THE UNIVERSITY *of* EDINBURGH

Improving the Modeling of Arabic Varieties in NLP

Amr Keleg

Doctor of Philosophy

Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh

2026

Abstract

Natural Language Processing (NLP) systems generally focus on supporting standardized varieties of languages. Developing systems for a non-standardized variety (dialect) requires finding/selecting samples in this variety, to create customized mixtures of pretraining data or to develop dialect-specific benchmarks. To this end, Dialect Identification (DI) is typically employed. This thesis explores some limitations of the long-standing approach that frames DI as a single-label classification task, where each sentence is linked to a single dialect. I specifically focus on Arabic, a language with a rich diversity of regional dialects. Arabic also exists in a diglossic state where two varieties co-exist within the same speaking community—Modern Standard Arabic (MSA) and local varieties of Dialectal Arabic (DA).

The thesis's main contributions are twofold. First, the different levels between *pure MSA* sentences and highly colloquial sentences are operationalized as a continuous variable—in range [0,1]—termed *Arabic Level of Dialectness (ALDi)*. ALDi estimation is modeled as a regression task, with a fine-tuned BERT-based model achieving an RMSE of 0.18. Second, Arabic Dialect Identification (ADI) is reframed as a multi-label task, where the validity of sentences in the different regional varieties is independently assessed. This is based on finding that in 66% of the cases where a single-label ADI model made errors, both its predictions and the gold-standard labels were valid.

Accordingly, each sentence has a set of regions in which it is valid, and an ALDi score to indicate how it diverges from MSA. By definition, MSA sentences are expected to be labeled as valid in all the considered regions, with almost-zero ALDi scores. Following the newly proposed framing, I created the first multi-label ADI dataset of 1,120 sentences (tweets), labeled by 33 annotators from 11 Arab countries, with ALDi ratings. This dataset allowed for investigating some widely adopted assumptions about Arabic. For instance, I show that Arabic dialects overlap considerably at both the country and regional levels. Additionally, the conscious *Dialect Level* choice that Arabic speakers make—operationalized as ALDi—is a better predictor of the number of dialects in which a sentence is valid than its length. Lastly, signs of systematic differences in the ALDi ratings provided by speakers of different dialects for the same sentences show the need for further investigations of ALDi's annotation.

ALDi is a valuable variable for many applications. For instance, I used it to identify stylistic differences in Arab presidents' speeches—previously only possible through qualitative analysis. For tasks requiring data annotation, I found that high-ALDi samples need to have a higher priority of being routed to speakers of the samples' dialects.

Lay Summary

Have you interacted with devices or programs—like *Alexa*, *Siri*, or even *Google Search*—and found yourself carefully choosing the terms you use to make sure these systems do not utterly fail? I will try to explain why these failures could happen.

For these programs to accept inputs in a language, they must first be exposed to large amounts of data in that language—textual data for this thesis. Books and newspapers are typically written in a standardized language. To collect dialectal sentences, researchers have been building *Dialect Identification (DI)* tools that can automatically infer the dialect of sentences on the internet. Current DI tools in general and Arabic DI tools in particular (the focus of this thesis) do not perform well.

Arabic is a language with many diverse dialects. Arabs have local dialects, but share a common standard dialect termed *Fus-ha* فصحى. This thesis tries to remedy the limitations of current Arabic DI tools.

First, these DI tools consider a speaker to be either using *Fus-ha* or their dialect. Hence, they are unable to determine how Arabic speakers use different degrees of mixing between *Fus-ha* and their dialect according to the situation. I define the *Arabic Level of Dialectness (ALDi)* score to differentiate between these different speaking styles. Another limitation is that the DI tools assume a sentence should be assigned to only a single dialect.

For the first limitation, I introduced a tool that can automatically estimate a sentence's ALDi score. To show its effectiveness, the ALDi estimation tool was used to automatically identify the various styles employed by Arab presidents in their speeches.

To investigate the two aforementioned limitations further, I collected 1,050 non-*Fus-ha* tweets from different Arab countries. I then asked 33 participants from 11 Arab countries to guess if these tweets could have been written by speakers originating from each participant's respective country. If they guessed *Yes* for a tweet, then they assigned the tweet an Arabic Level of Dialectness (ALDi) score, which rates how different the tweet is from *Fus-ha*.

I first found that 75% of the tweets could have originated from two countries or more. I also found that the participants, to some extent, agree on rating how different the tweets are from *Fus-ha*, especially for participants from neighbouring geographical regions.

In conclusion, I show that a DI tool should identify all the dialects from which a sentence could have originated, and ALDi scores can be used to show how different the sentence is from *Fus-ha*.

Acknowledgements

I grew into the habit of reading the *Acknowledgments* section of theses, and here I am finally writing mine! Reflecting on what I achieved by the end of my PhD, I think I could not have wished for more than I had. That said, I would not be where I currently am without the help of many people, and I will try to mention some of them as a small token of appreciation for that.

