Insights from Tom Pedtke

Text of MT Summit keynote address, with introduction by Bob Clark

Bob Clark: Tom Pedtke has worked for the United States National Air Intelligence Center (NAIC) for over 30 years. He had a short career in the US Air Force and then returned as a civilian. He has played a fundamental role with the NAIC. His current status is Assistant Chief Scientist and he heads the NAIC Machine Translation project. He started working with Machine Translation four or five years ago and he is convinced that it is the way to the future.

Tom delivered the Keynote Address at the MT Summit VI conference in San Diego on 30 October, 1997 (the conference was reported in *LT4*). The speech was full of optimism and gave us a rare insight into US Government Machine Translation initiatives. Bearing in mind that systems being developed by the US Government (usually behind closed doors) eventually find their way into the marketplace, what Tom had to say was extremely encouraging. In addition, the issues that Tom freely discussed illustrate the distance that the US has come in adopting a much more open, less secretive approach to research and development. We were so impressed with the delivery and content of this speech that we have obtained a transcript from All Star Tapes, from which -with their permission- the following text is extracted.

The full text of Tom Pedtke's talk, or indeed any of the talks from the MT Summit (see our report in LT4) can be ordered from All Star Tapes by calling +1 619 270 8741.

Tom Pedtke: I have to admit that I am not an expert in Machine Translation. I am not a linguist. I struggle greatly with English. I speak no other languages. I was actually quite flattered when Muriel [Vasconcellos] asked me to be the keynote speaker and I accepted rather quickly. Later on she showed me the rest of the agenda and I became very concerned. Because some of the greatest names in Machine Translation are here at this conference and the gentleman [W. John Hutchins] that follows me wrote the book. So I figured, boy, am I in a lot of trouble, this caused a great deal of anxiety and then, I happened to be leafing through *Webster's Dictionary* and I came across 'keynote address'. It said, 'keynote address', the opening, throw some issues out on the table, stand by and watch all the experts argue. I can do this. I am eminently qualified to be a non-expert in Machine Translation. So, thank you very much for inviting me. We are going to talk about a whole variety of things and we will just put them out on the table and we will move around, talk about some general things in the United States Government, a little bit about the NAIC Programme.

Tom Pedtke

I met John Hutchins last night, he is a fine gentleman and I have to admit that three very key documents have a lot to do with this speech, besides Dale Bostad, who wrote most of it for me. John Hutchins' book was very stimulating. I read it and incorporated a lot of comments from it. The FCCSET study by the White House Office of Scientific and Technical Policy, they did a study in 1993, we are going to talk a little bit about that. Benoit and Jordan did a report in *Mitre Quarterly*, which was a very interesting exposé in Machine Translation.

This is the 50th anniversary for Machine Translation, that was 1947. A lot of wonderful things happened in 1947. The United States Air Force that I work for started in 1947. So did the Central Intelligence Agency. The National Air Intelligence Center was 30 years old in 1947. We actually trace our roots in the Foreign Technology Division all the way back to the Army Signal Corps in 1917. Machine Translation and, actually, even the

National Basketball Association, the United States 'Other' Air Force, traces its roots to 1947. So, a lot of celebrations. There is something unique about some of these organisations. They all have crests and emblems and their crests and emblems signify a little bit about what they do. The United States Air Force, for example, has got an eagle in high flight and, obviously it signifies the mission of the Air Force. The CIA's emblem is radiating spokes of the compass and that is to signify their world-wide focus. At the National Air Intelligence Center, in the middle of our crest is a sphinx, which is our pursuit of intelligence. Those organisations, of course, have got a lot of respect and one of the problems in Machine Translation is that we have had a little bit of Rodney Dangerfield type of problem. We do not get any respect. People talk about howlers all the time. Now, what are howlers? Howlers are those things when the computer translates something and it makes no sense. People break out into uncontrollable laughter. Those are howlers. We felt, in order to generate some respect for Machine Translation, we needed a logo. For this conference, I have designed a logo. There are some very key things about this logo. There is our personal computer with MT right in the middle. There are some spooky eyes there that probably indicate some of the intelligence communities interested in Machine Translation. Then there is a horseshoe and four grenades. There is a reason for this logo and that is because all these things have a lot in common. It turns out that, in horseshoes, grenades and Machine Translation, 'close' counts.

I want to talk a little about history. It is not an exhaustive history of Machine Translation and it is my perspective. I must tell you, as a non-linguist and as a non-Machine Translation person, I have, as the Japanese would put it, a perspective that 'the bottle is half-full', not that it is 'half-empty'. And that is because I do not understand anything about some of these foreign languages and anything I get out of Machine Translation is very positive for me.

