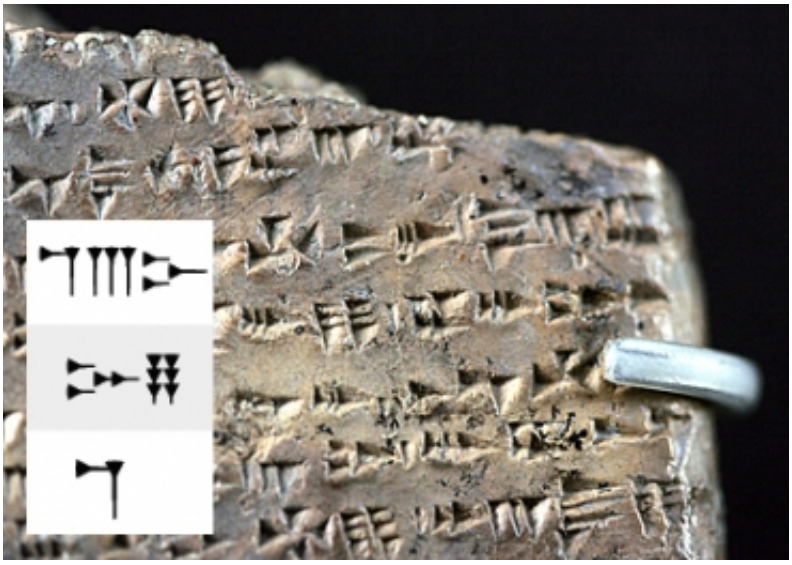
Computer automatically deciphers ancient language

June 30 2010, by Larry Hardesty



An incidental challenge in developing a computer system that could decipher Ugaritic (inscribed on tablet) was developing a way to digitally render Ugaritic symbols (inset).

In his 2002 book Lost Languages, Andrew Robinson, then the literary editor of the London Times' higher-education supplement, declared that "successful archaeological decipherment has turned out to require a synthesis of logic and intuition … that computers do not (and presumably cannot) possess."

Regina Barzilay, an associate professor in MIT's Computer Science and

Artificial Intelligence Lab, Ben Snyder, a grad student in her lab, and the University of Southern California's Kevin Knight took that claim personally. At the Annual Meeting of the Association for Computational Linguistics in Sweden next month, they will present a paper on a new computer system that, in a matter of hours, deciphered much of the ancient Semitic language Ugaritic. In addition to helping archeologists decipher the eight or so ancient languages that have so far resisted their efforts, the work could also help expand the number of languages that automated translation systems like Google Translate can handle.

To duplicate the "intuition" that Robinson believed would elude computers, the researchers' software makes several assumptions. The first is that the language being deciphered is closely related to some other language: In the case of Ugaritic, the researchers chose Hebrew. The next is that there's a systematic way to map the alphabet of one language on to the alphabet of the other, and that correlated symbols will occur with similar frequencies in the two languages.

The system makes a similar assumption at the level of the word: The languages should have at least some cognates, or words with shared roots, like *main* and *mano* in French and Spanish, or *homme* and *hombre*. And finally, the system assumes a similar mapping for parts of words. A word like "overloading," for instance, has both a prefix — "over" — and a suffix — "ing." The system would anticipate that other words in the language will feature the prefix "over" or the suffix "ing" or both, and that a cognate of "overloading" in another language — say, "surchargeant" in French — would have a similar three-part structure.

Crosstalk

The system plays these different levels of correspondence off of each other. It might begin, for instance, with a few competing hypotheses for alphabetical mappings, based entirely on symbol frequency — mapping

symbols that occur frequently in one language onto those that occur frequently in the other. Using a type of probabilistic modeling common in artificial-intelligence research, it would then determine which of those mappings seems to have identified a set of consistent suffixes and prefixes. On that basis, it could look for correspondences at the level of the word, and those, in turn, could help it refine its alphabetical mapping. “We iterate through the data hundreds of times, thousands of times,” says Snyder, “and each time, our guesses have higher probability, because we’re actually coming closer to a solution where we get more consistency.” Finally, the system arrives at a point where altering its mappings no longer improves consistency.

Ugaritic has already been deciphered: Otherwise, the researchers would have had no way to gauge their system’s performance. The Ugaritic alphabet has 30 letters, and the system correctly mapped 29 of them to their Hebrew counterparts. Roughly one-third of the words in Ugaritic have Hebrew cognates, and of those, the system correctly identified 60 percent. “Of those that are incorrect, often they’re incorrect only by a single letter, so they’re often very good guesses,” Snyder says.

Furthermore, he points out, the system doesn’t currently use any contextual information to resolve ambiguities. For instance, the Ugaritic words for “house” and “daughter” are spelled the same way, but their Hebrew counterparts are not. While the system might occasionally get them mixed up, a human decipherer could easily tell from context which was intended.

Babel

Nonetheless, Andrew Robinson remains skeptical. “If the authors believe that their approach will eventually lead to the computerised ‘automatic’ decipherment of currently undeciphered scripts,” he writes in an e-mail, “then I am afraid I am not at all persuaded by their paper.” The

researchers' approach, he says, presupposes that the language to be deciphered has an alphabet that can be mapped onto the alphabet of a known language — "which is almost certainly not the case with any of the important remaining undeciphered scripts," Robinson writes. It also assumes, he argues, that it's clear where one character or word ends and another begins, which is not the case with many deciphered and undeciphered scripts.

"Each language has its own challenges," Barzilay agrees. "Most likely, a successful decipherment would require one to adjust the method for the peculiarities of a language." But, she points out, the decipherment of Ugaritic took years and relied on some happy coincidences — such as the discovery of an axe that had the word "axe" written on it in Ugaritic. "The output of our system would have made the process orders of magnitude shorter," she says.

Indeed, Snyder and Barzilay don't suppose that a system like the one they designed with Knight would ever replace human decipherers. "But it is a powerful tool that can aid the human decipherment process," Barzilay says. Moreover, a variation of it could also help expand the versatility of translation software. Many online translators rely on the analysis of parallel texts to determine word correspondences: They might, for instance, go through the collected works of Voltaire, Balzac, Proust and a host of other writers, in both English and French, looking for consistent mappings between words. "That's the way statistical translation systems have worked for the last 25 years," Knight says.

But not all languages have such exhaustively translated literatures: At present, Snyder points out, Google Translate works for only 57 languages. The techniques used in the decipherment system could be adapted to help build lexicons for thousands of other languages. "The technology is very similar," says Knight, who works on machine translation. "They feed off each other."

More information: Paper on deciphering Ugaritic - people.csail.mit.edu/bsnyder/p … /bsnyder_acl2010.pdf

Provided by Massachusetts Institute of Technology