First, Walid and Sharon, I will forever be grateful for all the invaluable comments, tips, feedback, and guidance you have given me throughout this four-year journey. To Walid, you have always believed in me, sometimes more than I believe in myself. You treated me like a younger brother of yours, and I know I was not always an easy brother to deal with. You showed me the way to improve, encouraged me to explore and dream, gave me the chance to develop, and listened with full intent to my ideas and concerns. Your enthusiasm for research is exemplary and was crucial in times of doubt. To Sharon, I aspire to one day reach your levels of dedication, discipline, organization, and commitment. I learned what 'leading by example' means, thanks to you. You showed me how to pay attention to details, avoid overclaiming, and always aim to be concise and clear. It was a pleasure to have you as my supervisor.

Second, I had the opportunity to be surrounded by great researchers and dear friends from the *CDT in NLP*, *SMASH*, and, more recently, *Agora*. It is remarkable how human connections evolve. Thanks, Aida, Anna, Sandrine, Steph, Anil, and Ibrahim, for being kind to me during the early stages of this journey. Thanks, Henry, for your valuable feedback and the enthusiasm you showed in my early research. Danyang, Wanqiu, and Matthias, it was a pleasure working with you on the group project under the supervision of Bonnie Webber. To Matthias, I hope you would not mind that I involuntarily considered you my mentor and dear friend. Dialra and Silviu, you were two exceptional researchers and kind human beings. To Silviu, I hope you will forgive me for the sleepless nights we had in EMNLP 2022! To Youcef and Youssef, thanks for generously accepting me as a friend. Georgios, I am glad to have met you in Croy, and I am sure we will cross paths again soon. To Bálint, Gautier, Jonas, Oli, Nick Ferguson, Nick Sanders, and Coleman, sharing the office with you allowed me the chance to get to know you on a deeper level. To Coleman, it is strange how we have sat next to each other even after we moved offices, and how our connection has grown stronger over the years. To Nick Ferguson, you are a true example of what a gentleman is. I am now a Cityzen, thanks to you first and Marmoush second. Thanks, Rohit, Agostina, Verna, and Guillem for being my conference pals. Getting to know you was one of the highlights

of attending those conferences. I hope I had the chance to talk to more members of the CDT in NLP, but I am glad I crossed paths with you, Nickil, and Argyrios. You are a unique pair of friends. A special recognition also goes to Sally Galloway for all her efforts, help, and kindness.

I would also like to thank the participants who assisted with the various annotation tasks I conducted, as well as the members of the Arabic NLP community for their assistance at different stages of this thesis, whether through discussions or by generously sharing their datasets. Thanks, Nizar Habash, Nora Alturayeif, Manal Alshehri, Hamdy Mubarak, Nuha Albadi, Nedjma Ousidhoum, and Hala Mulki. Lastly, I am glad I had Adam Lopez as a member of my annual review board.

Throughout the years, I was honored to be a volunteer for *ACE IT* (special thanks to Campbell and Iain) and *The Meadows Community Garden*. Both communities welcomed me with warmth and understanding, exposing me to new experiences and teaching me lifelong lessons. Lastly, I would like to thank all my friends, past colleagues, and mentors for being a significant part of my journey.

Lastly, I am grateful for my family. Dad, thanks for everything you did. You put everything you could and more into getting us where we are now, and I want you to know that I will never be able to repay you for that. Novy, you have been, are, and I wish you would continue to be my number one fan. You have always believed in me, allowing me to explain the nitty-gritty details of topics that are utterly irrelevant to you while showing interest and encouragement. Mera, you have always been the person I have leaned on and the guide in my life. You have shared my celebrations, rants, concerns, fears, sorrows, and successes. Mammah, your kind words and unconditional love will always stay with me.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Amr Keleg)

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

To achieve certainty and control, we (humans) are in continuous pursuit of understanding the complex objects and systems of our world. Identifying different features to taxonomize the diverse objects (e.g., the various breeds of birds) provides a helpful way to define and identify these objects. Human languages represent another example of our world's fascinatingly diverse and complex systems. Similarly, taxonomizing human languages helps us situate the similarities and differences between them. This entails drawing boundaries—sometimes arbitrarily—to define the different languages and the varieties within each language. To build Natural Language Processing (NLP) models, their developers need to decide the languages and the varieties of each language to represent in their training and evaluation data. This decision relies on adopting taxonomies of language varieties for each language, which again requires finding a way to define boundaries between the different varieties. Given that NLP models still fail to generalize beyond the set of language varieties that they are designed to support, the underlying assumptions of the consulted taxonomies directly impact the ability of these models to represent the different varieties of the language.