I want to go over a few things in order to establish some foundation.

I divide up the history of Machine Translation into five periods, the conceptual years, that is because we had no concept of what was going on. Warren Weaver, in 1947, started talking about Machine Translation and using these new computers that had been developed during World War II for possibly translating language. In 1952 was the first MT conference and in 1954 the Georgetown experiments began. I have to tell you an interesting story. I met Muriel and she asked me if I knew Dorothy Pedtke. It turns out that she is my second cousin and she was involved in some of these early Georgetown experiments as a graduate student, working on those at Georgetown. It just goes to show you, every now and then, a Pedtke pops up about every 40 years in MT.

The early work started in about 1955 through 1966. I call it the early work because the Air Force first got interested. We had been hearing about some of these concepts for Machine Translation and we actually gave a task order to the Rome Air Development Center to develop a Machine Translation system for the Foreign Technology Division, at that time called the Air Technical Intelligence Center. By 1956, Mr Hutchins reports, Machine Translation research was really going on extensively world-wide. There was a lot of work being done. In 1963 we actually had the first operational system that we know of that was delivered to us at the Foreign Technology Division and it was the IBM Mark II system. Unfortunately, there was that dark moment in Machine Translation history when the ALPAC Report came out from the National Science Foundation. Now, we could argue forever as to whether or not there were hidden agendas there or whatever, but some scientists basically said, 'We have been investing very heavily in Machine Translation for ten years, we have not got anywhere, we are not going to get anywhere, and you should not spend any more money on this'. Unfortunately, that world-wide research really toned down to a few people who were just belligerent about it and absolutely believed that it would happen. People like Peter Toma said, 'No, we are going to continue to work on Machine Translation'. It is a good thing that he and other folks like him and Logos and other companies did, in fact, continue working on Machine Translation.

The Dark Ages started in 1967. But there were a few bright spots. One of the things was that we decided to abandon the IBM system and decided to contract with Systran. By 1969 we had our first version of the Systran Russian-English system. Obviously, being in the Intelligence community, we are very interested in the translations from foreign languages into English so that we can look at some of the technology that is being reported world-wide. By 1978, we developed an Editsys system, which allowed us to do some postediting so that we could do much more rapid translations.

The Renaissance sort of started in 1979, in my opinion. If you think back then, we were going through the Cold War and there was a lot of emphasis on intercontinental ballistic missiles. The United States and Russia were introducing multiple re-entry vehicles, multiple independently targeted re-entry vehicles, there were discussions of all kinds of exotic

technology, and it was a renewed interest that came in Machine Translation because someone had to start looking at all the things that were going on there. In 1983, we acquired French and German Systran systems. In 1985, another watermark decision, the Foreign Broadcast Information Service decided to team with Systran and the United States Air Force and develop a Japanese system. That development has continued since then. In 1987, we had our first interactive MT system where we were bringing Machine Translation to the analysts' desktops so that they could do things like translate titles and glossaries and indices, but it was still tied to the IBM mainframe. By 1990, we were actually using some raw Machine Translation in test and evaluation.

The Golden Era began in 1991. The reason I call it the Golden Era because it was fuelled with gold. We started to put a little bit more money into Machine Translation in our program. That started with the Drug Enforcement Agency that gave us some money to develop the Spanish system; we purchased the Spanish, Italian and Portuguese systems. The Community Open Source Program Office was formed and these are people in the Intelligence Community that look at open-source information. They were formed in about the 1993 time period. They came and visited the National Air Intelligence Center and they said, 'What could we do to invest some of this money that we have in some of your programs?'. We said, 'We really need to take Machine Translation, move it off of that mainframe computer and put it onto these UNIX and Windows PC environments. They funded that. That turned out to be one of the most critical decisions that was ever made in Machine Translation. None of it came from Strategic Planning. It just sort of happened. The NSA and FBI started to fund the Chinese system, DARPA and Navy the Korean system, and we will be talking about these in a minute. In 1995, our Windows and UNIX stand-alone systems were delivered. In 1996, we decided that we had this network called Open Source Information System, which we will be talking about, and we developed a thing called the Web Translator that would interact with the Internet and it would actually translate HTML pages that were in foreign languages. In 1997, we started Ukrainian, Cantonese, some research in retrieval mechanisms called InfoRaptor, we had UNIX and NT networks. That last period is only six years compared to the 12 years for many of the previous periods and the things that have occurred since 1991, at least from my perspective and the Air Force's perspective, have been truly phenomenal.

