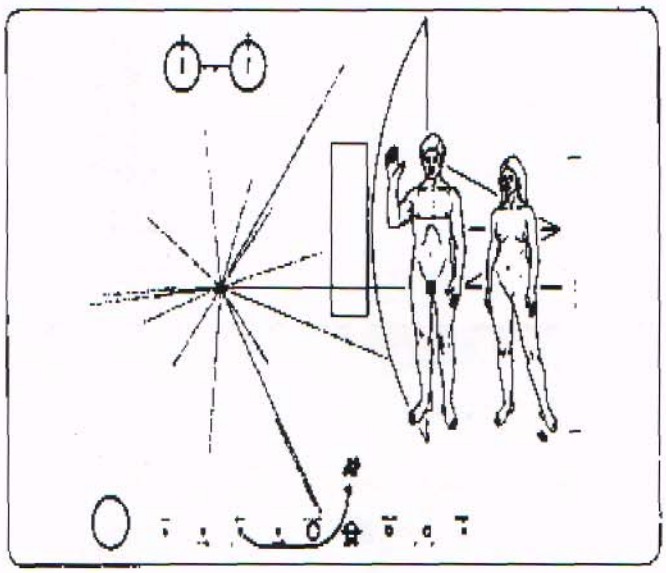
SPACE AGE AND MACHINE TRANSLATION

Dr. Michael Zarechnak



Launched by the Soviet Union in 1957, Sputnik I inaugurated the Space Age. The first man-made satellite to orbit the Earth shocked Americans. October 1 of this year marked the thirtieth anniversary of the birth of the Space Age. At Georgetown University (GU), government-funded machine translation (MT) research began in 1956.

Government support for the Georgetown research project ended in 1963, primarily because of the infamous ALPAC report, whose authors had no experience in large scale MT production efforts, and whose own flirtations with processing natural language fragments were complete failures. Yet, these were the individuals controlling the decision process during the investigation. Little doubt that their conclusion served their own interests – they authoritatively proclaimed that further machine translation improvement was not possible until linguistic research was separated from short-term machine translation goals, and directed to basic language studies, whether or not the results could ever be tested.

It is now evident that future MT development has to return to the assumption that basic language study cannot be improved until it incorporates the need to test hypotheses with large texts from various fields. The original GU MT dictionary was culled from texts of approximately ten million words. My current research is the further development of an intermediate language for multilingual machine translation. The paper I presented at the Berlin linguists conference in August includes the details of its design and development.

The rising interest in MT among linguists from many countries bears witness to the belief that eventually we shall overcome today's barriers to a higher quality of machine translation. How might we reach this goal? We shall take from experience what was successful, add the dreams of today's younger colleagues, and temper these with the need to handle massive amounts of language data using modern computers – and change the theoretical aspects until the discrepancies between real output and expected output reach a quality acceptable to an expert in the field. The user must be a partner in the future development of MT.

The basic assumptions at GU MT research included the following:

1. The source language should be analyzed structurally on word and sentence levels independently of the target language. The substitution operations were to be oriented around a variety of transfer routines using the output from the source language as an input for the synthesis of the target language.

2. The fundamental problem of MT is a linguistic one. In the 1950s we wanted to have larger memory in the the available IBM computers since we felt that the software based on the principles of contemporary structural descriptive linguistics (essentially a version of Bloomfieldian concepts as further developed by Zellig Harris) were quite suitable and useful for MT research.

Presently, there is a consensus that the computers have both enough memory and speed, while the linguistic software has not developed accordingly. The essential reason for this unwanted software failure is the wrong assumption that linguistic structures should be described in strict mathematical terms. At GU, we assumed that some segments of language are subject to strict formulation.

Yet, all things being equal, there was from the very beginning a strong feeling that language is a creative, open process, and some sort of a mixed system should be developed. The first signal for such a mixed system was the need to steer away from extreme procedures based on word-for-word translation and the completely formal phrase structure grammar which would produce all proper sentences and no deviant ones. We assumed that if the input sentences were incorrect, the output might be wrong and therefore that there should be an a priori postulation for post-editing.

Linguistic software has not kept pace with technology because its orientation is fragmented. It does not aim at large texts processed with dictionaries built around heavy depth coding under the control of a grammar which has a double track in its analysis: local operations and global operations.

3. Research in machine translation should be restricted for the foreseeable future to a limited field of investigation. In the GU MT research effort, the subfields were selected by the sponsoring agencies, NSF and CIA. We were examining approximately ten subfields from scientific Russian texts. The working procedures contained guidelines: while the research was text-oriented, it was not text limited. Any general observations known from the existing linguistic literature or uncovered during the research itself were formulated and programmed. We used a special programming language developed by Dr. A.F.R. Brown called LSC, (Linguistic Simulated Computer). This language is very easily used and debugged by a trained linguist.

Prior to Dr. Brown, our "liaison officer" with the programming team was Peter Toma. Toma later organized his own company to sell his SYSTRAN system, which was essentially based on the GU system. Its essential defect is what we called "clustering" – building the dictionary on a list of thousands of clustered entries reflecting the prevailing strings of words in a given field. This would be self-evident if the dictionary were printed along with the text generated. The GU MT is based on a split entry basis; so, given the more than 50,000 dictionary entries, we can generate over a quarter of a million linguistic forms plus the list of multiple choices and idioms.

The essential contribution of the SYSTRAN system is its flag-waving during all the years other MT groups did not try to develop working systems, and spent their efforts on fragments of the MT goal – naively believing that by building parts, they eventually will build a whole system. It looks as if we could have a long wait for this to happen.

The future promises a society of information services which certainly will include a good deal of multilingual translation. As in anything else, one can recognize the focal mainstreams in the information field. In our case, it is a flow via translations from a variety of languages belonging to different language families, Japanese, Chinese, Arabic, Russian, Spanish – into English.

These languages are each spoken by more that fifty million native speakers. Imagine that instead of translating in pairs from the source to the target language, we could analyze and synthesize the set of languages independently from one another. Then each could be translated once into the Intermediary Language (IL). From this language we could synthesize any of them when called for – and manage with 11 instead of 20 language pairs.

Ideally, the IL metalanguage should have the capacity to function as an algebraic representation of both paradigmatic units and their relationships, as well as iconic and/or indexical pointers to refer to extralinguistic fragments via the items from the linguistic text. Only with this kind of attitude can we solve such knotty problems as to why we have a smoke or a drink, but not an eat.

Michael Zarechnak Ph.D. is a pioneer of machine translatio research – he was a pariticipant in the original Georgettown project. Currently, hs is Associate Professor of Linguistics at Georgetown. He has published frequently on the subject of MT, and will serve as head grammarian, lexicographer and phonologist on the new GU MT project.