Throughout this thesis, I show that there is room to improve how our NLP tools represent the variation within languages beyond supporting their standard varieties by embracing more nuanced linguistic/sociolinguistic theories. As a case study, this thesis focuses only on how the varieties of Arabic are taxonomized to develop NLP models that analyze textual inputs in some of these varieties. Nevertheless, my findings could still be helpful for other languages and modalities (e.g., speech inputs). Similar to many other human languages, Arabic is not a single monolithic language. To the

contrary, it is known for its diverse set of varieties. The wide geographical area over which its speakers are distributed is one of the reasons for the rise of local varieties of Dialectal Arabic (DA). Before the rise of digital communication and social media, these varieties of DA were mostly spoken. Modern Standard Arabic (MSA) is the standardized variety of the language, traditionally being the variety for written work (e.g., books and newspapers) (Habash, 2010). However, the different varieties of DA are increasingly written online. Early Arabic NLP models were built for MSA (e.g., Diab et al., 2004; Diab, 2004; Habash and Rambow, 2005; Abdul-Mageed and Diab, 2011). Unsurprisingly, these models could not, by default, support DA inputs. It did not take too long for researchers to aim at building models that can support some of these dialectal varieties (e.g., Chiang et al., 2006; Habash et al., 2012b; El-Beltagy, 2016). In this process, different taxonomies of the Arabic varieties were sought and used to define specific varieties that the new models would support. Most of these taxonomies rely on geography (e.g., borders between countries) to define the different varieties (Habash, 2010; Abdul-Mageed et al., 2020b). Moreover, simplifying assumptions were made when adopting the taxonomies to develop NLP models.

I identify two major limitations of the currently adopted taxonomies of Arabic varieties, used to build NLP models for textual inputs. The first limitation is **Binarization**, which assumes that an Arabic speaker either uses pure MSA or a variety of DA, grouping all dialectal sentences into a single category (i.e., as non-MSA), while some sentences are clearly more dialectal than the others. The second limitation is **Disjointedness**, which assumes that the Arabic varieties (sometimes considering MSA one of these varieties) are mutually exclusive. Accordingly, a single Arabic variety is assigned to each piece of text. Both these limitations oversimplify the reality. This, in turn, limits the capabilities of the NLP models that adopt them. The two main themes of this thesis can be summarized into the following hypotheses:

Hypothesis 1 *Written Arabic sentences of a specific dialect exist over a spectrum of Dialect Levels, with pure Standard Arabic and highly Colloquial Arabic as its extremes.*

Hypothesis 2 *A substantial number of written sentences are valid in more than one variety of Arabic, especially when fine-grained geographical taxonomies of these varieties are used (e.g., country level or city/province level).*

The results of studying the two hypotheses can also be relevant to languages other than Arabic. For instance, I provide a new modeling of dialect levels that can better mitigate the limitations of using token-level code-switching tagging, which fails to

model the token-level divergence from Standard Arabic. This limitation could also apply to other languages that share a single standard variety. Moreover, parallels of the findings related to the validity of a sentence in multiple varieties are expected to exist for languages other than Arabic.

1.2 Research Questions

Following the two hypotheses in §1.1, this thesis addresses the research questions related to improving the representation/adoption of the two types of variation within Arabic in NLP models: *intraspeaker variation* (i.e., linguistic differences in the speech of the same speaker in different contexts) and *interspeaker variation* (i.e., linguistic diversity between different speakers). Lastly, I show how alleviating the *Binarization* and *Disjointedness* limitations is crucial for improving the modeling of DA, in relation to capturing intraspeaker variation that could not be achieved using the existing methods, by studying how they can improve on earlier guidelines for routing dialectal samples to Arabic-speaking annotators.

Theme #1 - Binarization Split of Arabic into MSA and varieties of DA

RQ1 *How can the concept of Dialect Levels be operationalized in a way that can be effectively estimated?*

RQ2 *What are some applications of automatically estimating Dialect Levels, in text analysis and data annotation?*

RQ3 *Do speakers of different dialects (varieties) share similar perceptions of a sentence's Dialect Level?*

Theme #2 - Disjointedness of Arabic Varieties

RQ4 *How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?*

RQ4.1 *Do most sentences valid in multiple dialects have a short length?*

RQ4.2 *Are current ad-hoc lists of dialectal lexical terms distinctive enough to ensure that a sentence is valid in a dialect and not valid in other dialects if it contains a term of this dialect's list of lexical terms?*

The following section explains how the chapters of this thesis progressively investigate and address these questions. Earlier chapters investigate each hypothesis

independently, while the later chapters study multiple research questions related to the two hypotheses in parallel.

1.3 Thesis Structure

The rest of this thesis is structured as follows:

Chapter 2 - Background An overview of the different varieties of Arabic, the potential reasons for this variation, and examples of how the variation is manifested in the various linguistic aspects. Afterward, I will discuss the linguistic theories proposed to define the different sources of variation, their limitations, and how the different Arabic varieties were represented and modeled in NLP.

