Day 3- April 5, 2006

Automatic Machine Translation from Poorly Studied Languages - John Goldsmith, Professor, Departments of Linguistics and Computer Science, University of Chicago

Goldsmith's research, which is partly funded through JTAC, seeks to meld advances in linguistic science and computer science to develop a tool to better understand the structure of languages that may not have received significant attention. He began his presentation by providing an overview of the historical progress of computer development in linguistically-related areas. In the 90s, for example, there was a change in computational linguistics based on data-driven statistical techniques (i.e. statistical machine translation that allows for word to word matching and common word alignment). In 1999, a breakthrough occurred with the Egypt project, which provided a platform for an easy to use mechanism to not just translate sentences, but also provide a way to understand sentence construction (how words are patterned/ordered), which can vary from language to language, and result in "null" translations (e.g., "le chien", translated into English, would result in the "le" being a "null"). The University of Chicago project (Linguistica) is using this platform to create a pathway for users to learn complex structures of languages (understand the grammar) and can address difficult compound words (i.e. the morphemes within a word), and how those words correspond to one another across languages. It is seeking to create an automatic mechanism to analyze morphological structures and is now being tested on Swahili. Currently 20,000 words can be analyzed in about 15 seconds. This project can play an important part in improving language awareness in foreign situations and, in doing so, would improve understanding of larger cultural issues as well.

Interesting insights from the Q/A:

- Software can work with mixed languages (i.e. a mix of Kurdish and Iraq) because the computer does not know that the languages are mixed.

- Goldsmith would like to look at a "non-standard" dialect of Arabic. The matter of getting this data is not trivial.