Vaulting the Language Barrier

Computers are helping to search texts and data now shrouded in linguistic differences

By JANET RALOFF

Marjorie Hlava can't read Russian, but that doesn't stop her from learning the contents of a document printed in the Cyrillic alphabet. She simply places each page under the cover of the flatbed scanner in her Albuquerque office, presses a button, and waits as her computer displays an English-language version.

Using only English, she can also search Russian databases, such as files of published scientific reports. She types in the key words or phrases that describe her interests, then lets a series of computer programs take over. After converting her request into Russian, they sift through data files for references to documents that seem to match, convert those matches back into English, and display them on her computer.

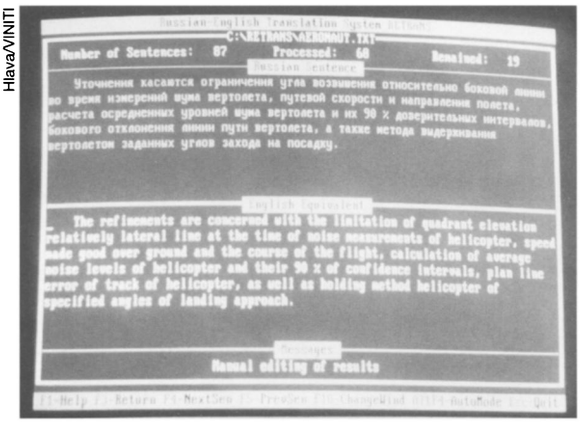
More than once she has even conversed via her laptop computer—on a plane, for instance—with Russians who know no English. She types her side of the dialogue in English, which the computer converts into a Russian display. The other party types his or her responses in Russian, which the computer translates for Hlava. They can chat for hours that way, provided they restrict their words and phrases to those in the thesauruses, or set lists of words, on her machine.

That isn't too hard, Hlava notes, since the Russian-to-English portion currently contains some 750,000 words and phrases and the English-to-Russian one nearly 600,000.

Most of the software programs that allow fairly inexpensive, off-the-shelf computer hardware to translate Russian are preliminary versions being developed by Gerold G. Belonogov and Boris A. Kuznetsov at VINITI, the All-Russian Institute for Scientific and Technical Information in Moscow. Hlava's company, Access Innovations, helped channel some U.S. government financing into the creation of those systems.

As the Internet has been demonstrating over the past few years, "we now have access to an enormous amount of information that didn't used to be available," notes Douglas W. Oard of the University of Maryland in College Park. "But it's only accessible to those who speak the language. And as the World Wide Web's name indicates, not everything on the Internet is in English."

Hlava/VINITI



Russian scientists are developing translation software, such as Retrans, that does not require special computer hardware.

Because users seldom pay for data they find on the Web, there is little incentive for those who post the information to invest in expensive, time-consuming multilanguage translations or indexing. What a user needs to make full and efficient use of a foreign database or the Internet, Oard explains, is a system that translates among languages, searches effectively for answers to a user's query or stated interests, and then ranks any matches by the likelihood of their satisfying a particular user's needs.

For many persons interested in focused areas of science or engineering—such as the microwave heating of plasmas or drugs to treat cancer patients—"the things that Marjorie Hlava [and her VINITI colleagues] do are just as good as you would like," observes Oard. "The limitation is that humans can find them difficult to use"; that is, they need to be trained in effective search strategies.

He and a host of others are working to make foreign data and files easily accessible to an even broader audience, one with little training in data searches. Unfortunately, he says, "we're only about half as good as you'd like at doing this. And getting halfway turns out to have been rather easy." It's the second half that will prove costly in both time and money, he maintains.

The payoff could prove substantial, he and Hlava agree. Such efforts could uncloak a world of research and data for people who don't speak a foreign language.

Today, computer technologies are being developed to translate a wide range of mother tongues. At the behest of the European Parliament, for instance, several ambitious programs are working to make documents prepared in English or French intelligible to those who read any of the other nine official languages of the European Union. Even more challenging projects around the world seek to pair English with languages written in non-Roman characters—such as Japanese, Chinese, Greek, Arabic, Russian, Korean, and Vietnamese.

Few of these efforts are designed to provide full machine translation of the documents; rather, their aim is a more limited rendering of some important aspects—such as titles, key words, or abstracts.

Indeed, this may be sufficient if the goal is merely to identify a few particularly valuable documents that a user might then choose to have translated in full, Oard observes. The projects could also help electronic browsers identify more circumscribed information, such as images posted on the Internet with captions in a foreign language, names and affiliations of foreign scientists who have conducted research on a topic of interest, or newly coined foreign terms or short quotations in a text.

Even limited cross-language identification and retrieval of electronically stored text represents a tall order, Oard notes.

For instance, even within a single lan-

guage, commercial database searching remains a fairly unscientific, "seat-of-the-pants thing," observes Richard S. Marcus, an information scientist at the Massachusetts Institute of Technology. What's not well recognized, he says, is that unless someone is an expert in searching or has the services of a good librarian, "you typically are able to retrieve only about 5 percent of the relevant documents available."

By employing certain computer techniques that he says are available only on experimental systems, "you can bring the comprehensiveness of a search as close to 100 percent as you like." With several interactions, sophisticated programs can prompt a user to find the most effective words for a query. Marcus maintains that this extra effort "can make all the difference between getting almost nothing and getting everything you want."