Chapter 3 - Arabic Level of Dialectness This chapter introduces the first theme of the thesis, related to the *Disjointedness* limitation, studying **Hypothesis 1**. The concept of Arabic level of Dialectness (ALDi) is formalized to quantify how much a sentence diverges from Standard Arabic, after examining the few attempts to model this intraspeaker variation computationally. An effective regression-based model is presented to address (*RQ1*), demonstrating its ability to generalize to different Arabic dialects.
Published in: (Keleg et al., 2023)

Chapter 4 - Applications of Arabic Level of Dialectness Two different applications are presented to demonstrate how ALDi provides more nuanced information than ADI, addressing (*RQ2*). The first shows how ALDi can automatically identify the different styles employed in the speeches of different Arab presidents, which is potentially useful in future sociolinguistic studies. The second concludes by refining previously proposed annotation guidelines for Arabic datasets, which recommended routing dialectal sentences to speakers of the samples' dialects. More specifically, it recommends that sentences with high ALDi scores should be prioritized and routed to native speakers of the samples' dialects, which are easier to identify for these high-ALDi samples.
Published in: (Keleg et al., 2023, 2024)

Chapter 5 - Limitations of Single-label Arabic Dialect Identification (ADI) This chapter presents the first attempt to quantitatively investigate the pervasiveness of sentences valid in multiple dialects (**Hypothesis 2**). First, I explain how assigning a single dialect label to a sentence valid in multiple dialects impairs the accurate evaluation of Arabic Dialect Identification (ADI) systems. This hinders progress toward building

better ADI models. Then, the errors of a single-label ADI system are manually analyzed to have a more detailed understanding of the limitations of the single-label framing of ADI, providing a preliminary study of (*RQ4*).
Published in: (Keleg and Magdy, 2023)

Chapter 6 - Redesigning Arabic Dialect Identification (NADI 2024) The results of the previous chapter highlighted the need for framing ADI as a multi-label classification task. Consequently, this chapter describes the joint efforts to build the first multi-label ADI evaluation dataset, introduced as part of the NADI 2024 shared task, where each sentence is labeled by 33 annotators representing 11 different Arab countries (3 each). Each sample is also labeled for its *Dialect Level* whenever an annotator labeled it as valid in their country-level dialect. Additionally, the chapter summarizes the multi-label ADI and ALDi estimation systems developed by the participating teams in the shared task. Lastly, *RQ1* and *RQ4* will be revisited using the newly introduced dataset.
Published in: (Abdul-Mageed et al., 2024)

Chapter 7 - Revisiting Common Assumptions about the Arabic Dialects in NLP This chapter presents *MLADI*, an extension of the dataset introduced in chapter 6. The newly introduced dataset allowed for investigating *three* widely held assumptions about the Arabic dialects, related to (1) the disjointedness of the Arabic dialects when grouped into macro-regional/country-level ones (*RQ4*), (2) the relation between the sentence's length and the number of dialects in which it is valid (*RQ4.1*), and (3) the ability to curate distinctive dialectal words to infer the dialect of sentences containing any of them (*RQ4.2*). Additionally, the chapter investigates (4) if speakers of different dialects share similar perceptions of a sentence's *ALDi* (*RQ3*).
Published in: (Keleg et al., 2025)

Chapter 8 - Conclusion The chapter summarizes the findings of the two main themes investigated in the thesis, related to the *Disjointedness* of the Arabic varieties, and the *Binarization* of Arabic sentences into MSA and varieties of DA. I then discuss some limitations and suggestions for mitigating them in future work.

1.4 Thesis Outcomes

Published Papers

1. Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic Level of Dialectness of Text. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10597–10611, Singapore.

2. Amr Keleg and Walid Magdy. 2023. Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification. In Proceedings of ArabicNLP 2023, pages 385–398, Singapore (Hybrid).

3. Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In Proceedings of the Second Arabic Natural Language Processing Conference, pages 709–728, Bangkok, Thailand.

4. Amr Keleg, Walid Magdy, and Sharon Goldwater. 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), pages 766–777, Bangkok, Thailand. Received an Outstanding Paper Award.

5. Amr Keleg, Sharon Goldwater, and Walid Magdy. 2025. Revisiting Common Assumptions about Arabic Dialects in NLP. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria.

Released Artefacts

1. **Sentence-ALDi model**: the ALDi estimation model introduced in chapter 3 hosted on HuggingFace. The model was downloaded more than 210,000 times as of January 2026.
   Link: `https://huggingface.co/AMR-KELEG/Sentence-ALDi`

2. **ALDi Estimation Space**: a HuggingFace space that allows testing the *Sentence-ALDi* model to automatically estimate the ALDi scores of sentences entered by the users.
   Link: `https://huggingface.co/spaces/AMR-KELEG/ALDi`

3. **MLADI Dataset**: the first multi-label Arabic Dialect Identification dataset, with country-level dialect labels and ALDi scores, manually provided by 33 annotators from eleven different Arab countries.
   Access Form: `https://forms.gle/gdgTToxG2tH5xT27A`

