Incorporating Sociolinguistic Theories for a Better Modeling of the Arabic Varieties

Amr Keleg

Working with: Walid Magdy, Sharon Goldwater

Cardiff University

8 May 2025

Final-year PhD Student at the University of Edinburgh

Arabic 101

- Spoken in 22 Arab Countries.
  - + Minority groups in other non-Arab countries.

Arabic 101

- Spoken in 22 Arab Countries.
  - + Minority groups in other non-Arab countries.

- Vast geographical area ⇒ Variation.
- Additionally, impact of:
  - Other local languages in the region (e.g., Tamazight, Coptic)
  - Colonial languages (e.g., English, French, Italian)
  - Contact languages (e.g., Greek, Persian)

Example of Lexical Variation

English
CAT

Source: `https://www.youtube.com/shorts/WXL4xITDreI`

Example of Lexical Variation

⚠️ Non-negligible Degrees of Variation in:

1. Phonology
2. Morphology
3. Lexicon
4. Syntax (e.g., Word Order)
5. Semantics (e.g., False Friends)
6. Culture?

Keleg, Amr. “LLM Alignment for the Arabs: A Homogenous Culture or Diverse Ones.” C3NLP 2025 - NAACL 2025.

⚠️ Non-negligible Degrees of Variation in:

1. Phonology
2. Morphology
3. Lexicon
4. Syntax (e.g., Word Order)
5. Semantics (e.g., False Friends)
6. Culture?

Interspeaker Variation ❗

Keleg, Amr. "LLM Alignment for the Arabs: A Homogenous Culture or Diverse Ones." C3NLP 2025 - NAACL 2025.

Theory of Intraspeaker Variation #1: Diglossia (Ferguson, 1959)

Diglossia

“a language state in which two varieties of the language co-exist within the same speaking community:

- a high variety linked to higher prestige
  - Modern Standard Arabic (MSA)
- a low variety perceived to be of lower status.
  - Arabic Dialects”

Ferguson, Charles A. “Diglossia.” WORD, 1959.

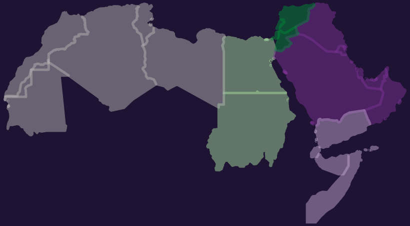
- MSA was the language of literary
    - e.g., books, newspapers, ...
- Arabic Dialects increasingly written online
    - e.g., texting, social media, ...

Operationalization of *Diglossia* in NLP

- Dialect Identification (i.e., sentence -> dialect)

- Dialect Identification (i.e., sentence -> dialect)

- Dialect Identification (i.e., sentence -> dialect)



- MSA as an independent dialect.

❌ Two Limitations

1. Disjointedness
   - A sentence valid in a dialect **can not** be valid in another dialect.

Keleg, Amr and Magdy, Walid. “Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification.” ArabicNLP 2023; Abdul-Mageed, Muhammad et al. “NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task.” ArabicNLP 2024.

❌ Two Limitations

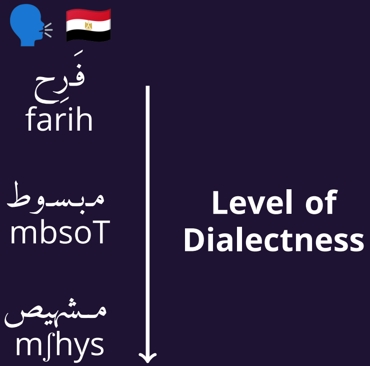
1. Disjointedness
   - A sentence valid in a dialect **can not** be valid in another dialect.

2. **Binarization**
   - Dialectal sentences are equally divergent from MSA.

Keleg, Amr and Magdy, Walid. “Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification.” ArabicNLP 2023; Abdul-Mageed, Muhammad et al. “NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task.” ArabicNLP 2024.

🔬 The Binarization Limitation

Different ways of saying: happy 😀



🗣️ 🇪🇬

فَرِح
farih

مبسوط
mbsoT

مشهيص
mʃhys

Level of Dialectness

Different ways of saying: happy 😀

MSA root meaning	Root	🗣️ 🇪🇬
to be happy	فرح frh	فَرِح farih
		مبسوط mbsoT
		مشهيص mʃhys

↓ **Level of Dialectness**

Different ways of saying: happy 😀

MSA root meaning	Root	🗣️ 🇪🇬
to be happy	فرح frh	فَرِح farih
extend - cheer	بسط bsT	مبسوط mbsoT
N/A	شهيص ʃhys	مشهيص mʃhys

↓ **Level of Dialectness**

Theory of Intraspeaker Variation #2: Dialect Levels (Badawi, 1973)

Dialect Levels

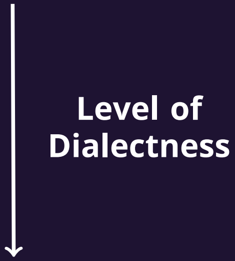
- focused on spoken language
- identified five different dialect levels
- with examples of linguistic features for each level

Badawi, As-Said Muhámmad. *Levels of Contemporary Arabic in Egypt* (*مستويات العربية المعاصرة في مصر*.) Dar Al-Maarif (1973).

Dialect Levels

Note: Fus-ha فصحى is the term Arabs use for the standardized classical and modern varieties.

1. Heritage Fus-ha
2. Fus-ha of the age (we live in)
3. Dialect of the (well-)educated
4. Dialect of the Literate
5. Dialect of the Illiterate



Level of Dialectness

Badawi, As-Said Muhámmad. *Levels of Contemporary Arabic in Egypt* (مستويات العربية المعاصرة في مصر.) Dar Al-Maarif (1973).

Operationalization of *Dialect Levels* in NLP

Few efforts mainly proposing guidelines and annotating limited data:

1. (Habash, 2008), (Elfardy, 2012)
   - token-level annotations mapped to sentence-level
   - data mostly MSA
   - not publicly available

2. (Zaidan, 2011)

Habash, Nizar et al. “Guidelines for annotation of Arabic dialectness.” Workshop on HLT & NLP - LREC 2008.

Elfardy, Heba and Diab, Mona. “Simplified guidelines for the creation of Large Scale Dialectal Arabic Annotations.” LREC 2012.

Zaidan, Omar F. and Callison-Burch, Chris. “The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content.”

