After attacks, company gains in translation

By TOM WALSH

As a former CIA analyst, Joel Ross knows all too well the frustration Americans felt after Sept. 11, 2001, when they learned U.S. intelligence agencies had reams of unprocessed data that might have helped stop the terrorists' attacks.

Ross left the Central Intelligence Agency last year to work for Basis Technology Inc., of Cambridge, which recently released a language analyzer it believes can help intelligence officers sift through the data.

"There is not a problem (in the intelligence community) in collecting the data — that problem was solved long ago," Ross said. "There's just more data than translators and analysts can get to quickly."

Government intelligence experts agree.

"One of the most pressing issues facing the intelligence community today is the need to quickly and accurately identify, analyze and extract information in foreign languages and scripts," said Glenn Nordin, assistant director of intelligence policy in the Department of Defense.

Part of the difficulty working in the Middle East comes from the vast differences between Arabic and English When peo ple translate from one to the other, mistakes are often made, or there is no definitive English way to spell a particular Arabic word.

To illustrate the problem, Ross said, you just need to go online and do an English-language search of the term "Libyan strongman."

"When we did that, in the first 100 hits, 35 came back with a different spelling of (Libyan leader) Moammar Gadhafi," said Steve Cohen, vice president of product development at Basis.

That means that if you search a database — or if a CIA or FBI analyst searches their files — for a particular English spelling of Gadhafi, many alternate spellings will be missed.

In Arabic, though, there is only spelling.

"If there is someone entering our country and his passport or papers have his name a certain way — say 'bin Ladin' instead of 'bin Laden' — it wouldn't match against a list of people to stop," Ross said. "But if you had our phonic software, it would match."

Other problems arise from the basic structure of Arabic.

"If you look up the word for 'airplane' on a normal search engine, for example, it will miss an enormous amount," Ross said.

The reason is that in Arabic, entire phrases are appended to the root word. For example, "in the airplane" or "from the airplane" would be unique words in Arabic. And though they contain the word "airplane," a normal search engine would not find it.

The Arab Language Analyzer, however, roots it out, Cohen said.

Basis' software attempts to overcome the problems by using what's called a "phonics fuzzy search."

Instead of looking only for an exact match, the program has a phonic fingerprint of what the foreign words sound like in English, and returns all those hits.

When they applied it to the Gadhafi search, Ross said, they picked up nearly every English variation of the spelling.

The program would be useful, for example, when agents come across notebooks or computers with Arabic names they want to check against the names of known terrorists, such as in the recent arrest in Pakistan of Khalid Shaikh Mohammed, the suspected al-Qaeda mastermind.

In such a situation, the field agents often transcribe — or have someone else transcribe — the Arabic into English, which is sent back to the United States for analysis.

Much of that information is run through computers looking for hits on various names of people and places.

The vagaries of translation, however, can lead to misses, said Ross, who spent 21 years in the CIA.

"This is not just software and technical stuff," Ross said, "but something that could save lives if used properly."

24, 2003