4. **MLADI Leaderboard**: a leaderboard to benchmark the performance of multi-label ADI systems, using the MLADI dataset.
   Link: `https://huggingface.co/spaces/AMR-KELEG/MLADI`

Other Papers not Included in the Thesis

This thesis focuses on the linguistic aspects of Arabic varieties, studying how to represent them more effectively in NLP. These regional varieties are a clear manifestation of the cultural variation within the Arab world, another aspect that I have also been pursuing, publishing the following three papers:

1. Amr Keleg and Walid Magdy. 2022. SMASH at Qur'an QA 2022: Creating Better Faithful Data Splits for Low-resourced Question Answering Scenarios. In Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, pages 136–145, Marseille, France.

2. Amr Keleg and Walid Magdy. 2023. DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 6245–6266, Toronto, Canada.

3. Amr Keleg. 2025. LLM Alignment for the Arabs: A Homogeneous Culture or Diverse Ones. In Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025), pages 1–9, Albuquerque, New Mexico.

Chapter 2

Background

This chapter elaborates on the two dimensions of variation in Arabic: interspeaker variation and intraspeaker variation. While the rest of this thesis focuses on the modern varieties of Arabic, having some basic knowledge of the language's history will allow for a better understanding of its current state. Afterward, I discuss how the dialectal variations are manifested in different linguistic aspects, demonstrating the *interspeaker variation* in the different Arabic-speaking regions. For the *intraspeaker variation* within the Arabic-speaking communities, I describe three different linguistic theories that have been proposed to model it. With these two types of variation (interspeaker and intraspeaker) in mind, I then discuss how the different Arabic varieties are currently represented in NLP models that operate on textual inputs. I conclude by identifying the research gaps, paving the way for the following chapters in which I elaborate on how the language variation within Arabic could be better represented in our NLP models.

2.1 The Arabic Varieties

Arabic is a Semitic language, alongside other languages such as Amharic, Aramaic, and Hebrew. It is spoken by more than 420 million people all over the world (Bergman and Diab, 2022), and is the official language of 22 Arab countries. Unlike many modern languages, Arabic speakers can still fairly comprehend the rich literature of the language, from poetry to scientific and religious books dating back more than 16 centuries. This fact was the idea on which an online show[1] was based, where a Bahraini and an Omani chefs followed food recipes from a book written during the Abbasid dynasty in the 9th century, an example of which is provided in Figure 2.1.

[1] `https://www.youtube.com/watch?v=M2p4wbjeJfs`

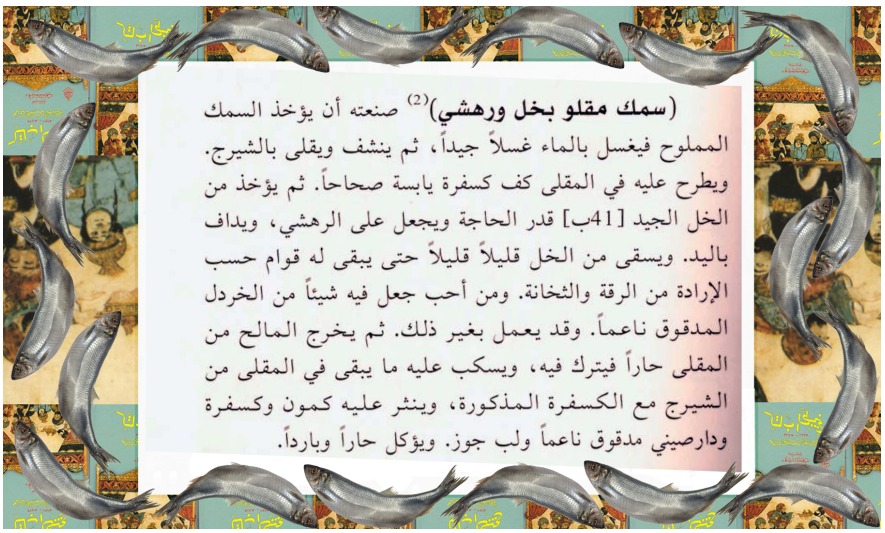


Figure 2.1: A sample recipe of fried fish mixed with vinegar and ground sesame, from *The cookbook* (كتاب الطبيخ) by (أبو إسحاق إبراهيم بن المهدي العباسي), which dates back to the 9th century (i.e., the Abbasid dynasty). **Source**: https://www.vice.com/ar/article/مال-أو ل-و صفة-السمك-المملو ح-من-كتاب-طبيخ/

In NLP literature, a distinction is made between three main varieties of the language (Habash, 2010; Darwish and Magdy, 2014). (1) *Classical Arabic (CA)* refers to the language of *Qura'n*—the holy book of Islam, believed to be revealed to *Prophet Muhammad (Peace be upon him)* in the 7th century—and the language of poetry that dates to the pre-Islamic era. (2) *Modern Standard Arabic (MSA)* is believed to be a descendant of CA, which keeps most of its grammatical properties while adapting new terms and phrases into its vocabulary to keep up with the advancements of modern life. (3) Varieties of *Dialectal Arabic (DA)* are a set of local varieties that exist throughout the Arab world.

