# U.S. GOVERNMENT SUPPORT AND USE OF MACHINE TRANSLATION: CURRENT STATUS

Thomas R. Pedtke, Associate Chief Scientist
National Air Intelligence Center, USAF
4180 Watson Way
Wright-Patterson AFB, OH 45433-5648
trp59@naic.wpafb.af.mil
937-257-6121

## Abstract

The United States Government has filled a key role in the development and application of Machine Translation technology for over four decades. A recent study by the White House Office of Science and Technology has reaffirmed the importance of this role. Two key world events, the emergence of Internet technology and the collapse of the former Soviet Union, have stimulated rapid changes in the status of Machine Translation requirements and applications.

A continuing need for Machine Translation systems in the United States military along with the application of Machine Translation systems on key United States Government networks has made Machine Translation systems available to tens of thousands of users. Advances in automating textual information processes and in testing and evaluation of the technology has further stimulated Machine Translation development and applications.

Although budget reductions will impact this continuing growth, renewed cooperation will ameliorate some of the impact and the emerging widespread use of Machine Translation could reverse the budget trends. Age old arguments between linguists and Machine Translation advocates seem to be giving way to recognition of mutual dependence and the potential for Win/Win outcomes.

The past five years have witnessed an accelerated exposure and application of Machine Translation technology in the United States Government unequaled in its 40 year history. However, with some budgetary adjustments, the next five years could be truly phenomenal. Advocates for Machine Translation technology and its applications are poised to meet the 21st Century and the Information Age with renewed vigor and practical applications which promise to end the debate over Machine Translation's viability forever.

## I. Introduction

Historically, the United States Government has played a critical role in the development of Machine Translation. The National Science Foundation, CIA, and the Air Force were all involved in extensive funding of Machine Translation in the United States prior to the ALPAC report in 1966. It is estimated that the United States Government and military agencies spent approximately $13 Million in the period 1956- 1965 on Machine Translation research. That may not seem like a large investment by today's standards but if we do a present value calculation it becomes $40 Million FY97 dollars which is significant. Money went primarily to eleven United States agencies. My organization, the National Air Intelligence Center (NAIC), which for most of the Cold War was known as the Foreign Technology Division or FTD, was the beneficiary of this sponsorship and development.[1]

In 1964, the IBM MARK II Russian-English System was installed at Wright-Patterson AFB. This system has historical significance as one of the first operational Machine Translation systems. As is well known, the MARK II System was followed by the Systran Russian-English system, and I am proud to report that some 34 years later my organization still has a large-scale Machine Translation production operation and that we are still engaged in promoting and developing Machine Translation systems.

## II. United States Government's Role Today

The United States Government continues to play a primary role in the field of Machine Translation. The document that most clearly states the case for Machine Translation from the Government's point of view is "Machine Translation Technology: A Potential Key to the Information Age.[2]" It was written in 1993 by the Committee on Industry and Technology of the Federal Coordinating Council for Science, Engineering and Technology. This group is referred to and better recognized as the FCCSET Committee on Industry and Technology. They are sponsored by the White House Office of Science and Technology. The document, despite some shortcomings, is about as close as the United States Government has come to reaching consensus on a Machine Translation policy. The document reflects the inputs from eight government agencies. I will summarize the committee's significant conclusions.

First, the report states that the United States Federal Government should play a catalytic role in targeting Machine Translation as a critical technology and promoting it and the involvement of United States industries in its development. Promotion of Machine Translation also advances the state of several generic technologies such as natural language processing, artificial intelligence, and optical character recognition. As a critical information technology, Machine Translation can spin off benefits for society in such areas as health, science, technology, and the environment. Additionally, a strategic investment by the Federal Government in planning, coordinating, and supporting Machine Translation development will also contribute to America's defense posture and make it more competitive in the Global Information Age.

Specific recommendations include: sponsoring of research; sponsoring enhancements to existing systems to meet validated government needs; evaluation of the performance of existing systems; sponsoring of Machine Translation workshops; and identifying new public and private sector needs in terms of language pairs, domains, and projection of future requirements.