AOC Dataset

🗄️ **Arabic Online Commentary Dataset (Zaidan et. al, 2011)**

- 💬 Comments to news articles
- 📰 Three publishers (Egypt, Jordan, Saudi Arabia)
- 🎯 127,835 sentences (3👤annotations each)
- 🌟 Popular Dialect Identification (DI) labels

Zaidan, Omar F. and Callison-Burch, Chris. “The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content.”

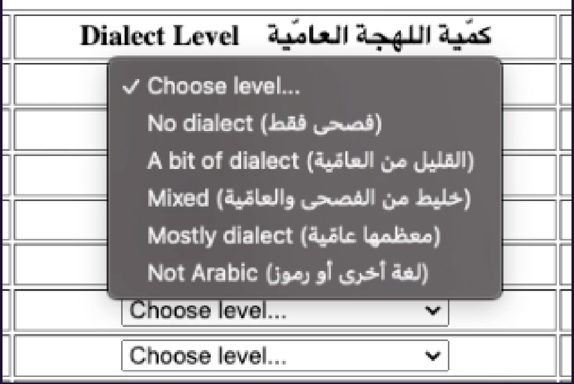
AOC Dataset

🗄️ **Arabic Online Commentary Dataset (Zaidan et. al, 2011)**

- 💬 Comments to news articles
- 📰 Three publishers (Egypt, Jordan, Saudi Arabia)
- 🎯 127,835 sentences (3 👤 annotations each)
- 🌟 Popular Dialect Identification (DI) labels
- 🗑️ Ignored *Discrete* Level of Dialectness labels!

Zaidan, Omar F. and Callison-Burch, Chris. "The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content."

Tell us how much dialect (عامية) is in the sentence.



Dialect Level كمّية اللهجة العامّية

✓ Choose level...

No dialect (فصحى فقط)

A bit of dialect (القليل من العامّية)

Mixed (خليط من الفصحى والعامّية)

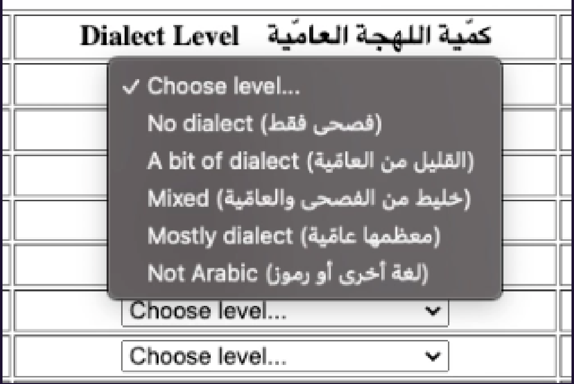
Mostly dialect (معظمها عامّية)

Not Arabic (لغة أخرى أو رموز)

Choose level...

Choose level...

Tell us how much dialect (عامية) is in the sentence.



Dialect Level كمّية اللهجة العاميّة

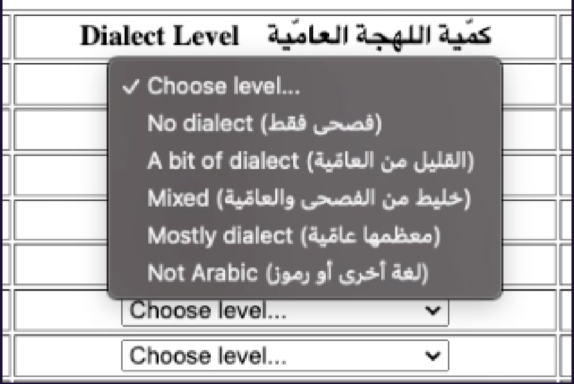
- ✓ Choose level...
- No dialect (فصحى فقط)
- A bit of dialect (القليل من العامية)
- Mixed (خليط من الفصحى والعامية)
- Mostly dialect (معظمها عامية)
- Not Arabic (لغة أخرى أو رموز)

Choose level...

Choose level...

🤝 Fleiss' κ = 0.44

Tell us how much dialect (عامية) is in the sentence.



Dialect Level كمّية اللهجة العامّية

✓ Choose level...

No dialect (فصحى فقط)

A bit of dialect (القليل من العامّية)

Mixed (خليط من الفصحى والعامّية)

Mostly dialect (معظمها عامّية)

Not Arabic (لغة أخرى أو رموز)

Choose level...

Choose level...

🤝 Fleiss' κ = 0.44

🤔 Embrace annotators disagreement!

Sentence with two valid pronunciations

نبتدى بقى الشغل الصح فى تطوير المدارس وتوفير
المراقبين عليها

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

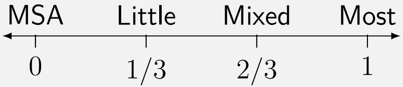
Sentence with two valid pronunciations

نبتدى بقى الشغل الصح فى تطوير المدارس وتوفير المراقبين عليها

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

♻️♻️♻️ From AOC to AOC-ALDi ♻️♻️♻️

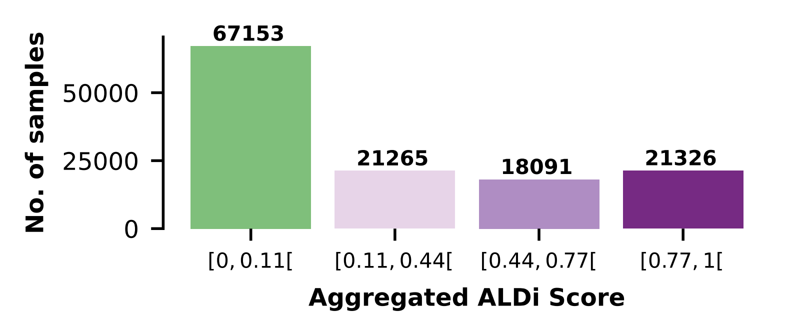
1️⃣ Labels into numeric values 2️⃣ Algebric Mean



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

🤝 Krippendorff’s α (interval) = 0.63

ALDi scores Distribution



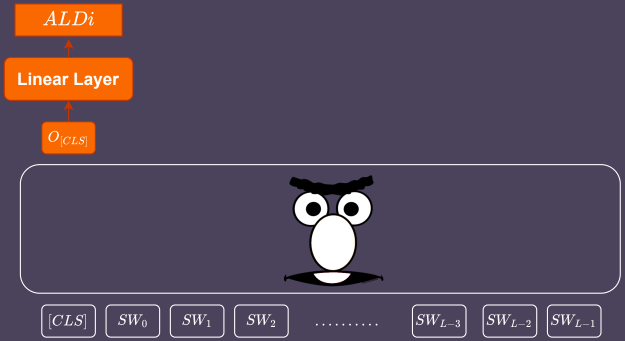
⭐⭐⭐ Our Operationalization ⭐⭐⭐

- **Arabic Level of Dialectness (ALDi):** Divergence from Standard Arabic (MSA).
- Continuous score in [0, 1].
- Sentence-like level.