So, let me throw a few things out on the table, things that you might want to discuss. The first is MT policy in the United States Government. Perhaps the only document that is out there that really talks to policy is the FCCSET report. This is a subcommittee of the White House Office of Scientific and Technical Policy. They did a study in 1993 and they came up with five very good conclusions. They said that the United States Government should be involved in sponsoring research, sponsoring workshops, evaluating the performance of these MT systems, sponsoring enhancements and identifying requirements. Those are pretty good and there has been some response to it. For example, in research, the National Security Agency does extensive research. The Department of Defense, DARPA, Advanced Research and Projects Agency, has been doing a lot of MT research. We in the National Air Intelligence Center have been doing research. The Central Intelligence Agency has been doing research in MT.

In terms of workshops, the National Air Intelligence Center and a consortium that I happen to also chair actually put on a Machine Translation workshop a couple of years ago and it was very widely attended. The National Security Agency just had a big conference on MT and the Defense Intelligence Agency had a Foreign Language Day where MT was really featured. And, of course, we in the United States Government have been supporting the International Association for Machine Translation and the Association for Machine Translation in the Americas in the conferences that have been done.

In terms of evaluating performance, as I believe a lot of you know, there have been various studies. DARPA did one just a few years ago. There is a new federal laboratory in the United States Government called the Federal Intelligent Document Understanding Laboratory. Their mission is to evaluate technologies associated with the handling of textual information. This includes looking at the performance of scanners, OCR systems, Machine Translation systems and other tools. They have been very active in evaluating performance and at the National Air Intelligence Center we have even established an evaluation program, our comparator and our test corpus, that we use to evaluate the performance of the systems that we develop with Systran.

There has been a lot of sponsoring of enhancements. I think it is really phenomenal, last night I was next door here, looking at some of the demonstrations by the different contractors. It is obvious that there has been some very recent investments in improving a lot of Machine Translation systems. We have been doing a lot of that.

Identifying requirements, that has been a little tougher, because, unfortunately, the way that the United States Government funding works is that it tends to be a little isolated in programs and departments. Machine Translation is, unfortunately, something that cuts across the needs of all parts of the United States Government, but there seems to be no mechanism for looking at resources from that. I am going to talk about that later because I think that is a challenge that we have and that I personally want to take on. To somehow put together a more comprehensive program.

The study was great by FCCSET but had a couple of major shortfalls, in my opinion. Number one, it never said anything about resources and, number two, it did not talk about co-operation. Fortunately, a lot of co-operation has occurred but the unfortunate thing is this policy did not translate, so to speak, into a program response. And that is the unfortunate thing. We have some good policy here but a lot of individual type things occurred.

That is enough about policy. Let me talk a little bit about why we are doing Machine Translation in the Military and Intelligence. You know that sometimes people think we are kind of dirty because we are in Intelligence or in the Military. Actually, we believe that our role is to help preserve the peace. One of the ways to preserve the peace is to have a lot more understanding. If you cannot communicate and you do not understand each other, it is a little difficult to have peace. There has been a major information explosion throughout the world and it is necessary for us to be able to look at the information of the world and to put that into our policy-making and into the needs of our military and our decision makers. Unfortunately, at the same time there has been a major loss of linguistic skills. Not that the United States was ever necessarily the greatest at linguistic skills in the world. Being isolated over here in between the two big ponds, we have tended to focus on one language. However, in past years, I think that, unfortunately for Machine Translation, there has been a lot of animosity, a lot of competition between human translation and Machine Translation. I think that is the wrong paradigm. To me, it is not a question of Machine Translation versus human translation, it is a question of Machine Translation versus no translation. That is the problem that we are running into from the consumer world. We either do not get anything or we get Machine Translation. So, one of the big things that we are trying to do with Machine Translation is to be able to use it to gist through large amounts of information so that we can tell where we want to use this ever-dwindling, precious resource of human translators, where we really need their skills. Now at the same time, I think that Machine Translation has matured to the point where it is also an excellent tool for the translator, the linguists themselves. Because it is speeding them up, it is reducing the cost, and, by the way, it is also adding to the accuracy. You know, we talk about howlers in the Machine Translation world, but it is kind of interesting to see the howlers in the human translation world. We have had some cases where human translators have started to translate something for us and it is like they get a nostalgic trip back to the old country and all of a sudden, they are off to the Milky Way, translating something that is not even on the paper. So, I think it happens also in the human translation world and the nice thing about that computer is, it may be right, it may be wrong, but it is going to consistently be translated the same way. I think that is good because then the human translator can make some judgements, whether or not, in this particular instance, he ought to stay consistently with the same translation, or he ought to change it.

