Carnegie Mellon releases data on Haitian Creole to hasten development of translation tools

January 27 2010

In response to the humanitarian crisis in Haiti, scientists at Carnegie Mellon University's Language Technologies Institute (LTI) have publicly released spoken and textual data they've compiled on Haitian Creole so that translation tools desperately needed by doctors, nurses and other relief workers on the earthquake-ravaged island can be rapidly developed.

Since Carnegie Mellon began to make the data publicly available last week, a team at Microsoft Research has used it to help develop an experimental, web-based system for translating between English and Haitian Creole (http://www.microsofttranslator.com/).

Translators Without Borders (http://www.tsf-twb.org/), a not-for-profit association based in Paris, plans to distribute a medical triage dictionary to doctors in Haiti once that data has been converted into a readable format. LTI researchers, likewise, have begun working on their own translation system for Haitian Creole.

Although French is the official language of Haiti and is spoken by elites, Haitian Creole is the most widely spoken language in Haiti, said Robert Frederking, LTI senior systems scientist. Haitian Creole is based on French, but has evolved substantially since Haitians overthrew the French colonists more than 200 years ago. Word meanings have drifted and the language incorporates some African syntax.

"French speakers can sort of puzzle through it, but Creole isn't penetrable if you don't know French," Frederking said. Few translation resources are available for the language, he added.

The Carnegie Mellon data base for Haitian Creole was created in the late 1990s for Diplomat, a project sponsored by the Defense Advanced Research Projects Agency. The project was headed by Jaime Carbonell, LTI director, and focused on developing portable, speech-to-speech translation devices that could be deployed rapidly for Haitian Creole and other languages of special interest to the Department of Defense. Frederking and Alex Rudnicky, principal systems scientist in the Computer Science Department, served as co-principal investigators.

A prototype Haitian Creole translation system was delivered to the U.S. Army, but "as far as we know, nobody ever field-tested it," Frederking said. The project ended in the late 1990s, but LTI retained the data compiled and produced for the project.

Since the Jan. 12 earthquake, LTI researchers decided to begin work on an updated translation system for Haitian Creole that would incorporate the latest translation technologies. To aid other groups pursuing parallel efforts worldwide, they also opted to release the data publicly at www.speech.cs.cmu.edu/haitian/, making it available with minimal restrictions. In addition to the Diplomat material, other data developed by researchers at LTI and elsewhere are being added to the site as they become available.

Given the extreme poverty of Haiti, "nobody is going to make money on a Haitian Creole translator," Frederking said. "But translation systems could be an important tool, both for the relief workers now involved in emergency response and in the long-term as rebuilding takes place."

Provided by Carnegie Mellon University