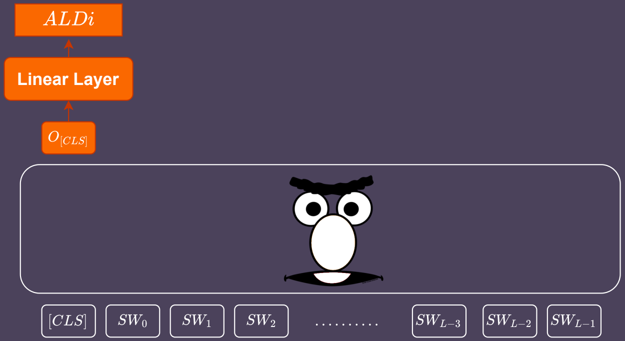
Keleg, Amr, Goldwater, Sharon, and Magdy, Walid. “ALDi: Quantifying the Arabic Level of Dialectness of Text.” EMNLP 2023.

🚧 Building a model to estimate ALDi automatically

Sentence-ALDi model

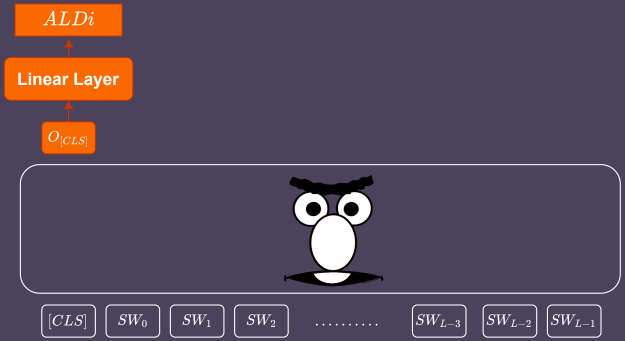


Sentence-ALDi model



🌍 Dialect-agnostic

Sentence-ALDi model



🌍 Dialect-agnostic

🔬 $\text{RMSE}(\text{test set}) = 0.18$

📐 Applications of ALDi

1. **Analyzing Intraspeaker Variation (Presidential Speeches)**

Keleg, Amr, Goldwater, Sharon, and Magdy, Walid. “ALDi: Quantifying the Arabic Level of Dialectness of Text.” EMNLP 2023.

Case Study - Presidential speeches

- Arab presidents use:
  - Modern Standard Arabic (MSA) - authority
  - Dialectal Arabic (DA) - compassion and belonging
- Replicated for speeches of former Tunisian and Egyptian presidents.

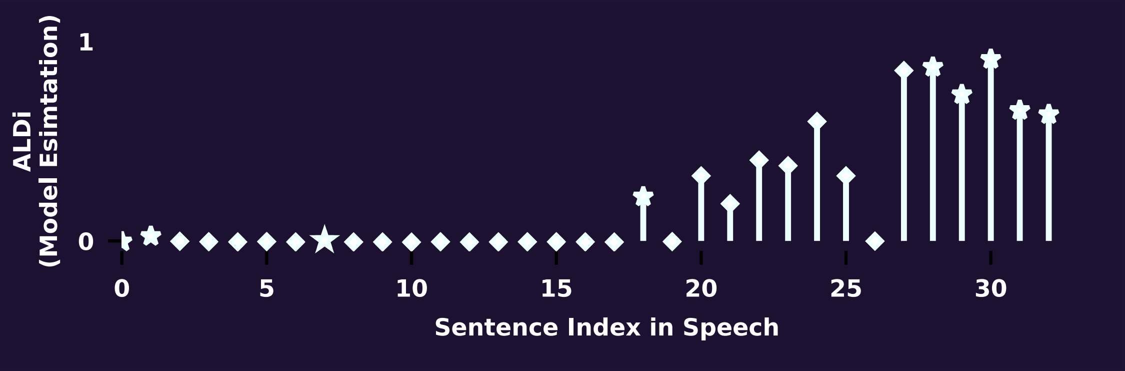
Lahlali, M. “The Arab Spring and the discourse of desperation: shifting from an authoritarian discourse to a “democratic one”.” (2011).

Egyptian President (El-Sisi) - 18/07/2022

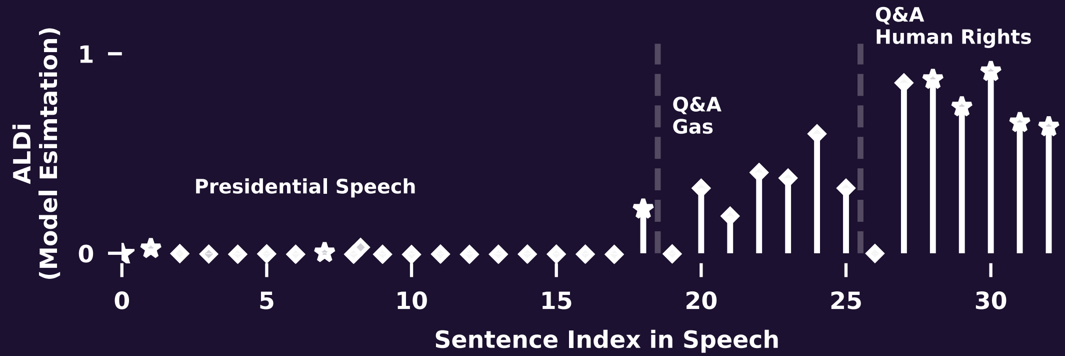
Source: `https://beta.sis.gov.eg/en/presidency/presidents-monthly-activities/presidents-activities-during-july-2022/`