The other thing that we are going to talk about a little bit more is networks and the need for communications. You know, the fact is that we are having the globalisation of organisations. There is more contact going on because of things like the fall of the former Soviet Union, suddenly it is not a bi-polar world, it is a multi-polar world. The Internet has created just an amazing ability for people all over the world to communicate and Machine Translation is going to play a major role in networks from now on.

Let us talk about a few of the military applications. I do not want to go through an exhaustive list of them. Just to give you some idea of the ways that the Military uses communications is C4I Communications. We have put together a couple of programs called the Common Coalition Language System and the Text English Korean Machine Assisted Translation system. These are just a couple of examples of many that are being used with Coalition Forces in Korea so that the South Koreans and the Americans can talk to each other. So, it is adding to some of the rough linguistic skills and allows pilots to talk to other pilots and it allows some communication of different presentations to be made, some viewgraphs to be translated back and forth.

MT is being integrated with speech recognition. There is a Navy program called the Multilingual Interview System. Technically, it is not a Machine Translation system, but it has some aspects of Machine Translation. A non-linguist can go in with this system and there are a lot of different questions

and answers that are recorded, using a native speaker, and they can actually communicate to some native people. For example, if you are on a peace-keeping mission and there are some medical things that need to be done, they will ask them to answer in 'yes' or 'no' and they will go through a whole interview process. It is just fantastic. I saw a demonstration of it about a month ago.

One of the most fantastic developments, I call it another one of the great benchmarks of Machine Translation, has just occurred recently with Machine Translation integration with OCR in an Army system called the FALCON. We have set it up so it is going to be demonstrated here in the lobby for a couple of days. It is an excellent system. What it allows us to do is to take paper documents, feed them into a scanner, go through an OCR and come up with a translation. It is packaged in a lightweight thing. We developed a Serbo-Croatian Systran language system for this in six months and we have it in the field in Bosnia. There are six units that are helping with the peace-keeping mission out there. One of the key things that is really critical about FALCON is that a lot of the development that has been going on in Machine Translation, particularly within the Intelligence Community, has been there to support the intelligence analyst and so forth. This system is actually getting Machine Translation out to some of the military forces.

The other thing that I want to point out is, obviously, since the fall of the former Soviet Union, there is a lot of activity going on relative to NATO Partnership for Peace, there is a lot more interest in languages like Polish and Czech and so forth as NATO expands and the Partnership for Peace things occur. A NATO commander said in a recent article, 'If you cannot communicate, you cannot fight together. If you cannot communicate, you cannot do exercises together. If you cannot communicate, you cannot do peace-keeping missions together'. There are a lot of things that we are trying to do together that we have to be able to communicate. I think that Machine Translation, very soon if I have my way, is going to find itself in much greater use in NATO and in Partnership for Peace.

This is the FALCON system. It is a Pentium laptop computer that is integrated with a scanner, OCR and the Systran Machine Translation system. With this particular system, it used to take about 19 steps to be able to go from a paper document, feed it in and get a translation. They have automated part of that process and they have actually got it down to five steps so that even someone like myself can go along the function keys and press them. There are six of them in Bosnia. One of them went to Uzbekistan for a Partnership in Peace exercise that occurred last month. And they have some near-term plans for some upgrades to this. They are actually going to reduce this weight from 35 pounds down to 20. They are trying to get the cost down from about 18,000 to 8,000. They are probably going to get the number of steps down to either one F-key or two. In which case, I will be an expert in FALCON also. The next prototype is due out in March of 1998.

But they even have longer term plans in the Army. They are talking about things like wearable Machine Translation systems, hand-held Machine Translation systems. If you do not think that Machine Translation has a future, you have to be thinking about those kinds of things.

Let me jump now to Machine Translation in networks. Why networks? Networks are inherently a textual medium. And there has been an incredible growth; not just the Internet, there are local area networks in every company, government organisations, and there are intranets cropping up all over the place. Extranets are becoming extremely popular. There is a lot of textual information that is being made available to a lot of people. And if you stop and think about these intranets and extranets, you know if a company or if a government has an intranet they are talking within their own organisation and maybe they are all talking one language within their own organisation, but if they start to talk about extranets then they are starting to talk to suppliers, customers and other folks, other divisions that might be in a foreign country with a foreign language, and the ability to do things like e-mail and to be able to translate documents that are on the intranet among the extranet partners becomes extremely critical. There are major implications in these networks relative to culture, politics, economics, and Machine Translation, to me, has a tremendous opportunity. What I was very happy to see last night was I went next door to the demo room and it is obvious that a lot of the people that are in the business of supplying Machine Translation have already got the message on this network thing. Because there are a lot of interesting things going on in several companies. I was very pleased to see that.

