

MURIEL VASCONCELLOS: MT'S GREAT SURVIVOR

By Andrew Joscelyne

For you, the mid-1950s might conjure up a cozy technoscape of scratchy Buddy Holly 78s and Brylcreem. But for misty-eyed veterans of the American machine translation scene, it was a golden age, when government agencies were doling out dollars aplenty to researchers at universities throughout the land.

Never-never-land? Ask Muriel Vasconcellos, now head of Machine Translation at the Pan American Health Organization (PAHO). She was there. In 1956, she personally typed the proposal from Georgetown University to the CIA, which led to a million-dollar grant for the development of a Russian-English MT system.

Officially, Vasconcellos was working in an administrative capacity. But her natural curiosity for what was going on brought her close to some of the founding fathers of the discipline, and she's been carrying on the good work ever since.

LIFE BEFORE ALPAC – AND AFTER

"There were lots of ideas around at the time," she remembers. "There were at least three separate Russian-to-English projects going on while I was there. Apart from the main General Analysis Technique (or GAT) research, Paul Garvin was working on his FULCRUM analysis project – plus there was another team working on what was called code-matching techniques.

"For five years, I was assistant to Leo Dostert, who was the architect of the simultaneous interpretation system used at the Nuremburg trials. At Georgetown, he would brainstorm ideas every Friday morning and then get people to work on a project. Anybody could work on anything they wanted to, so there were people doing stuff on Chinese, Slavic languages, and French.

"Dostert was very much attuned to using corpora – real texts – so he started up an English-to-French MT system



"I'D LIKE TO BUILD UP THE MT AWARENESS OF INDEPENDENT TRANSLATORS, SO THAT REAL MT WILL BECOME A GENUINE OPTION FOR THEM. THERE IS NO QUESTION THAT THE MAJOR SYSTEMS OVER THE NEXT FIVE YEARS WILL BE PC-BASED, AND IT IS MANAGEMENT RATHER THAN TECHNICAL QUESTIONS THAT ARE GOING TO BECOME IMPORTANT"

based on a twin-text corpus of biochemistry data. Above all, though, Leo was a great publicist for the MT cause.

"We were a regular nursery for computational linguists. There was Peter Toma, with the four-language translation system he had programmed at night at the company he worked for in La Jolla. He came to Georgetown to work on morphological analysis and stayed for two years. Four years after he'd left, the first version of what came to be Systran came out."

Then disaster struck. Muriel Vasconcellos was still at Georgetown in 1966 when the ALPAC blow fell. This notorious "no future" evaluation of the prospects for MT suddenly spelled the end of public funding of GAT.

In fact, the GAT system had become operational in 1964 at Atomic Energy Commission facilities both in the US and Italy, and it continued to be used over the next twenty-odd years to provide rapid raw translation of Russian physics documents. However, any idea of further development was deep-sixed, the Georgetown research team broken up, and dust settled on the piles of punch cards.

Vasconcellos, though, struck lucky. She soon found herself working for the only non-military public organization in the Americas still prepared to put money into MT: Washington-based PAHO, the Pan American Health Organization.

In 1975, PAHO asked Tabor Corp., the research company that now markets Globalink translation software, to perform a feasibility study on how to reduce the vast costs of producing all its public-health-related documentation in English and Spanish. The result was a project, begun in 1976, to develop an MT system using PAHO's onsite IBM 360. And a rocky road it proved to be.

NOT JUST RAW BUT BLEEDING

"I was the PAHO staff member," recalls Vasconcellos, "with the glorious title of 'terminologist and coordinator' of the project. At first, we used a team of consultants who had already worked on the old Georgetown project, including Allen Tucker and Bedrich Chapulka.

"Bedrich had been a dictionary coder, and he arrived with an image of the dictionary record in his memory. Since he had never worked on the algorithm, he set himself to work backwards, thinking up the rules that would be needed to access the information coded in the dictionary."

Three years later, a first Spanish-English algorithm was in place. Written in PL/1 and accessing the 48,000-word dictionary, it turned Spanish input into raw English text. Correction: not just raw but *bleeding*, apparently.

"When it came to be evaluated, this system was frankly a dog's dinner. We were about to give up the ghost and see if something like Systran could be used. But developing new dictionaries would have been as expensive as continuing with our own work. So I asked Yerrick Wilks, who had been working on MT at the University of New Mexico, what he thought. He

figured it could be fixed – and he was the first to really encourage us to use what we had."

