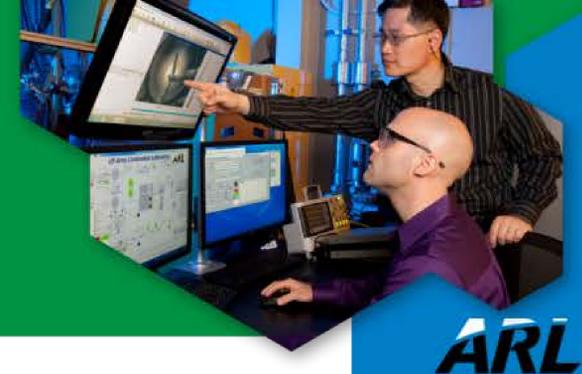


# Linguistic Core for Analysis and Machine Translation of No/Low-Resource Languages



**S&T Campaign: Information Sciences**  
*Intelligent Systems: Information Understanding*

Dr. S. Tratz (301) 394-2305, [stephen.c.tratz.civ@mail.mil](mailto:stephen.c.tratz.civ@mail.mil)  
 Mr. J. Laoudi (301) 394-5629, [jamal.laoudi.ctr@mail.mil](mailto:jamal.laoudi.ctr@mail.mil)  
 Dr. C. Voss (301) 394-5615, [clare.r.voss.civ@mail.mil](mailto:clare.r.voss.civ@mail.mil)

## 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

 User 1 ! o_O ! هني مجتمعي التكم عن السياسة .. فيلم رعب Expand	Modern Standard Arabic <i>Arabic script</i>
 User 2 @user_2 @user_1 eh 3ande9 l7a9 nas kiykhafo yhedro f siyassa Expand	Moroccan Arabic <i>Romanized</i>
 User 1 @user_1 @user_2 !!! انا كنهدر في السياسة نيشان ماما كتقولني سكتي !!! Expand	Moroccan Arabic <i>Arabic script</i>
 User 2 @user_2 @user_1 7ta ana kiy9ololi chi nhar ghatjibiha f rassek hhhh Expand	Moroccan Arabic <i>Romanized</i>

Arabic dialects are widely spoken and now appear online in both Arabic and Roman scripts, as shown above. Code-switching appears frequently within conversations.

## 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

here Darija shows S-V-O word order  
 → like English

Darija	!!! سكتي	كتقولني	ماما	نيشان	السياسة	في	كنهدر	انا
Gloss	(yourself)+silence	me+tells+(she)	my+mom	freely	politics+the	about	talk+(I)	I
Syntax	(O <sub>3</sub> )+V <sub>3</sub>	IO <sub>2</sub> +V <sub>2</sub> +(S <sub>2</sub> )	S <sub>2</sub>				V <sub>1</sub> +(S <sub>1</sub> )	S <sub>1</sub>

English Reference Translation: [Whenever] I talk freely about politics[,] my mom tells me to be quiet!!!

Google MT: I Knhdr in politics Nishane Mama Ktcola Scotty!!!

→ unlike MSA (V-S-O word order)

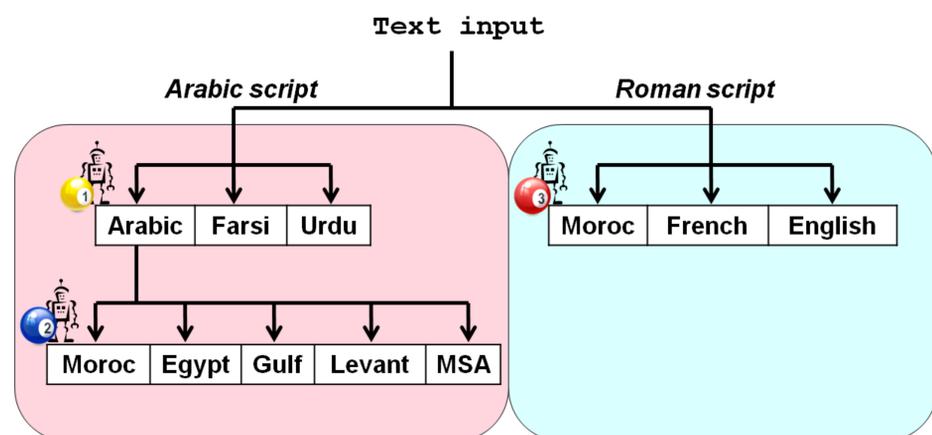
Modern Standard Arabic (MSA)	!!! اسكتي	أمي	لي	تقول	صراحة	بكل	السياسة	في	أتكلم
Gloss	(yourself)+silence	my+mother	me+to	tells+(she)	frankness	all+with	politics+the	about	speak+(I)
Syntax	(O <sub>3</sub> )+V <sub>3</sub>	S <sub>2</sub>	IO <sub>2</sub>	V <sub>2</sub> +(S <sub>2</sub> )					V <sub>1</sub> +(S <sub>1</sub> )

Google MT: Speak in politics frankly says me SHUT mom!!!

Grammatical analyses of Arabic dialects suggest that their syntax follows Subject-Verb-Object (SVO) word order, as opposed to the dominant word order of Modern Standard Arabic (MSA) which is Verb-Subject-Object (VSO).

## 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



ARL Arabic language/dialect classifiers