Now, there is a thing called the Open Source information System. It is an unclassified, but for Official Use Only network within the Intelligence Community. There are about 110,000 users on it, 26 locations all over the globe and we have put Machine Translation on that network that is available to those users. There is also a thing called Intelink, which is a Secret and a Top Secret system and there are over 100,000 users on the Intelink systems in the United

States Government. They include people in the Intelligence Community, people within the policy-making community and people within the Military community. We have also put our Web Translators on those two networks. So we have instantly got exposure to a 110,000 people with Machine Translation. A lot of those 110,000 do not even know that the tool is up there. But already, we have gone from a few hundred translations a month to 5,000 a month on the networks. That is 60,000 a year. I will tell you right now that I believe that, within a year to two years, there will be a half a million to a million translations a year on these networks. Now these translations could be anywhere from a couple of sentences to entire documents. That is what I call exposure for Machine Translation. And if that does not get a lot of interest from the user community and help stimulate the resources and the further advancement of the technology, I do not know what will.

FILTER is another system. It is a program by the Department of Commerce, the National Technical Information Service, with SRA and Systran, in order to put a tool on to the Internet, where people can go in and use the SRA Namefinder system, go into foreign sites, look for the subject areas that they are interested in, and then, if they need to, they can translate the information. The first one is being done in Japanese. We are also replicating our Web Translators in other United States Government networks. We have already had interest from the Department of Energy, NASA and the FBI to make copies of these same Web Translator tools that we have on the Intelink networks in their environment. One of the unique features of the Web Translator version, on the classified systems there is no connectivity to the Internet, so everything on there is basically in English unless people input it themselves, but on the unclassified system there is a gateway to the Internet. So, one of the things that they did was to put in what we call the Web Translator. It allows you to literally just paste a URL of an HTML page that can be in a foreign language and it will go out, translate that page, bring back that Web page, holding all the graphics and the layout but substituting the English text in for the foreign text.

Let me give you a couple of examples of this. It is in ten languages on the system but here is a Spanish site and I am not sure what the HTML address is, but you can see that it is definitely Spanish. Then, if you press the button, thirty seconds later, it looks basically the same but is has substituted English for the text. Spanish

is one of the better languages, a lot of our Spanish system has been built around scientific and technical literature, not general news type information and some domain work needs to be done on this. But still, for someone like myself, who is such a linguistics expert, I do not know what I would do without something like this because I certainly could not understand the previous chart. Here is a Chinese site. This is a Chinese HTML page (*from the audience:"Japanese! Japanese!"*). I do not know the difference! This is a Japanese page. OK, press the button, 30 seconds later, there we go. So, the Japanese is pretty good. Again, it needs additional work but this is good enough for me to navigate around that foreign site and find out the kind of information that I am interested in. I can then maybe bring back the documents, run the specific ones that I am interested in. I might be happy with the raw Machine Translation, I might go to a linguist and say, 'Give me a little postediting on this', or 'Here, take this whole electronic file and give me a good human translation of it'.

Let me change gears here for a minute and talk to you a little bit more specifically about the National Air Intelligence Center Machine Translation program and also the FTD program. The things that we have been working on for the last six years, very hard, are the networks that you just saw, getting it on the Open Source Information System, getting it on Intelink, allowing other United States Government organizations to replicate the system. We have worked on general enhancements and research. In the general enhancements area, one of the systems that we developed was the Chinese system. There was a tremendous interest in Chinese, particularly in the Pacific Rim. So, we put together, with FIDUL and with the Community Open Source Program Office a package, which was a PC, a Pentium computer, the Systran Machine Translation system, the ECI OCR system. The algorithms were developed in Beijing, but then they were marketed by ECI world-wide. We integrated this entire package together, we deployed it to the Pacific Command and to Embassies throughout the Pacific Rim and to the Foreign Broadcast Information Service.

We have also been working on new language pairs. You know, if you go back to 1980, we had Russian to English. And then we got German, French, Italian, Portuguese, but now, we have been adding Serbo-Croatian, Chinese, Korean, Cantonese, Ukrainian, and I will talk to you in a minute about some additional ones. We have also tried to improve our dictionary domains. I have heard people say that transfer systems have sort of reached their limit and we need some kind of brand new exotic technology in Machine Translation in order to improve these translations. I heard them saying that five or six years ago when I first got involved in Machine Translation. And then I saw us make some significant investments in domain development, in the dictionary development, in the linguistics and, let me tell you, we have not pressed transfer systems to their ultimate limit. There is a lot more that can be done to improve those systems. We have been making some investments in that area.