So in an attempt to beef up SPANAM, as the system was called, in 1979, PAHO hired Marjorie Léon, a computational linguist, to work fulltime on development. For Muriel Vasconcellos, Marjorie Léon is the "real architect" of SPANAM.

"Marjorie is a brilliant computational linguist. She managed to streamline the whole thing. For example, we had been left with a full-form dictionary – which meant that any conjugated verb form in Spanish was displayed as unknown, entailing endless dictionary updating.

"Within a month of arriving, Marjorie had written a morphology module to process the 52 forms of the source verbs. What I learned from her was that you could take something pretty basic and use what you had without wiping it out and starting from scratch each time."

Since then, SPANAM and its sister program ENGSPAN, which translates in the opposite direction (and was developed with a two-year grant from the US Agency for International Development in 1983), have gone from strength to strength, offering the translation of many different text types – a "try-anything" system if ever there was one.

"People were continually telling me that MT couldn't really work. So I accepted the challenge and decided to find out just what we could do over a given period." The result proved that 80% of the English documentation could be processed by the machine and 60% of the Spanish, using scanned or e-mailed input text.

"Less than 1% of the words in the documents could not be found, out of one million words per year. Where previously we were just one component of the PAHO Language Services department, we've now proved we can use machines to do the main job. The tail is beginning to wag the dog!"

COLLABORATION A MUST

The encouraging results of the eight years that Vasconcellos has devoted to promoting MT at PAHO have borne other fruit. There is, for example, a project in the air to work with the 13 centers of the pan-American Consultative Group for Agricultural Research (CGAR) in a consortium of public sector agencies interested in exchanging public health information.

One plan is to co-develop an English-to-Portuguese module. Another is to get funding to add an English-to-French module to the PAHO system. Muriel Vasconcellos smiles knowingly: "I know the world needs another French-to-English system like it needs a hole in the head – but ours might just be the best around."

Vasconcellos's wide experience of the ups and downs of MT-land has taught her how necessary it is both to generate solidarity within the MT community and educate laypeople on the outside.

She is particularly vehement about the need to combine resources via internationally accepted standards for docu-

ment structure and dictionary entries. "Right now," she says, looking back on ALPAC and the struggle to get SPANAM off the ground, "MT people need mutual support. We should stick together and try to exchange much of the effort that's going into building machine dictionaries around the world.

"It's so labour-intensive that it's ridiculous to have different MT vendors building up what amount to very similar dictionaries. Look at the way Japanese MT teams get together to work out mutual problems. We're simply going to have to build some basis for understanding beyond mere gossip about what 'the others' are doing.

"A combined effort on lexical data could still leave each system free to do its own thing at the parsing and semantic processing levels."

ALPAC AFTERBURN

Apart from her work with PAHO, Muriel Vasconcellos has returned to her old stamping ground Georgetown University, giving courses each year on Machine Aids to Translators at the Department of Interpretation and Translation.

She also chairs the Committee on Translation and Computers in the American Translators Association.

"I'd like to build up the MT awareness of independent translators, so that real MT will become a genuine option for them. There is no question that the major systems over the next five years will be PC-based, and it is management rather than technical questions that are going to become important.

"If we're going to have a bunch of \$500 systems on the market, what'll happen to the real systems? We're going to have to educate people to understand the difference between them."

In spite of her successes at PAHO and the radical improvements that have affected user interfaces and the general technological environment for machine translation over the years, Muriel is still haunted by the stigma that ALPAC left on the MT community.

"MT is not necessarily getting a better press now than it did just after ALPAC," she explains. "There's still not a cent of public money around in the States for MT. When people at a cocktail party ask you what you do, and you say you work in MT, they invariably come out with 'Oh I know all about that – they did a study about it, didn't they, and found out it can't work?'"

And she's sick and tired of hearing the same old jokes about dumb machine-made translations of the "Spirit is willing, flesh is weak / the vodka is alright but the meat is off" variety, flashily updated to the eighties, with the output coming from some futuristic automatic telephone interpreter.

Muriel Vasconcellos clearly feels there is still a long way to go. "Trying to get rid of ALPAC's fallout," she concludes, "is like pushing jelly up a hill." Maybe she should show the Cassandras just how ENGSPAN can translate that evocative little simile.