Research Objective

- Develop semi-automated methods for detecting Arabic dialect text written in Arabic & Roman scripts
- Adapt detection algorithm for “code-switching,” the alternation of languages/dialects within text, as commonly found in social media

Challenges

- Historically, few texts are written in Arabic dialects and there are no established standards for spelling and grammar
- As currently written, dialects appear in very noisy & informal contexts, such as social media, and are frequently mixed with other languages/dialects
- Code-switching (alternating between two or more languages) in texts precludes use of “state-of-the-art” statistical machine translation systems: these require training sets of large parallel-aligned, monolingual corpora

ARL Facilities and Capabilities Available to Support Collaborative Research

- Dialect classification experience
  - at segment and at token levels
  - on Arabic and on Roman script
  - best published cross-validation accuracy
  - on Zaidan & Callison-Burch’s (2011) Arabic Online Commentary dataset
- first published Moroccan dialect (Darija) classifier
- Arabic morphological and syntactic parsing software provides important pre-processing capabilities
- Dialect id software for viewing, annotating, and running in-house classifiers on social media conversations
- Native language expertise

Complementary Expertise/ Facilities/ Capabilities Sought in Collaboration

- Visualization expertise—demonstrate the value of highlighting code-switching and other linguistic phenomena to intelligence analysts
- Social network analysis expertise—leverage social network for improved language/dialect detection
- Prowess in active learning
- Additional native language experts
- Access to large datasets