I strongly endorse these recommendations. However, I believe the report should have taken a step further in two areas. First, it failed to address the insufficient and sporadic investment that the United States has made on Machine Translation in recent years. For example, from 1978 until 1993 when the report was issued, the United States put $20 million into Machine Translation R&D and considerable less into maintenance of existing systems. Europe, on the other hand, spent approximately $70 million and Japan approximately $200 million.[3]

My organization has had good budgets for Machine Translation development the past five years, first due to Defense initiatives sponsored by General Minihan, then at the Air Staff, and

Mr. Berbrich at the Defense Intelligence Agency and in the past two years due to initiatives sponsored by Dr. Markowitz of the Community Open Source Program Office on behalf of the Director of Central Intelligence. However, suddenly the future looks bleak in fiscal year 1998 and beyond with all but a small amount of funds being eliminated. The urgency seems to be missing for supporting Machine Translation research and development at a reasonable figure of $5 million or more a year. This amount would hardly seem to present an insurmountable budget challenge for the Intelligence Community, much less the United States Government, but it has.

Secondly, the report failed to exhort cooperation among United States organizations and come out as a strong advocate for a unified government approach to Machine Translation. From the earliest days of government sponsorship, there have been several different organizations sponsoring their own research and developing their own systems. This way of doing business simply reflects the way money is divided up in the Department of Defense and the Intelligence Community. However, this can and does lead to redundancies, rivalries, and profligate the wasting of scarce resources.

I am happy to say that there is now much more cooperation among Machine Translation developers in the United States Government. This, I believe, is largely attributable to downsizing and diminishing resources that have affected many government organizations in the last five years and stimulated this cooperation. Nonetheless, all this cooperation cannot completely obviate the impact of the severely reduced budgets and the consequences we will suffer in 1998 and beyond if we do not recognize the problem and fix it.

## III. Fast-Developing Technologies and Geopolitical Changes

Two events have recently had a significant impact on United States policy decisions concerning Machine Translation. The advent of the Internet and the impact of Internet technology in the last few years have had an enormous effect on the United States Government as well as the commercial world. Suddenly, there is a new way of gaining instantaneous access to world-wide users and providers of information. A portion of that information is in non-English languages in spite of English being the principal language of the Internet. Machine Translation has begun to play a significant role in this application. Because of networks and Internet technology, tens of thousands of analysts, librarians, researchers, and other government people are now aware of Machine Translation and can use Machine Translation on-line. The phenomenon is dynamic and its implications are enormous.

The other significant change is the political scene. The collapse of the former Soviet Union and the changing face of the political map have created new countries with languages of importance that only a few people speak. The lessening threat of global war, but the increased prospect of limited wars, peace-keeping missions, skirmishes, insurgencies, and anti-terrorism have increased the need for diverse linguistic tools, including Machine Translation, to increase communication and quickly assess and distribute information. Multinational coalition forces must communicate in many languages and deployments will be in countries with so-called third-tier languages. The nomination and probable selection of Army General Hugh Shelton as the next Chairman of the Joint Chiefs of Staff is reflective of the changing thinking about the makeup of future forces. He commanded the Special Operations Command composed of 47,000 troops, including Navy Seals, Green Berets, and other elite commandos. This force responds to hit-and-run tactics, is highly mobile, and can be deployed anywhere in the world. As such, they require up-to-date technologies like Machine Translation for communication and information acquisition.

## IV. Interesting, New or Dramatic Uses and Developments of Machine Translation in the United States Government

It is beyond the scope of this paper to present an exhaustive list of the extensive research and development and the many small-scale applications of Machine Translation in the government. Machine Translation R&D, mainly sponsored by DARPA and NSA, is well presented by Benoit and Jordan.[4] This is especially true in the period of the middle 80's through 1993.

I would like to concentrate on several specific areas: (1) Machine Translation in military operations; (2) Machine Translation on networks in the government; (3) Machine Translation testing and evaluation in the United States Government; and (4) current developments and future applications of Machine Translation at NAIC.