Before computers were in wide use, librarians indexed documents with a few key words—the ones that appeared in a card catalog. Such limited indexing "is not very good for detailed analysis of articles and documents," Marcus says, "because a few terms won't cover all of their information." Moreover, unless the wording of an indexed portion of some text—often the title or abstract—is restricted to terms in a thesaurus, an indexer might employ words that a later searcher wouldn't think to use.

With computers, "you can now index all of the words in a document" for full-text querying, Marcus notes. Yet even this does not always prove satisfactory. If an author used the word "Cessna" in his text and a searcher attempted to retrieve it by asking for references to small planes, even a full-text search would miss what conceptually should have been a valid match.

"So our research over the past 20 years has been to make key-word use smarter" by getting the computer to suggest synonyms, Marcus says. Not only might it point out that a Cessna is a type of small plane, it might also ask whether it should expand the ongoing search to include other small planes, perhaps helicopters—even dirigibles.

Alternatively, the computer may attempt to narrow an overly broad search by soliciting feedback on its first few matches. The computer can then look for a pattern in what was rejected or ask the user why certain choices were rejected, then refine subsequent searches based on the response.

British computer scientist Steven Pollitt of the University of Huddersfield's Centre for Database Access Research is taking a similar tack. His computer-aided searches ask the user what terms he or she would like to begin with and use them as a departure for identifying related search terms—some broader, some narrower in focus.

If a searcher typed in Alzheimer's disease, for example, the computer would flash a list of related terms, such as Alzheimer's syndrome and Alzheimer fibrillary lesion. A number next to each term shows how many documents match it.

The computer can also search simultaneously for texts fitting additional categories—such as a country (where clinical trials may have occurred), drugs, or other treatments (such as acupuncture)—and count or display all texts that match the combination.

The key to making this approach work is a comprehensive list of index terms that have been organized into hierarchies, Pollitt explains. Degenerative disease, for instance, would contain a file of terms for Alzheimer's and other chronic illnesses. Choosing Alzheimer's would allow the computer to suggest broader terms, such as degenerative disease, or narrower ones.

For searching to work effectively, the developers of a database must have indexed all texts using an agreed-upon vocabulary—and the more specific the vocabulary, the better.

The European Parliament has a list of 6,000 terms, known as EUROVOC, to index all subjects in its documents, from politics and law to science. Only a few dozen of these EUROVOC terms deal with medicine. In contrast, the National Library of Medicine has compiled a working list of more than 17,000 words for indexing articles cited in its MEDLINE database.

Searching success also improves, Pollitt notes, when each starting thesaurus is tailored to the vocabulary of a particular field, such as medicine or physics. This will limit confusion among terms common to both but having quite different meanings—such as plasma. To physicists, it's an ionized gas, whereas to biochemists it's blood minus its cellular components.

Belonogov, who is a linguist, has embedded 21 such thematically organized dictionaries (covering such subjects as ecology, geophysics, and foreign trade) within his thesauruses. To limit confusion further, the thesauruses treat as a single term many commonly used phrases up to 13 words long. In fact, about 75 percent of the English entries involve word combos, such as "bottom line," "ballistic missile," or "might be interested in."

When they surveyed the field last year, Oard and Maryland colleague Bonnie J. Dorr found few commercial systems that ranked potential matches. So if 20,000 potential matches are identified, a user must sift through them all to find the few that might be valuable.

Though the VINITI browser does rank its responses, "the drawback is that those responses are in Cyrillic," Hlava says. Nonetheless, it can prove useful when coupled to VINITI's translator programs. Together, the pair can search and retrieve documents from Russia's scientific holdings, which include not only Russian documents but also those published by Russia's trading partners, such as the former Soviet republics, North Korea, Syria, Iran, and Iraq.

MIT's experimental system attempts to rank matched terms on the basis of how they were used or where they appeared. For instance, Marcus says, "we have demonstrated that the title words are most important." So if a queried term appears there, the document will be ranked higher than another in which the same term is buried in the text.

Pollitt has tested his searching system on a database of 600,000 medical citations written in a host of European languages. He has also tested it by querying and retrieving citations—in English or Japanese—from INSPEC, a British bibliographic database covering texts on physics, electronics, and computing. He says the system can now be developed commercially.

Similarly, Marcus believes the system his team has developed is ready for commercialization.

Though VINITI's systems are still under development, working versions are available from the institute in Moscow and from Hlava. However, Hlava notes, money to refine them has all but dried up. The software programs that she marries into working systems still have a way to go before they offer "transparent" translation capabilities to both English and Russian readers, she says.

"It breaks my heart," she told SCIENCE NEWS, "that we can't get these technologies off the ground." Hlava says $20,000 would enable the VINITI team to develop a version of the translation and searching programs that would be compatible with Microsoft Windows, the primary organizing software on desktop computers today. The Moscow researchers have no money to invest in it, however: Not only are they working without pay, they don't have money to heat their offices this winter.

Indeed, most of these programs suffer from a paucity of both financing and visibility. Oard hopes to counter the latter through a symposium he's organizing under the auspices of the American Association for Artificial Intelligence. He plans this month not only to showcase what seems to work reasonably well today but also to highlight where future challenges lie.

Among the challenges, he says, are programs to revise thesauruses automatically as languages grow and change, to identify words in languages like Chinese and Vietnamese, which do not put spaces between words, and to insert verbs in languages, such as Arabic, that frequently use nouns in place of verbs. □