Until the rise of digital communication and social media, MSA was the variety of literary works, such as books, newspapers, and official documents. It also has a standardized orthography and documented grammar rules, and is the variety taught in schools. Conversely, the varieties of DA are mostly spoken and do not have a standardized orthography. Nowadays, these varieties of DA are being used in online communications, reshaping the relationship between the language and its speaking communities. The following subsections discuss the three main varieties of Arabic.

2.1.1 Classical Arabic and Modern Standard Arabic

Defining CA and MSA as two varieties assumes that each is one homogenous variety. This is more problematic for CA as it used to refer to the varieties of Arabic spoken over a long period, dating back to even before the 7th century. Moreover, it assumes a single CA variety (i.e., a source variety that is the ancestor of all modern varieties), which is inaccurate given that dialectal variations existed in the Arabian Peninsula even before the dawn of Islam (Holes, 1995, p.24). These dialectal variations could be a potential reason for the presence of seven/ten canonical recitations of the Qur'an (Denny, 1989). They are also referenced in some interpretations of a saying by Prophet Mohammad (shown below[2]), to explain what the seven *modes* (أحرف - ÂHrf) of the Qur'an are. It needs to be mentioned that some also believe that these dialectal variations were not significant (Alghamdi, 2021).

> وعن ابن عباس رضي الله عنهما قال: إن رسول الله صلى الله عليه وسلم قال: «أقرأني جبريل على حرف فراجعه فلم أزل استزيده ويزيدني حتى انتهى إلى سبعة أحرف» . قال ابن شهاب: بلغني أن تلك السبعة الأحرف إنما هي في الامر تكون واحدًا لا تختلف في حلال ولا حرام. - متفق عليه (الألباني)
>
> Ibn 'Abbās reported God's messenger as saying, "Gabriel taught me to recite in one mode, and when I replied to him and kept asking him to give me more, he did so till he reached seven modes." Ibn Shihāb said he had heard that those seven modes are essentially one, not differing about what is permitted and what is prohibited. (Bukhārī and Muslim.)

Fus-ha In reality, Arabic speakers use the term *Arabiya Fus-ha* عربية فصحى or just *Fus-ha* فصحى to refer to both CA and MSA. MSA is a term popularly used in the NLP literature, and not a concept commonly known by Arabic speakers (Holes, 1995; Parkinson, 1991).

> "MSA is merely a handy label used in Western scholarship to denote the written language from about the middle of the nineteenth century, when concerted efforts began to modernise it lexically and phraseologically. Most western scholars refer to the formal written language before that data, and par excellence before the eclipse of Arab political power in the fifteenth century as *Classical Arabic*". (Holes, 1995, pp. 4,5)

In his book, Alghamdi (2021) argued that many of MSA's syntactic structures and stylistic features are foreign to the language. More specifically, they result from the waves of translations of scientific materials written in foreign languages (mainly English

[2]**Source**: https://sunnah.com/mishkat:2214

and French) into Arabic. While these translated books have helped in improving the people's access to knowledge, these translations were strongly affected by the syntax and structure of the languages in which they are written initially —a phenomenon that is termed as *Translationese* (Gellerstam, 1986 as cited in Koppel and Ordan, 2011). Throughout his book, Alghamdi (2021) identified multiple aspects (e.g., sentence structure) by which MSA stylistically matches English and diverges from CA. One of the examples he provided is shown in Figure 2.2. Although both the original and backtranslated texts would be considered *Fus-ha* فصحى, an Arabic speaker would perceive the original text as of higher eloquence than the backtranslated one, with the backtranslated version being more stylistically similar to the Arabic variety found in modern literature.

Original Text	مما أنعم الله به على أهل الحجاز هذا البن
Translation	One of the blessings of Allah on the inhabitants of Hijaz was coffee beans
Backtranslation	إحدى نعم الله على سكان الحجاز هى حبوب البن

Figure 2.2: An excerpt from a book written in the 17th century (i.e., before the wave of translating books from foreign languages into Arabic), with its English Translation and a backtranslation that matches the translation's stylistic structure and expressions. **Source:** (Alghamdi, 2021, p. 164)

These stylistic differences might explain why Arabic speakers can still differentiate between MSA and CA, even when they refer to them both as (*Fus-ha*). For instance, when Parkinson (1991) asked an Egyptian woman to respond to the questions of a proficiency test in *Fus-ha*, she responded in a way that implies that she makes the distinctions above:

> "Do you want me to use fusha or fusha fusha?", apparently implying "Should I write simple, but grammatically correct sentences, or should I make them flowery and fancy as well (i.e., should I try to adopt a classical style)?" (Parkinson, 1991, p. 34)