### A. Machine Translation for Military Operational Needs

In recent years, the United States Government has increasingly used Machine Translation in military operations. United States deployments in Somalia, Haiti, Saudi Arabia, Macedonia, and Bosnia have acted as the catalyst for developing new Machine Translation systems. The basic applications have been Machine Translation in Command, Control, Communication, Computers, and Intelligence (C4I) Systems, Machine Translation systems paired with speech recognition, and Machine Translation systems coupled with OCR's. Some interesting prototypes have been developed, some of which have been successful enough to become major projects with deployment.

One dramatic system is the Serbo-Croatian Speech Recognition and Translation System developed by Carnegie-Mellon. It was funded by DARPA and presented at the AMTA Symposium in Montreal last October. I will discuss in some detail three other systems.

#### 1. CCLINC

This is the Common Coalition Language System developed at Lincoln Lab by DARPA. The project is an ongoing and funded DARPA project. The system does two-way English-Korean text and speech translation, but the recent focus has been English-Korean translation of military material. The system uses a Lincoln Lab developed English-Korean Machine Translation engine, but Systran is used for Korean-English. Naval messages, command-and-control, and Commander's briefings for the CFC Korea have been the main focus of the development. The system was demonstrated in Korea in September 96 and April 97 and will go back in October 1997 for further testing and user feedback.

#### 2. TEKMAT

The Text English Korean Machine Aided Translation system was developed by the Naval Research and Development group and was funded by the United States Marine Corps. It was tested in Korea in August 1996. It has a combination of COTS and GOTS software, including Systran's English-Korean Machine Translation engine. TEKMAT is used to translate intelligence summary messages, briefing slides, unclassified memos, etc. At this time, it is unclear if funding is available to develop the system beyond a prototype.

### 3. FALCON II

The Forward Area Language Converter is a Machine Translation system developed as a lightweight field translator to allow a non-linguist to determine the military significance of an enemy document. The program, headed up by the Army Research Laboratory in Maryland, includes a unique partnership of other government agencies. DARPA, NAIC, and CIA are major contributors to the program as well as several other Army groups.

Falcon I, the precursor to Falcon II, was fielded and tested for French and Spanish Machine Translation in the Haiti deployment in August, 1995. The FALCON II has been deployed with the V Corps G2 in Bosnia. The system is a combination of COTS and GOTS hardware and software that has been integrated into a special product. The unit, weighing 30 lbs., consists of a Pentium computer, a built-in scanner, the Cuneiform OCR, six Systran Machine Translation systems including Serbo-Croatian, and communications ports.

## B. Machine Translation on United States Government Networks

This is an area of intense interest and activity and an unqualified success. The positive feedback from users who access Machine Translation on these systems has far exceeded expectations. Several of these systems will be discussed.

### 1. Open Source Information System (OSIS)

The OSIS is an Intelligence Community sponsored network that provides access to open source information for Intelligence Community users and other organizations interested in openly available information. All material is unclassified but the network is For Official Use Only and protected by a firewall to ensure adequate protection of copyrighted materials. There are now 25 major nodes on the network with about 10,000 users. All major production elements of the Intelligence Community are online and all Unified Commands have been or are being connected. Currently there is a very strong push to make this network available to every United States embassy and Defense Attache Office worldwide.

My organization, NAIC, is responsible for providing Machine Translation tools on this network. We call the Machine Translation package *The Web Translator*. Currently there are ten Systran systems available which translate from foreign source languages into English. Languages include Russian, German, French, Spanish, Italian, Serbo-Croatian, Portuguese, Chinese, Japanese, and Korean. In July of this year, English-French and English-Spanish systems were made available. Information to be translated can be input in several ways. Users can keystroke text directly into *The Web Translator,* paste electronic files into the graphical interface, or scan and

OCR data for porting to the system. In addition to being a private intranet, OSIS also provides gateways to the Internet and World Wide Web. *The Web Translator* has an additional feature which allows the user go to a foreign web site or type in the site's URL. The page will be returned with the text translated but retain the HTML page formatting and graphics.