Egyptian President (El-Sisi) - 18/07/2022

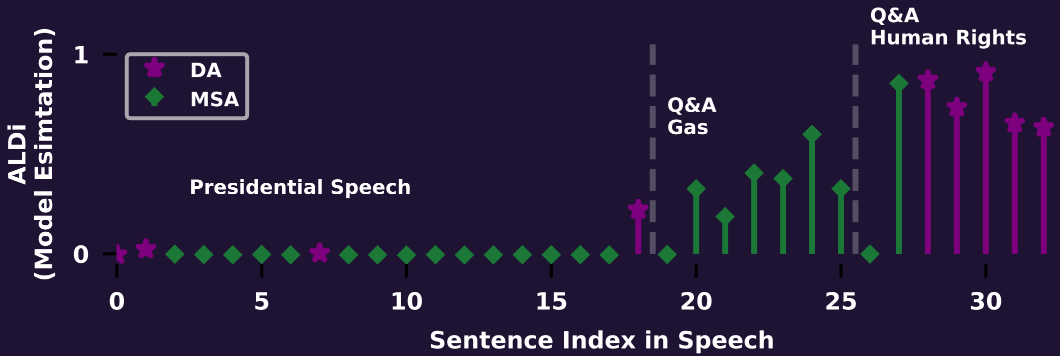
Egyptian President (El-Sisi) - 18/07/2022



Egyptian President (El-Sisi) - 18/07/2022



Egyptian President (El-Sisi) - 18/07/2022



📐 Applications of ALDi

2 Impact of Interannotator Agreement

Keleg, Amr, Magdy, Walid, and Goldwater, Sharon. “Estimating the Level of Dialectness Predicts Inter-annotator Agreement in Multi-dialect Arabic Datasets.” ACL 2024.

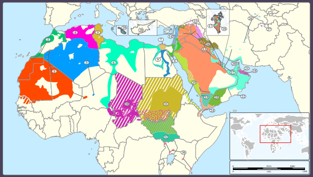
2) Annotating Multi-Dialect Arabic Datasets

Common Practice: 🎲 randomly assign to Arabic speakers

☢️☢️☢️ **Annotator's dialect $\neq$ Sample's dialect** ☢️☢️☢️

📎 More strict annotating Hate Speech 🤬 (Bergman and Diab, 2022)

📎 Less accurate identifying Sarcasm 🤪 (Abu Farha and Magdy, 2022)



Bergman, A. and Diab, Mona. "Towards Responsible Natural Language Annotation for the Varieties of Arabic."

Abu Farha, Ibrahim and Magdy, Walid. "The Effect of Arabic Dialect Familiarity on Data Annotation."

Annotation Codebook (v1.0) 📚

🏷️ **Step 1:** Identify the dialect of each sample

🔀 **Step 2:** Route the sample to speakers of its dialect

Mubarak, Hamdy and Darwish, Kareem. “Demographic surveys of Arab annotators on CrowdFlower.”

Annotation Codebook (v1.0) 📚

🏷️ **Step 1:** Identify the dialect of each sample

🔀 **Step 2:** Route the sample to speakers of its dialect

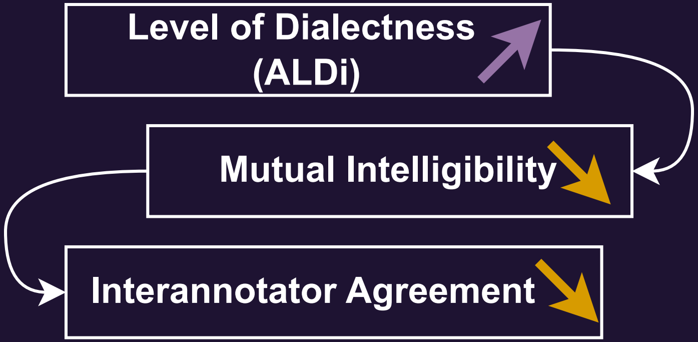
Hard to crowdsource speakers of some dialects (i.e., Limited resource 💎💎)

(Mubarak and Darwish, 2016)

Mubarak, Hamdy and Darwish, Kareem. “Demographic surveys of Arab annotators on CrowdFlower.”

🤔 Should some dialectal samples be priotrized?

💡 Intuition



Level of Dialectness (ALDi) ↗

Mutual Intelligibility ↘

Interannotator Agreement ↘

🧪 Analysis

📄 15 public datasets covering 6 Tasks:
Hatespeech, Sentiment Analysis, Dialect Identifcation, ...

(1) sentence-level classification datasets
(2) multi-dialect samples
(3) samples randomly assigned to annotators

(4) individual annotator labels

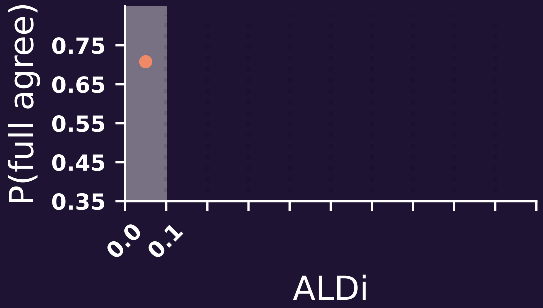
🧪 Methodology:

1. Estimate ALDi of samples.
2. Bin samples.
3. $\mathrm{P}_{\mathrm{bin}}(\mathrm{Full\ Agreement})$

$$\mathrm{P}_{\mathrm{bin}}(\mathrm{Full\ Agreement}) \approx \frac{\mathrm{N}_{\mathrm{(bin)\ Full\ Agreement}}}{\mathrm{N}_{\mathrm{(bin)\ Total\ Samples}}}$$

🧪 Methodology:

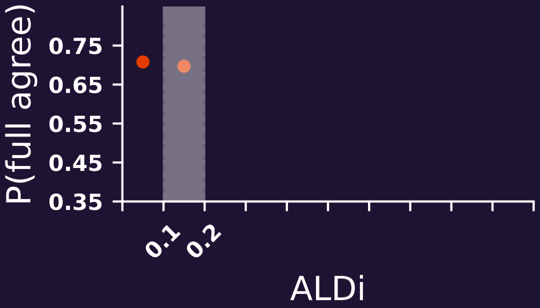
1. Estimate ALDi of samples.
2. Bin samples.
3. $\mathrm{P}_{\mathrm{bin}}(\mathrm{Full\ Agreement})$



$$\mathrm{P}_{\mathrm{bin}}(\mathrm{Full\ Agreement}) \approx \frac{\mathrm{N}_{\mathrm{(bin)\ Full\ Agreement}}}{\mathrm{N}_{\mathrm{(bin)\ Total\ Samples}}}$$