In terms of research, we have a product that we helped sponsor the research on called the InfoRaptor, which uses the parsing and the parts of speech to actually go out and retrieve information, so that you can improve your retrieval, in either the source or the target language. Because if you know the parts of speech, you can eliminate some types of documents you are not interested in. I will be talking about the Digital Library Information Input Processing System, or what we call DILIPS, in just a moment.

I said there were three key events that occurred, in my opinion. In 1993 COSPO paid for moving those systems off the mainframe, the Web Translator came around in 1995 and, all of a sudden, we are interactive, we are exposed to tens of thousands of people, and then the FALCON system in 1997. All that stuff has resulted in a whole variety of offerings that we now have for the systems that we have developed. The systems that we have developed for the United States Government by the Air Force are available as government off-the-shelf software to all United States Government agencies. We have them in stand-alone UNIX and Windows versions, network versions in Windows NT and UNIX. In UNIX we have SUN OS, DEC Alpha, HP, and we will soon have Solaris.

We can put in information in three different ways. We can keystroke it in and there are all kinds of things that help you with the transliteration and the keystroking of the information. You can attach an electronic file in a foreign language and put that into the system. Or, as I mentioned, you can just put in a URL.

The languages that we are interested in and have been developing are from the foreign language to English. I know there are a lot of people here whose focus is English to the foreign language because of commercial reasons. The operational systems that we have are Russian, German, French, Spanish, Italian and Japanese. We call them operational when they have achieved what we call 80% accuracy on our rating scale, which is based on a test corpus of about four hundred sentences and about ten different evaluations that we do on the performance of the Machine Translation system. We have two prototype Portuguese

and Chinese and the Portuguese is very close to moving into the operational category. The Chinese probably should be a pre-prototype but we had a huge dictionary that we added to it and it really performs a whole lot better than a system with only the few years of development in it. In pre-prototype there is Korean, Serbo-Croatian, Ukrainian and Cantonese. Normally, we would not deploy things that were not in the operational category but, because of the Pacific Rim interest in the Chinese and the Army's interest in the Serbo-Croatian, they are actually out on the street. That causes a little bit of a problem because there are more howlers out there. People are complaining, to some extent about, 'Boy, this is a terrible translation'. You know, we have had all of six months' investment in developing the Serbo-Croatian system! We have 30 years of developing the Russian system. So, it is a crude system but we have also had a lot of positive feedback. People have said, 'I'd be lost without this system'. One of the things that we have been able to do, because the Serbo-Croatian language is related to Russian, we have been able to develop a trunk parser for Slavic languages, which is allowing us to develop new systems. If there is a trunk language that we have already developed, we can develop the new languages in about half the time and half the cost.

Let me give you a couple of examples of this. People talk about, 'How good is Machine Translation?'. Well, I have talked to you about the NAIC quality control where there are 400 sentences and ten categories. ARPA did a study in 1994 and 1995. Dow Corning did a study in 1996. Buckman Labs did a study. All those studies are really neat. But do you know what? To me, quality is in the eye of the beholder. And today, you are the beholders, so let me show you a little bit about the quality of these systems.

The first is Spanish. There is our source text. Now, I do not have any linguistic skills, as I mentioned, but I did take high school Latin. Everybody in those Catholic schools had to take high school Latin. I can see a few things in here, like it might be 'Kashmir'. There is some kind of celebration going on, it might have something to do with the independence of India. But there were demonstrations and there was some kind of hostility in New Delhi. The violence might have caused 40 murders in a region in the northern part of India. Not too bad, for a guy that does not know anything. Let us see how good I was. Here is the human translation. 'Last Sunday the Muslim population in the region of Kashmir boycotted India's independence celebrations in an attempt to demonstrate its hostility to the central government in New Delhi. This happened the day after a wave of violence caused approximately 40 deaths in the region in the north of India'. I was pretty good! Let us look at what the computer did, whether we get a howler out of this. 'The Muslim population of Kashmir boycotted Sunday the celebrations of India's independence to demonstrate its hostility to the central government of New Delhi the day after a violence wave that caused about forty dead ones in the region of north of India'.

Now, that is not publication quality. You would have to do a little postediting on it. But can you tell me that we did not get the idea from that thing?

Let us try another one. They tell me that German and English are related. It must be a distant cousin. I think we have a problem here. Something is 'complex'. I can tell that it is 'representative' of something. It is 'exponentially related', I do not know if it is up or down. Let us see what the human translator did. 'Nearly all the interesting problems are so complex they cannot be solved by random trial. The number of decisions grows exponentially from the first intersection of the first decision point exponentially.' Let us see what the computer did. 'Almost all the interesting problems are so complex that one cannot solve them by arbitrary trying. The number of decisions rises from the first crossing of the first decision point exponentially.' My gosh! You know what? In my opinion, the computer did better than the human did. That one was really good.