## 2. Intelink

Intelink is a network put together by the Intelligence Community in December 1994. It uses Internet technology and can be thought of as a large intranet or, using current jargon, an extranet. A joint DIA, CIA, NSA organization manages the network. It is now made available at two security levels, Secret and Top Secret. (The intranet portion of OSIS is becoming known as Intelink-U although OSIS is more than an intranet with its Internet gateways). Intelink is world-wide and now has 35 intelligence organizations feeding it and many more using it. The user group with access now exceeds 100,000.

Although Intelink mainly provides finished intelligence reports, some tools have been placed online. *The Web Translator* with Systran systems was migrated up to Intelink this year. All the functionality of the OSIS web translator is provided in the Intelink versions except, of course, the ability to translate web pages on the Internet.

Use and user responses have been quite phenomenal. There are now several thousand translations being performed on *The Web Translator* each month. Customer demands and feedback on Intelink motivated NAIC to also put up Systran English source languages on both OSIS and Intelink. This is the beginning of establishing a two-way Machine Translation system on these networks for real time interaction.

## 3. CYBERTRANS

CYBERTRANS is a large enterprise-wide translation system used by the National Security Agency (NSA). It integrates commercial off-the-shelf and government Machine Translation products with text-manipulation facilities and NSA-supplied domain-dependent dictionaries. The system is used for quick gisting of information using NSA’s GISTER software (words and phrase translators), but Systran-based UNIX Machine Translation systems are the backbone of the CYBERTRANS suite of Machine Translation engines. Development is open ended. New language pairs are added as they are acquired or developed and dictionaries are constantly being expanded.

## 4. FILTER

FILTER is a program being developed by the National Technical Information Service (NTIS) of the Department of Commerce. NTIS’ mission is to ensure that appropriate United States Government information is made available to the public. FILTER is an acronym for Foreign Information Locator, Translator, and Electronic Retrieval. The purpose of the program is to provide government and private sector users of scientific, technical, and related information with the ability to locate, evaluate, and translate foreign language sources of information accessible via the Internet.

Existing government software developed for GISTER, CYBERTRANS, and TIPSTER will be used to produce an integrated system. In addition, NTIS has entered into a joint venture with SRA and Systran to use their software.

FILTER translates a user request from English using dynamically maintained lexicon tables. A search is made across the Internet limited to servers in the foreign language selected and narrowed by SRA's Nametag to the subject of interest. Then, the appropriate search findings are collected, collated, and cross-indexed. A "gist" of the findings is translated into English using NSA's GISTER word and phrase translators. Other FILTER features offered by the joint venture translation service providers include refined searches, a Systran raw Machine Translation, or post-editing of the raw Machine Translation.

A suite of Systran Machine Translation systems will be used, beginning with Japanese, Chinese, Russian, German, French, and Spanish. A Japanese prototype was tested in the spring of this year. The program will be made available to USG users for a nominal annual fee and to commercial customers on a per use, fee-based subscription. NTIS supplies products to more than 300 government agencies and in excess of 125,000 private sector customers.

## C. Machine Translation Testing and Evaluation in the United States Government

The United States Government's role in testing and evaluation of Machine Translation has been sporadic across several organizations. However, in some instances it has been quite dramatic, for example, the famous ALPAC report of 1966 written by the National Academy of Sciences.

Several organizations, including NSA and FBIS, have done internal evaluations of Machine Translation systems. Some government agencies have sponsored contract evaluations of systems and there have been individual reports and papers written by government Machine Translation researchers. My organization wrote a test plan in 1981 for evaluation of new Systran versions and attached it to all subsequent contracts. This activity fostered the development of the comparator program, the creation of a test corpus, and the random testing and scoring of new versions against old versions based on an acceptable improvement/degradation ratio.