Methodology:

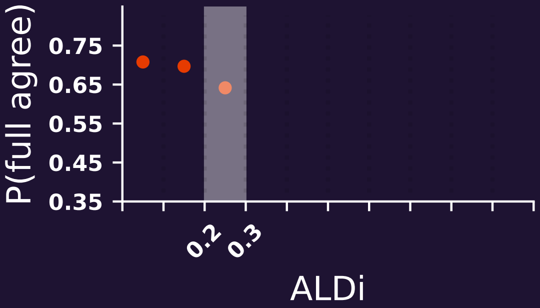
1. Estimate ALDi of samples.
2. Bin samples.
3. $P_{\text{bin}}(\text{Full Agreement})$



$$P_{\text{bin}}(\text{Full Agreement}) \approx \frac{N_{(\text{bin})\ \text{Full Agreement}}}{N_{(\text{bin})\ \text{Total Samples}}}$$

🧪 Methodology:

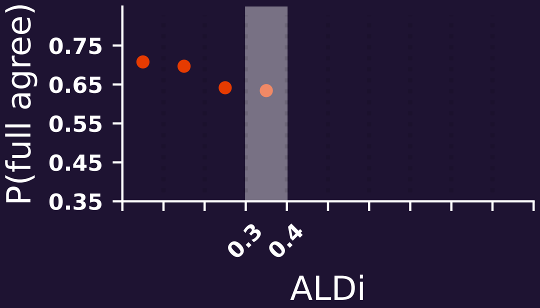
1. Estimate ALDi of samples.
2. Bin samples.
3. $P_{\text{bin}}(\text{Full Agreement})$



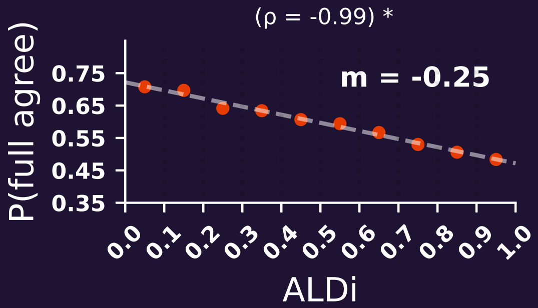
$$P_{\text{bin}}(\text{Full Agreement}) \approx \frac{N_{(\text{bin})\ \text{Full Agreement}}}{N_{(\text{bin})\ \text{Total Samples}}}$$

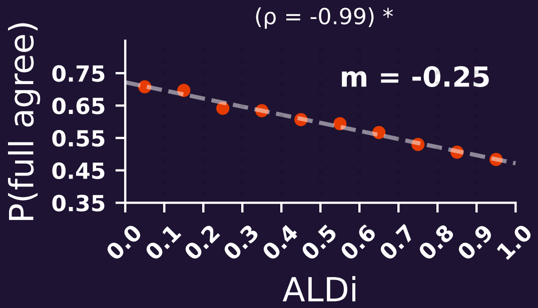
Methodology:

1. Estimate ALDi of samples.
2. Bin samples.
3. $P_{\text{bin}}(\text{Full Agreement})$



$$P_{\text{bin}}(\text{Full Agreement}) \approx \frac{N_{(\text{bin})\ \text{Full Agreement}}}{N_{(\text{bin})\ \text{Total Samples}}}$$


(ρ = -0.99) *
m = -0.25
P(full agree)
0.75
0.65
0.55
0.45
0.35
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ALDi



Finding (1) - For 8 of 12 non Dialect Identification datasets

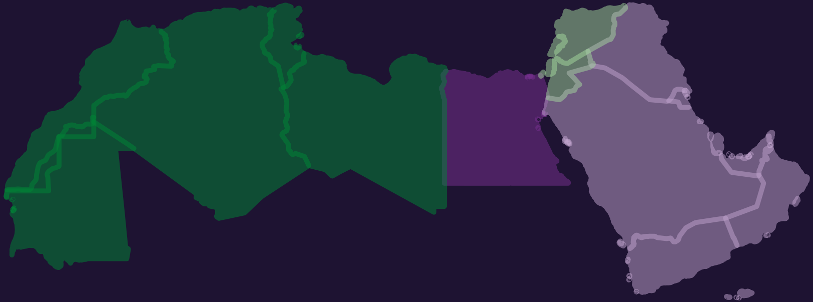
ALDi ↗ Interannotator Agreement ↘

with significant strong negative $\rho < -0.7$

IAA - Dialect Identification Dataset

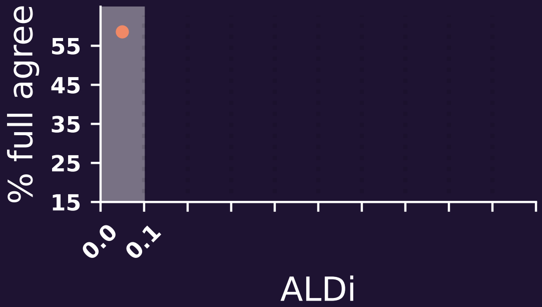
ℹ️ **Labels (Macro-regional):**

MSA, Maghreb, Egypt, Levant, Gulf



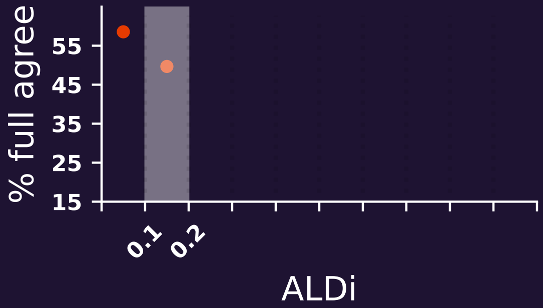
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



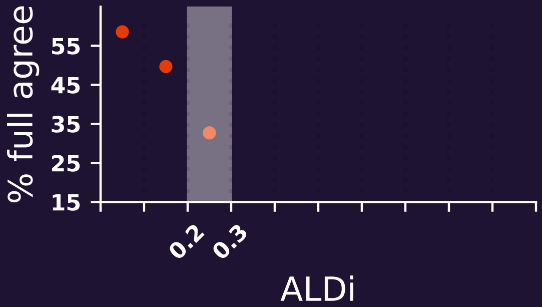
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