MSA In the NLP literature, MSA is discussed as an Arabic variety commonly shared across the different Arab countries (Chiang et al., 2006). Unlike English, where there is no standard variety used in all English-dominant countries, Arabic speakers are claimed to share a single standardized form of the language. To computationally investigate this claim, automatically identified MSA tweets from 18 different Arab countries were curated as part of the NADI 2021 shared task (Abdul-Mageed et al., 2021b). The

participating teams were tasked to predict the geolocated country label of each tweet. The top-performing team achieved a macro-averaged F1-score of 22.38, demonstrating difficulty in distinguishing between MSA tweets from different countries. Moreover, the non-zero F1 scores might be attributed to the models' ability to identify spurious correlations such as local named entities from other countries, which should not be tied to any dialect (Abdul-Mageed et al., 2020b; AAlAbdulsalam, 2022). Unsurprisingly, a couple of teams concluded that the task is practically impossible (El Mekki et al., 2021; Issa et al., 2021; Nayel et al., 2021). However, an Arabic speaker's native dialect might still be impacting their perception of what is in MSA and what is colloquial, as later demonstrated in chapter 7.

2.1.2 Modern Varieties of Dialectal Arabic

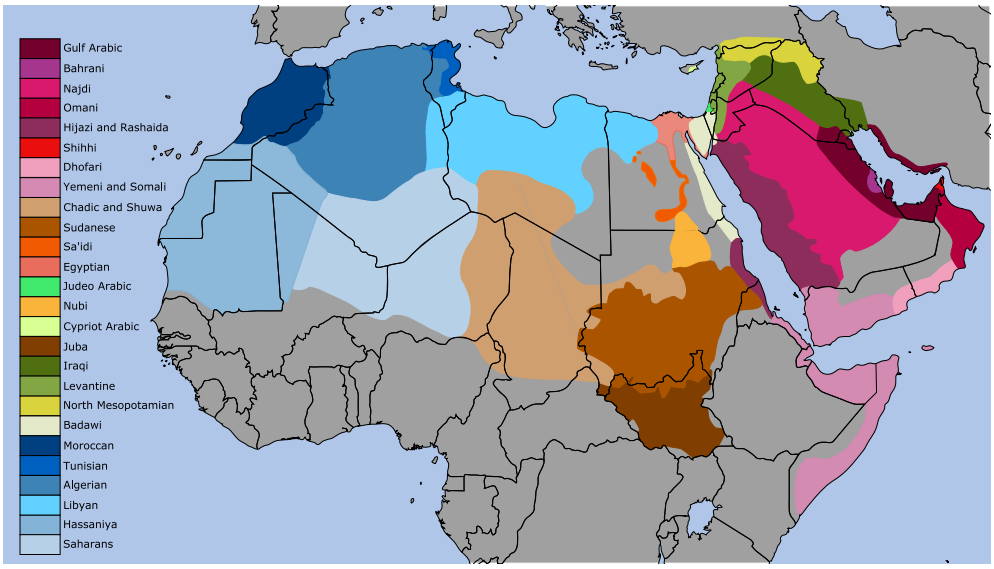


Figure 2.3: A map of the Arabic varieties shown in (Schmitt, 2020).

The local varieties of DA reflect the rich interspeaker variation across the Arab world. Examples of these varieties are those of Cairo in Egypt, Rabat in Morocco, and Baghdad in Iraq. Figure 2.3 provides an approximate visualization of dialect areas (i.e., areas that are assumed to share the same dialect). The different taxonomies of these varieties used in NLP will be discussed later in §2.5.1. Some researchers argue that the varieties of DA should be considered as independent languages (Kaye and Rosenhouse, 1997). The differences between the varieties of DA are sometimes compared to the differences between the Romance languages (Chiang et al., 2006), while their relation to MSA is compared to the Romance languages' relation to Latin (Alshargi et al., 2019). In

a similar fashion, Zaidan and Callison-Burch (2014) compared the relationship between the DA varieties to that between the North Germanic languages (Norwegian, Swedish, and Danish) and the West Slavic languages (Czech, Slovak, and Polish).

These analogies can be useful for non-Arabic speakers to imagine the relation between the Arabic varieties. However, the idea of considering them as independent languages just to abide by the norms of other languages is, at the very least, naive. The question of whether DA varieties should be considered as dialects of one language or as independent languages is multifaceted. For instance, considering the different varieties of DA as independent languages might result in the abandonment of MSA. This is problematic, as MSA could in some cases be a useful tool of communication between speakers of different Arabic dialects, as per the following:

> "In cases of dialectal contact of speakers from more widely separated areas the matter is a complex one, but depends basically on what the participants perceive as the minimum degree of switching to 'neutral' dialectal, MSA, or even 'hybridised' forms, which is necessary to ensure smooth communication in an appropriate style". (Holes, 1995, p. 5)

Despite being intriguing, considering the varieties of DA as dialects of Arabic or as independent languages is beyond the aims of this thesis, which focuses on improving how NLP models represent the different Arabic varieties, as opposed to linguistically studying the similarities and differences between these varieties.