Probably the most systematic Machine Translation studies were sponsored by DARPA under the Human Language Technology Program. They did some significant work on methodologies for evaluating Machine Translation, with reports coming out in 1991, 1992, 1993, and 1994. The Machine Translation evaluations from 1994 became widely known and the methodologies developed were presented and discussed at several Machine Translation forums.

I would cite the Federal Intelligent Document Understanding Laboratory (FIDUL) as the organization currently taking the lead in evaluation of Machine Translation and related technologies. FIDUL was chartered in 1994 and opened in 1995 under the sponsorship of CIA/ORD. Its mission is to conduct research, develop test methodologies, assess product development, promote standards, and foster the transfer of multilingual document understanding technologies.

FIDUL is now involved in a broad range of collaborative agreements with other government agencies to evaluate Machine Translation related projects. Because FIDUL is a relatively new organization that not everyone is aware of, I will go into some detail about the extent of their evaluation program.

The principal focus of FIDUL's methodology is to conduct comprehensive end-to-end assessments of document understanding technologies by evaluating a system in the context of a business or production process. These testing results are used to assess the readiness of the technology to transfer into operations. The methodology also identifies where continuing R&D should be focused. An example of this type of study is the Russian Language System Technology Transfer Assessment that involved automating scanning, OCRing, browsing, and verifying information.[5]

Other studies underway include:

(1) Prototype Chinese OCR and Machine Translation Technology Transfer Assessment and CHIOCR/Machine Translation Technology Insertion Project.

(2) Prototype black-box evaluation of an Arabic-English Machine Translation system using measures of adequacy, fluency, and informativeness of the English output.

(3) Russian Internet System Technology Transfer Assessment.

(4) Cross-language Text Browser Evaluation.

(5) Development of a Machine Translation Functional Proficiency Scale that will predict the Machine Translation system's output and whether its performance is useful for particular analytical text handling tasks.

(6) Partnership with ARL to conduct joint testing and defining of advanced research requirements of Army Foreign Language Processing Systems that can readily serve to meet similar needs in the Intelligence Community. Critical applications include high priority languages, particularly Chinese, Korean, and Serbo-Croatian.

Hopefully, the above discussion has provided an appreciation for the activities in which FIDUL is involved. I believe they are doing important work and making a significant contribution to real-time applications of Machine Translation and related technologies.

## D. Current Developments and Future Applications of Machine Translation at NAIC

Finally, I come to my organization, NAIC, and some of the key activities in which we've been involved. I will discuss on only a few of the significant highlights. NAIC has two primary roles in Machine Translation. First, NAIC is the Executive Agent for the DIA tri-service Foreign Language Machine Translation program. The Foreign Language Machine Translation program is under an overarching program called the Defense Intelligence Information Services Program (DIISP) and is called out in Chapter 5 of DoDIPP (Department of Defense Intelligence

Production Program). DIISP is the natural evolution of the original STIISP (S&T Intelligence Information Services Program). STIISP has been a model for interagency cooperation since it was first formed in the late 1960's to provide integrated information services support to S&T Intelligence production. When DoD production centers became integrated S&TI and GMI (General Military Intelligence) production centers in the early 1990's, STIISP changed to DIISP.

NAIC's Systran program was first designated as the Foreign Language Machine Translation program under STIISP in the early 1970's. Specific early tasking was to provide the Russian system to other production centers and to promote agreement on the greater use of computer and computer-aided translation systems. It is under the auspices and funding of DIISP that NAIC's development of the world renowned Systran Machine Translation systems are accomplished.

The second key role is from the Director of Central Intelligence Community Open Source Program Office (COSPO). It designates NAIC as the Lead Agency and Center of Excellence for providing Machine Translation tools on Intelligence Community networks (Intelink and OSIS discussed above).

Recently, after NAIC's Machine Translation systems were demonstrated to USAF General Henry Viccellio Jr. and Major General John Casciano, NAIC was encouraged to make the Machine Translation systems available to all United States embassies and United States Defense Attache Offices world-wide. We have been actively pursuing this task. The systems have already been placed at twenty USDAOs in Europe, the Far East, Africa, and South and Central America.