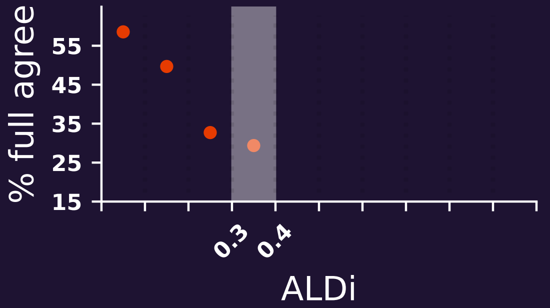
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



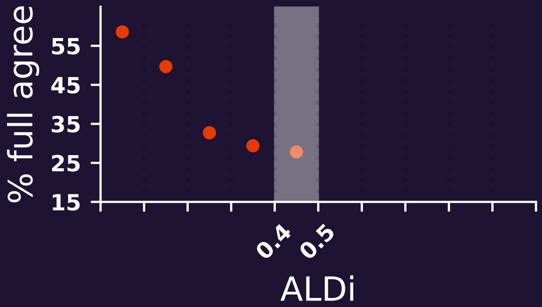
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



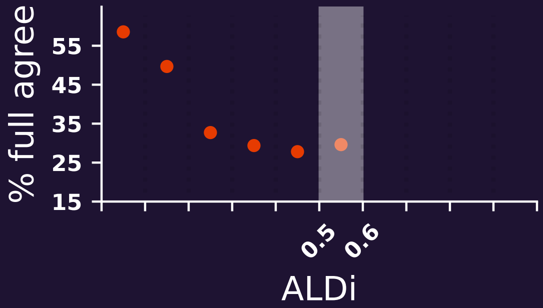
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



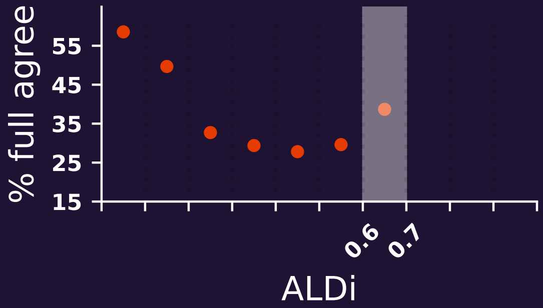
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



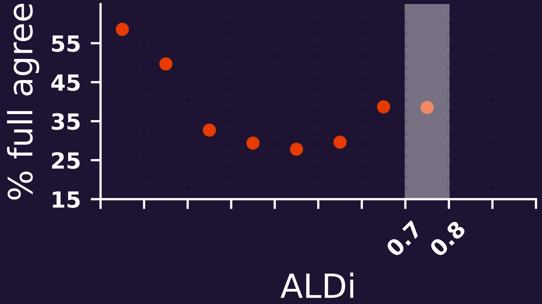
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



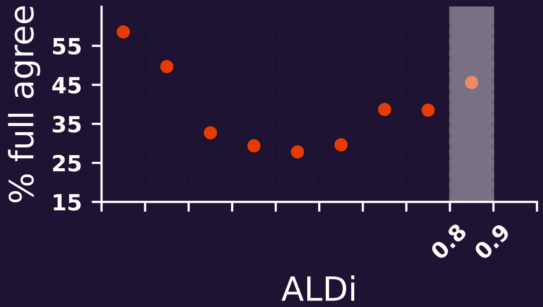
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



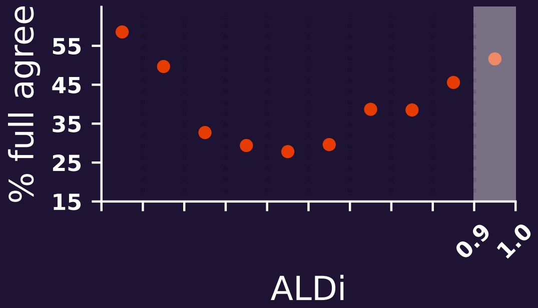
IAA - Dialect Identification Dataset

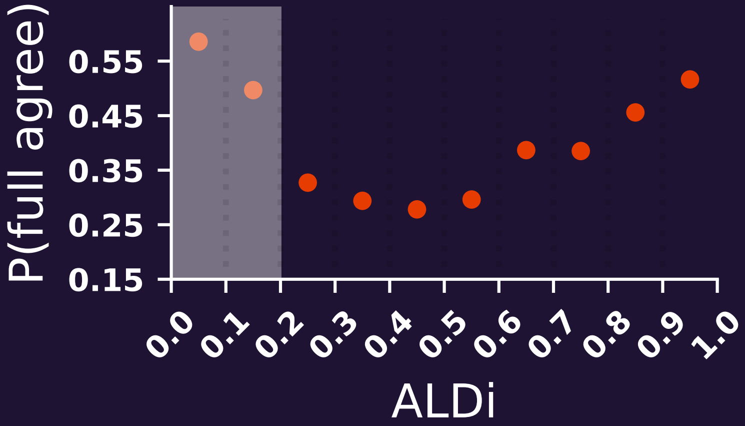
ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb



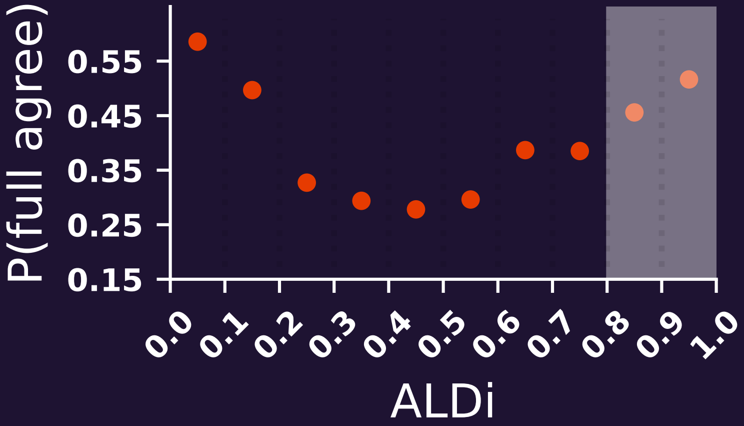
IAA - Dialect Identification Dataset

ℹ️ **Labels (Macro-regional):**
MSA, Egypt, Gulf, Levant, Maghreb

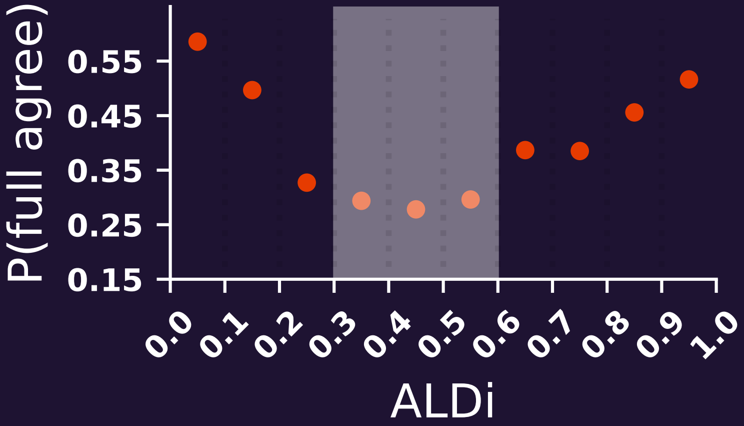



P(full agree)
0.55
0.45
0.35
0.25
0.15
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ALDi

📎 **MSA samples**



📎 **Highly DA samples (with distinctive cues?)**



📎 **1) Hard to determine the dialect?**

OR 📎 **2) Valid in multiple dialects?**

Annotation Codebook (v1.1) 📙

🚨 Prioritize routing high-ALDi samples to speakers of the samples’ respective dialects, (Finding 1)

🕵️ for which Dialect Identification is more accurate. (Finding 2)

Thanks!

🦋 @amr-keleg.bsky.social

Thanks!

🦋 @amr-keleg.bsky.social

Summary

1. Arabic sentences exist along a continuum
   - Pure MSA < —————- > Highly Dialectal
2. Adapting sociolingusitic theories can improve our NLP tools

🤩 Please enlighten me about variation in your native languages!

References I

- Abdul-Mageed, Muhammad et al. (2024). "NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task." In: *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)* (ArabicNLP 2024).

References II

- Abu Farha, Ibrahim and Walid Magdy (Dec. WANLP 2022). "The Effect of Arabic Dialect Familiarity on Data Annotation." In: *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*. Ed. by Houda Bouamor et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 399–408. DOI: `10.18653/v1/2022.wanlp-1.39`. URL: `https://aclanthology.org/2022.wanlp-1.39`.

- Badawi, As-Said Muhámmad (1973). *Levels of Contemporary Arabic in Egypt (مستويات العربية المعاصرة في مصر.)* Dar Al-Maarif.

References III

Bergman, A. and Mona Diab (May ACL (findings) 2022). “Towards Responsible Natural Language Annotation for the Varieties of Arabic.” In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 364–371. DOI: `10.18653/v1/2022.findings-acl.31`. URL: `https://aclanthology.org/2022.findings-acl.31`.

References IV

- Elfardy, Heba and Mona Diab (May 2012). “Simplified guidelines for the creation of Large Scale Dialectal Arabic Annotations.” In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)* (LREC 2012). Istanbul, Turkey: European Language Resources Association (ELRA), pp. 371–378. URL: `http://www.lrec-conf.org/proceedings/lrec2012/pdf/815_Paper.pdf`.
- Ferguson, Charles A (1959). “Diglossia.” In: *word* 15.2, pp. 325–340.

References V

- Habash, Nizar et al. (2008). "Guidelines for annotation of Arabic dialectness." In: *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world* (Workshop on HLT & NLP - LREC 2008), pp. 49–53.

References VI

Keleg, Amr (May 2025). “LLM Alignment for the Arabs: A Homogenous Culture or Diverse Ones.” In: *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)* (C3NLP 2025 - NAACL 2025). Ed. by Vinodkumar Prabhakaran et al. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 1–9. ISBN: 979-8-89176-237-4. URL: `https://aclanthology.org/2025.c3nlp-1.1/`.

References VII

Keleg, Amr, Sharon Goldwater, and Walid Magdy (Dec. 2023). “ALDi: Quantifying the Arabic Level of Dialectness of Text.” In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2023). Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 10597–10611. DOI: 10.18653/v1/2023.emnlp-main.655. URL: https://aclanthology.org/2023.emnlp-main.655.

References VIII

- Keleg, Amr and Walid Magdy (Dec. 2023). “Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification.” In: *Proceedings of ArabicNLP 2023* (ArabicNLP 2023). Ed. by Hassan Sawaf et al. Singapore (Hybrid): Association for Computational Linguistics, pp. 385–398. DOI: `10.18653/v1/2023.arabicnlp-1.31`. URL: `https://aclanthology.org/2023.arabicnlp-1.31`.

References IX

Keleg, Amr, Walid Magdy, and Sharon Goldwater (Aug. 2024). "Estimating the Level of Dialectness Predicts Inter-annotator Agreement in Multi-dialect Arabic Datasets." In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (ACL 2024). Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 766–777. DOI: `10.18653/v1/2024.acl-short.69`. URL: `https://aclanthology.org/2024.acl-short.69/`.

References X

- Lahlali, M (2011). "The Arab Spring and the discourse of desperation: shifting from an authoritarian discourse to a "democratic one"." In: *The Journ. of Arab Media and Society.—Cairo: American Univ. in Cairo* 13.
- Mubarak, Hamdy and Kareem Darwish (2016). "Demographic surveys of Arab annotators on CrowdFlower." In: *Proceedings of ACM WebSci16 Workshop "Weaving Relations of Trust in Crowd Work: Transparency and Reputation across Platforms*.

References XI

- S’hiri, Sonia (2002). “Speak Arabic please!: Tunisian Arabic speakers’ linguistic accommodation to middle Easterners.” In: *Language contact and language conflict in Arabic: Variations on a sociolinguistic theme*, pp. 149–174.
- Zaidan, Omar F. and Chris Callison-Burch (June 2011). “The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content.” In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 37–41. URL: `https://aclanthology.org/P11-2007`.

