Submitted by: MITRE
Presenter: Sherri Condon, Ph.D.
Additional Contributor/Authors: Luis Hernandez (ARL), Ghulam Hazrat Jahed (ARL), Dan Parvaz (MITRE), Mohammad Shahab Khan (MITRE), Michelle Vanni (ARL)
Topic: Producing Data for Under-Resourced Languages: A Dari-English Parallel Corpus of Multi-Genre Text

In Developers producing language technology for under-resourced languages often find relatively little machine readable text for data required to train machine translation systems. Typically, the kinds of text that are most accessible for production of parallel data are news and news-related genres, yet the language that requires translation for analysts and decision-makers reflects a broad range of forms and contents. The proposed paper will describe an effort funded by the ODNI FLPO in which the Army Research Laboratory, assisted by MITRE language technology researchers, produced a Dari-English parallel corpus containing text in a variety of styles and genres that more closely resemble the kinds of documents needed by government users than do traditional news genres.

The data production effort began with a survey of Dari documents catalogued in a government repository of material obtained from the field in Afghanistan. Because the documents in the repository are not available for creation of parallel corpora, the goal was to quantify the types of documents in the collection and identify their linguistic features in order to find documents that are similar. Document images were obtained from two sources: (1) the Preserving and Creating Access to Unique Afghan Records collection, an online resource produced by the University of Arizona Libraries and the Afghanistan Centre at Kabul University and (2) The University of Nebraska Arthur Paul Afghanistan Collection. For the latter, document images were obtained by camera capture of books and by selecting pdf images of microfiche records.

A set of 1395 document page images was selected to provide 250,000 translated English words in 10 content domains. The images were transcribed and translated according to specifications designed to maximize the quality and usefulness of the data. The corpus will be used to create a Dari-English glossary, and an experiment will quantify improvements to Dari-English translation of multi-genre text when a generic Dari-English machine translation system is customized using the corpus. The proposed paper will present highlights from these efforts..