Buoyed by good budget resources during the past five years, NAIC has rapidly expanded the number of Systran systems. NAIC has initiated the development or collaborated in the development of twelve Systran systems to date, the most recent in development being Ukrainian and Cantonese. Assuming reasonable funding, we expect to commence development of six to eight new systems in the next five years. Given the expansion of NATO, two of these systems will undoubtedly be Polish and Czech.

The conversion of Systran code to C++ from IBM Assembler Language greatly influenced the use and portability of Systran within the United States Government. This migration effort was principally funded by COSPO for Russian, German, French, and Spanish and by FBIS for Japanese. The remaining language systems are newer and were developed from the outset in C++.

The C++ software has enabled such diverse configurations as standalone Windows 95 and Unix workstations, Windows NT and Unix network configurations, and *The Web Translator.* NAIC's web translator now runs under Unix with Sun workstations and will soon be modified to run under Windows NT and be compatible with Hewlett-Packard systems.

NAIC has fostered and joined numerous joint venture developments with other United States Government agencies for the specific application of Machine Translation systems. I can say without exaggeration that during the last five years there has been an accelerated exposure of

Machine Translation and acceptance of Machine Translation across the United States Government like no other five year period since Machine Translation's humble beginnings in 1950's. I believe Machine Translation for quick gisting is now an essential application. Machine Translation has become well known on Intelligence Community networks and is rapidly expanding to other United States Government networks. Recent publicity, unlike other such coverage in the past, has been very favorable.[6] All this suggests that the next five years could prove to be phenomenal for Machine Translation applications. NAIC will continue to play a very major role in developing, publicizing, and deploying Machine Translation systems within the United States Government.

## V. Conclusion

Despite many years of getting little respect, I now believe that the future outlook for Machine Translation in the United States Government is extremely positive. Never has Machine Translation been so well known to so many people. Many of the decades-old arguments about Machine Translation's role in translation and information processing technology have been resolved. People have finally come to terms with what Machine Translation can do well and what it cannot do and where the principal payoff is in embracing Machine Translation systems. The application of Machine Translation systems on United States Government networks has had a very positive effect on Machine Translation's role.

Equally encouraging is the new relationship between the Machine Translation community and Linguists. Machine Translation advocates now understand that the task of programming a computer to provide the quality of a human translation is extremely challenging and will probably take decades more development before acceptable results are achieved...if then. On the other hand, linguists are realizing that Machine Translation systems have key applications for them not the least of which are relieving them of many gisting tasks, speeding up their translations through machine assisted translations, and adding consistency to their human translation. We will probably never have enough linguists to serve our needs; however, the tradeoff is not Machine Translation vice human translation, it's Machine Translation vice no translation!

Despite what I believe is a very significant reduction in United States funding for Machine Translation development, the integration of COTS and GOTS Machine Translation products into many applications has and will continue to expand tremendously. Machine Translation for information scanning has a bright future in the 21st Century. There will be...there must be...a very concerted effort by the United States Government to expand language pairs to cover low density or third-tier languages. Machine Translation's future can only get better and the United States Government is proud to have played a significant role in Machine Translation's interesting saga and will continue to provide Machine Translation leadership into the 21st Century.

1. Hutchins, W.J., Machine Translation: Past, Present, Future. Chichester: Ellis Horwood Limited, 1986.

2 . FCCSET Committee on Industry and Technology report: Machine Translation Technology...A Potential Key to The Information Age, 1993.

3 . Benoit, John W. and Jordan, Pamela W., A Survey of Machine Translation Technology and Products. Information Systems Engineering Journal, Mitre, 1993.

4 .Benoit, John W. and Jordan, Pamela W., IBID.

5. FIDUL report: Russian Language System Technology Transfer Assessment, 14 January 1997

6 .Dandar, Ed, Is Outsourcing the Answer?, INSCOM Journal, March/April, 1997.