Exposure is key for Intelligibility

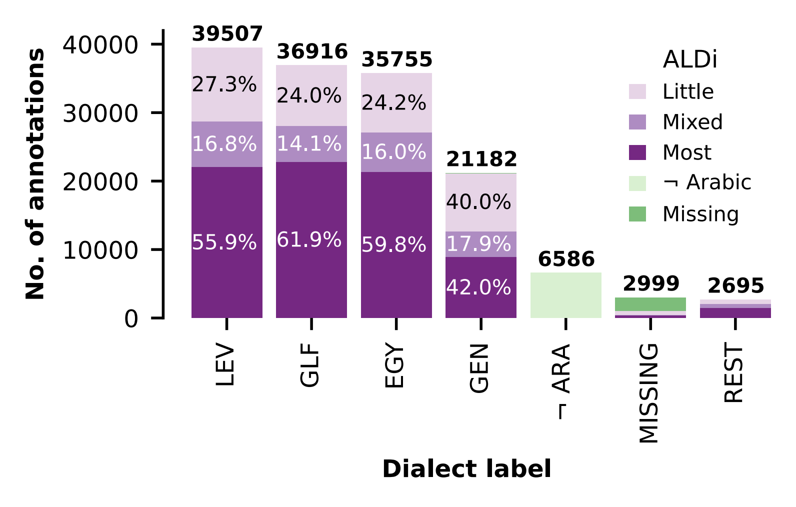
"Egyptian Arabic and to a lesser extent Levantine Arabic are widely understood because of the massive exposure to them through the media and the arts during the last generation or so." (S'hiri, 2002)

Sometimes compared to Romance Languages.

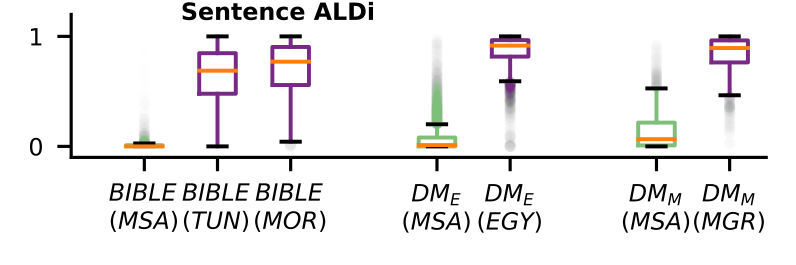
S'hiri, Sonia. "Speak Arabic please!: Tunisian Arabic speakers' linguistic accommodation to middle Easterners." Routledge (2002).

Why not different languages?

“MSA is a kind of communally-owned reservoir that Arabs use to ORANGEmake themselves understood to others from distant countries”. (Holes, 1995)


No. of annotations
0
10000
20000
30000
40000
39507
36916
35755
21182
6586
2999
2695
27.3%
16.8%
55.9%
24.0%
14.1%
61.9%
24.2%
16.0%
59.8%
40.0%
17.9%
42.0%
LEV
GLF
EGY
GEN
¬ ARA
MISSING
REST
Dialect label
ALDi
Little
Mixed
Most
¬ Arabic
Missing

ALDi scores automatically estimated:



NADI 2024 Dataset

Is it possible that the tweet is authored by someone who speaks one of your country’s dialects?

- 1,120 sentences.
- with geolocations uniformly distributed across 14 countries.
- 3 annotators from 9 different countries (total of 27)

Abdul-Mageed, Muhammad et al. “NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task.” ArabicNLP 2024.

NADI 2024 Dataset

Is it possible that the tweet is authored by someone who speaks one of your country’s dialects?

- 1,120 sentences.
- with geolocations uniformly distributed across 14 countries.
- 3 annotators from 9 different countries (total of 27)

Sentence	Valid in
وين يلعب هذا ما شفته	Algeria🇩🇿, Palestine🇵🇸, Yemen🇾🇪

Abdul-Mageed, Muhammad et al. “NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task.” ArabicNLP 2024.

IFF an annotator labels the tweet as written in one of their country-level dialects.

Please evaluate the Level of Dialectness of each tweet as:

L0 Sound MSA

L1 Formal Colloquial or Colloquial-influenced MSA

L2 Natural/Ordinary Colloquial

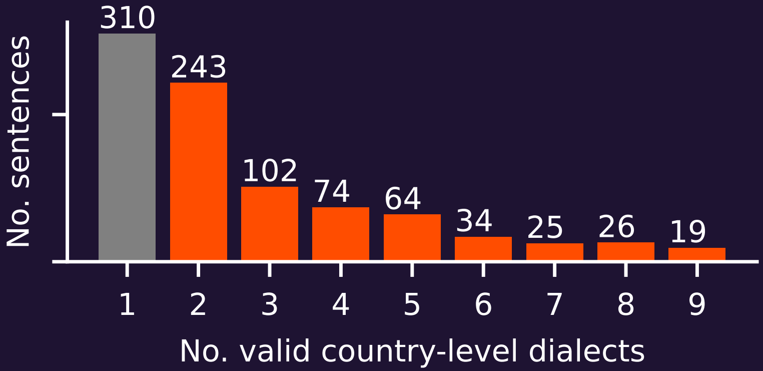
L3 Informal (or Vulgar) Colloquial

Note: The levels and their descriptions were provided in Arabic.

Country	N valid	Krip. α
Algeria	333	0.66
Morocco	230	0.74
Tunisia	189	0.75
Egypt	353	0.82
Sudan	393	0.66
Palestine	375	0.68
Syria	475	0.79
Iraq	271	0.73
Yemen	454	0.50

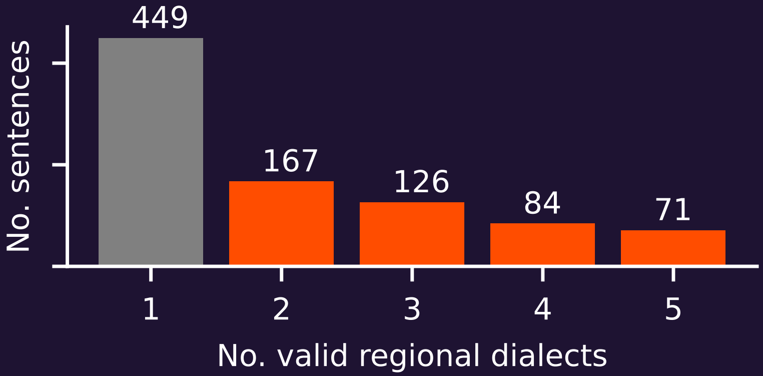
✅ Improved alpha scores than AOC-ALDi.

Multilabel samples in NADI 2024?



💡 All samples but 310 are multi-dialect (country level).

Multilabel samples in NADI 2024?



💡 > 50% of samples are valid in multiple regions.

❌ Not just because of within-region similarities!