Let us try another one. Let me see what I can tell about this one with my rudimentary linguistic skills. I am in a lot of trouble. I am not going to be able to tell anything about this one. Obviously, this is Cyrillic, Russian. Let us see what the human translator and what the Machine Translation says. 'On the whole, a system has been created and needs to be kept for which the construction of large ships is not necessary.' What did the computer do? 'The system, in essence, is created and it must be preserved for which the construction of large vessels is not required.' Again, that looks to me to be pretty good.

Now let me give you the last example. Talk about really being in trouble. This time I think it is Chinese. It is easy to understand under these circumstances why a non-linguist like myself thinks the bottle is half-full. Let us look and see what the translations were. Let us bear in mind that those first three, Spanish, German and Russian, are fairly mature systems. This one is a prototype system but it says, 'Chapter eight is an in-depth discussion of relational database theory, which not only helps provide a deeper understanding of relational methods, but also lays a solid theoretical foundation for future database design'. That translator was pretty good. How about the machine? 'Eighth

chapter thoroughly has discussed connection database theory, this not only is helpful as to deepen understanding which to connects method, but also also has established strong rationale for the next database design'. It is getting a little bit rough. But I tell you what. Ambassador Lyn Hanson, who is one of the senior folks in the Intelligence Community and happened to have been Dale Bostad's boss years ago when he was a Major in our translation shop, and who also speaks fluent Chinese, said, 'Hey, you know, if I am an expert in the particular field that we are translating in, I can get the gist of what this article is about from even this quality of Machine Translation'.

Ladies and gentlemen I think in the Machine Translation area we have arrived and let me tell you a couple of things we are going to do with some of this technology.

Digital Library Information Input Processing System, or DILIPS. We have a large database of a lot of collections of scientific information that was originally in foreign text. It is very expensive to translate it, to manually index it, and so forth, so we decided to automate the process a few years ago. We are going to have an input system that allows us to take hardcopy or digital information, English or foreign text, classified or unclassified. And we are going to process it. As necessary, if it is hard copy, we are going to scan and OCR it. If it is in a foreign language, we are going to machine-translate it. We are going to do all that, automate it, and then we are going to SGML-tag it and filter it and we are going to put it into a thing we call Information Space. Information Space is going to be the unstructured translated text, an imaged version of the original text, and a parametric relational database of the parametric information like the bibliographics and some of the deep indexing. We expect to run between one to five million documents through this system. We are then going to interface it with a variety of analytic tools. One of those is going to be what we call the MINS system, Multi-Information Notification System. Every one of our users will have a profile of their interest and, when an accession is made to the database that is in their interest area, it will automatically fill their inbasket. We will also use those profilers to go out and data -mine different sources within intranets or the Internet. We will have an interface to this, which is based on the Excalibur retrieval engine, augmented by visualisation tools, such as the Calvin system developed by Calspan, and we will use tools like Pathfinder from the Army and Information Dominator from the United States Air Force.

I know that this is a big mouthful that I have given you here. But I just want to tell you that this system and this type of an application, we are going to insert Machine Translationinto this process. These are the kinds of things that are going on.

Where is our MT program going? I believe that there are a lot of enhancements, a lot of tools that have to be done. We recently created a document. It lists all the strengths and all the weaknesses of our linguistics and our dictionaries in every one of the systems that we have and it lists the different types of projects that could be done to enhance and correct those things. We are going to develop post-editing modules. We are going to continue working on the main dictionaries. Where users come to us with both requirements and resources, we will develop those special dictionaries just for them. We also have become interested in the English to foreign systems so we are going to license some of those because of the opportunities for two-way communications on things like the Internet. We are going to continue working on national language retrieval systems, OCRs, and we now have a fair amount of interest in the Department of Defense. We have been asked to put together an initiative for some additional money to develop some more languages. Those additional languages that we intend to work on next are Polish, Czech, Hungarian, Swedish, Dutch, Norwegian, Danish, Greek, Macedonian, Slovenian, Urdu and Vietnamese. That will about double the number of languages that we have available now. Unfortunately, that is going to be a 1999 or 2000 initiative. Hopefully, it will make it through the funding process.

In the meantime, we do have some problems because some of our resources run out after financial year 1997. The good news is that the trunk-parser approach means that in a lot of these languages, for example Polish, Czech, Macedonian and Slovenian, we will be able to use the Slavic parser and develop the thing for about half the cost.

