ALDi: Quantifying the Arabic *Level of Dialectness* of Text

Amr Keleg, Sharon Goldwater, Walid Magdy

Demo

Amrkeleg

1 The Linguistic Variation of Arabic

Previous Work - Regional Variation

Egypt ⟷ MSA ⟷ Levant

أسعدنا الرجل
AlrjlÂsçdnA

Egypt	Levant	
الراجل أسعدنا ÂsçdnA AlrAjl	الزلمة أسعدنا ÂsçdnA Alzlmħ	Low
الراجل بسطنا bsTnA AlrAjl	الزلمة بسطنا bsTnA Alzlmħ	Medium
الراجل شهيصنا šhySnA AlrAjl	الزلمة نغنجنا nγnjnA Alzlmħ	High

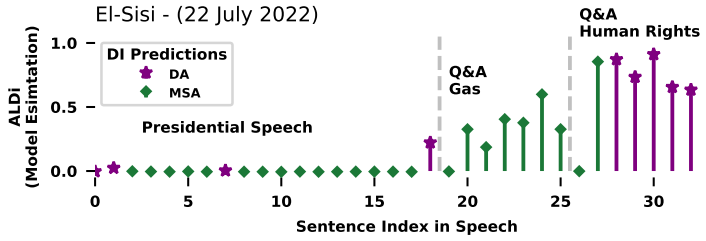
Our Work - Level of Dialectness

The man cheered us

2 ALDi as a Sociolinguistic Variable

Current Egyptian President

El-Sisi - (22 July 2022)



3 Arabic Level of Dialectness (ALDi)

- **Definition**: Divergence from Standard Language.
- **Operationalization**: Score in $[0, 1]$ on sentence-like level.

Arabic Online Commentary Dataset (Zaidan et. al, 2011)
Popular Dialect Identification (DI) labels.
Ignored *Discrete* Level of Dialectness labels!
Embrace annotators disagreement in AOC-ALDi!

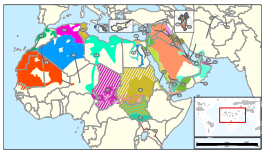
From AOC to AOC-ALDi

1. Labels into numeric values
2. Algebric Mean
3. Regression-head on top of MarBERT

MSA	Little	Mixed	Most
0	1/3	2/3	1

e.g., ALDi(MSA,MSA,Little)$=\overline{(0, 0, \frac{1}{3})} = \frac{1}{9} \approx 0.11$

Eager for more?



*** AOC-ALDi Dataset**
- 127,835 sentences (3 annotations each).
- Comments to news articles.
- Publishers from EGY JOR SA.
- Fleiss' $\kappa = 0.44$
- Krippendorff's α (interval) $= 0.63$

*** Example of Disagreement**

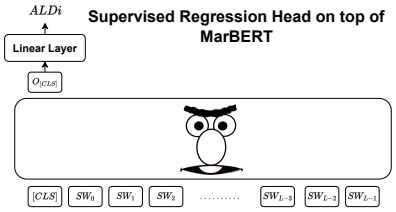
Observation: Pronouncability of sentence in a dialect impacts the perceived ALDi.

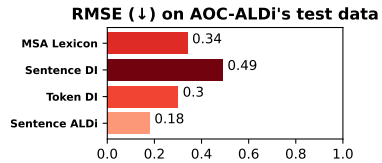
نبتدى بقى الشغل الصح فى تطوير المدارس وتوفير المراقبين عليها

We start with the right task of developing schools and providing observers over them

LOW LOW HIGH

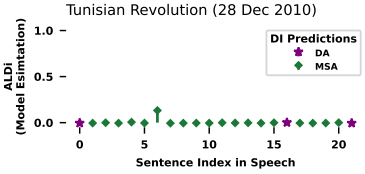
*** Models and Evaluation**



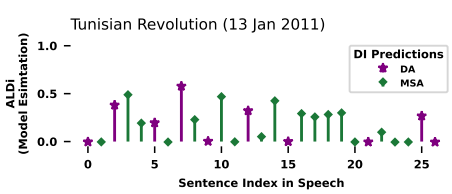


*** Former Tunisian President's Styles**

Authoritative (first speech)



Seeking Empathy (last speech)



THE UNIVERSITY of EDINBURGH informatics
ilcc Institute for Language, Cognition and Computation
NLP UKRI CENTRE FOR DOCTORAL TRAINING