That is a little bit about the NAIC program. Let me use the last few minutes to talk about some general future directions. I think you can count on it, and it looks like everybody over there in the commercial world has got the news, that there is going to be a tremendous proliferation of MT use on network applications and associated tools. As I said, not only the Internet itself, but I think this whole concept of intranets, interacting as extranets with other countries, with other commercial partners and so forth, is going to be a massive growth area. I think you are going to see a lot of integration of the tools into things like DILIPS, FALCON and CYBERTRANS at the National Security Agency, and major language pairs being expanded greatly. We need to do a lot more marketing of Machine

Translation because we need to get expanded use. We now have something, I think, in Machine Translation that is very viable. What we need to have is a much greater clientele of users that is going to buy the products, that is going to demand continuing advances in Machine Translation. I think that some of our research ought to focus on technology insertion. Some of the people in this room may not agree with me, but I do not think there is a Rosetta Stone out there that is suddenly going to give you an algorithm that is going to translate everything and do it wonderfully. I think transfer systems are going to continue to be our baseline. I think, if there are going to be breakthroughs, it is probably going to be in the computational world.

You know, the PC was built in 1980. In 1987 I had a hot computer at home. It was called the IBM 286. By 1990, it was a 386, by 1991 it was a 486. Then there was a Pentium in the middle '90s, which was the hottest thing going, literally hot, they were worried about whether or not they could cool them down. People were starting to say, 'Boy, are we going to get anything faster than a 100 MHz Pentium?' You cannot even buy chips for a 100 MHz Pentium anymore. 133 and 166 MHz are the baseline systems and I will bet you by early next year, a 200 MHz Pentium is going to be the low-end PC that is out there. Now, there is some real good news for Machine Translation there because, with all that speed and all that storage and so forth, suddenly we can start talking about inserting some advanced technology that universities and the corporations can do into our basic transfer systems. We can start augmenting them with statistical modules, and n-grams, and text-meaning algorithms that will improve readability of the systems, and automated information tools. So, I think that is where we ought to be heading and I think that there is a lot that we can do in that area.

Just to give you an idea of one of the government organisations and some of the things that they are putting into their strategic plan, the Community Open Source Program Office says this about Machine Translation: 'We want to see cross-linguistic browsers and foreign language browsers. MT and machine-assisted translation for all major languages and critical low density languages. We want dictionary domain growth in policy, in military, in education, law enforcement and so forth. Automatic SGML-tagging for enhanced retrieval and tools, lots of tools for the analyst to do cluster analysis, data-mining and summarisation.'

Well, we finally reached that slide that everybody has been waiting for, my last slide. There is a saying in the Military, 'Co-operate and graduate'. You know what is really interesting, in the United States education system we always talk about this rugged individualism, but you get into the Military and people realise that they have to work together in order to make things happen. So, in the Military, in their education programs, they talk about co-operating and graduating. We need to do a lot more of that. We need to do a lot more of that within government, between industry and government, between foreign countries. We need expanded use, as I mentioned before. Some of these things that I have seen next door are just fantastic. The Machine Translation associations of America, Europe, the Far East, the International Association, should be pressing very hard for the expanded use of Machine Translation. Because as more people buy the products of these fine companies that are out there providing these things, they are going to have the profitability and the resources to further advance Machine Translation technology. In the governments we need to work on our requirements and, somehow, in the United States Government we have to figure out a way to take something that affects everything in the United States Government and come up with a program that is responsive to it. And I am going to take that on as a personal challenge over the next few years. I think we need a balanced program. We need to enhance some of the current systems that are out there. We need to have a strong R&D program. The R&D program should look at new concepts, but it should also look at technology insertion into the current programs. In the last six years we have made tremendous advances, as far as I am concerned, in the National Air Intelligence Center program. We have put probably twice as much money in the last six years than we did in the first 20, 25 years into it. We see the results of it. But you know something that troubles me, is, when I look back to 1966 and the ALPAC Report, one can sit there and, if they gave them the benefit of the doubt, they could have said, 'Well, they had an intellectual difference with us relative to the viability of Machine Translation' and so they said, 'Don't spend money on it.' You know what bothers me about 1998? We know it is good. It is the bean-counters that are saying, 'We don't have the money for it.' That is ridiculous. We should somehow, in a thirty million dollar budget, be able to come up with ten million dollars for Machine Translation research. I cannot believe that we are doing this. We know better. At least in 1966 someone could have said, 'Maybe they didn't know better'. We know better and we are not making the investment. I will tell you this, ladies and gentlemen. If I could get ten million dollars a year for Machine Translation, I would set MT on its ear. Thank you. ■