Varieties of DA—sometimes even being mutually unintelligible (Abu Farha and Magdy, 2022; Bergman and Diab, 2022)—can diverge from MSA and each other in phonology, morphology, syntax, and semantics (Habash, 2010), as later explained in §2.2. In this subsection, I will provide some potential reasons for the rise of these local varieties of DA. More specifically, I will explain the impact of other local and foreign languages that were used across the current Arab world. Manifestations of the contact between Arabic and these languages are apparent in the DA varieties to different degrees. I refer the reader to (Holes, 1995, pp. 18-30) for a further discussion of the linguistic hypotheses that have been proposed to explain how Arabic was adopted in different regions where it was not originally spoken. These hypotheses discuss the **diachronic changes** in Arabic and its varieties.

Pre-Islamic Local Languages

Arabic was the language used in most of the Arabian peninsula, and some parts of Mesopotamia and the Levant (Holes, 1995). As indicated in Figure 2.4, other regions of

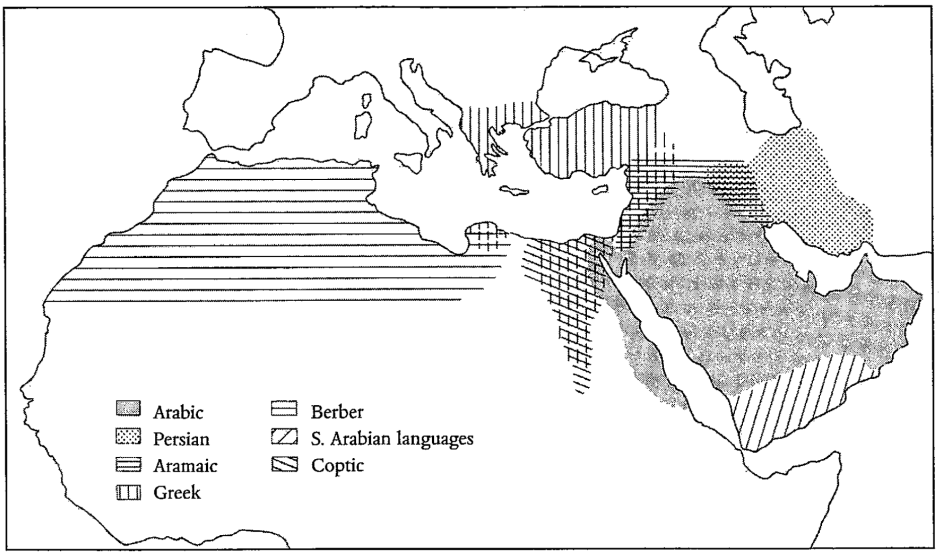


Figure 2.4: A map of the main languages used in different regions of the Arab world before the Islamic conquests. The map uses the term *Berber* that has negative connotations in Arabic, instead of *Tamazight*, a more appropriate term to describe the languages and people of west and central North Africa. **Source**: (Holes, 1995, p. 16)

the current Arab world spoke other languages: (1) Tamazight in western and central North Africa, (2) Coptic in Egypt and Sudan, (3) Greek in east North Africa and the Levant, (4) Aramaic in the Levant and Mesopotamia, (5) Persian in some parts of Mesopotamia, and (6) South Arabian languages in Yemen and Oman. The degree of impact of these languages on the modern varieties of DA varies. Their current usage across the contemporary Arab world is also variable. For instance, Egyptian Arabic has a few words that can be traced back to Coptic ones, such as أوطة *ÂwTħ* (tomatoes), coming from the Coptic word ⲟⲩⲧⲁ. Nowadays, Coptic is generally restricted to religious sermons in churches. Conversely, Tamazight had more impact on the lexicon of Moroccan Arabic, as in the word لالة *lAlħ* (lady). Tamazight is still widely used, to the extent that the standardized Tamazight is another official language of Morocco and Algeria. The same applies to Kurdish—a language of many dialects spoken in parts of Iraq and Syria—which is an official language of Iraq besides Arabic. The coexistence of Kurdish/Tamazight and Arabic in these countries suggests that they continue to influence each other to this day (Lahrouchi, 2018), in contrast to the influence of the Coptic language on modern Egyptian Arabic, which appears to have stagnated.

Judeo-Arabic dialects—varieties of Arabic spoken by Arab Jews—also demonstrate how different languages influence each other. Their speakers' knowledge of Hebrew impacted these varieties. An interesting example of how Hebrew impacted these dialects was that a modified version of the Hebrew script was adopted as a writing system for these Arabic varieties (Terner et al., 2020).

Colonial Languages in the Arab World

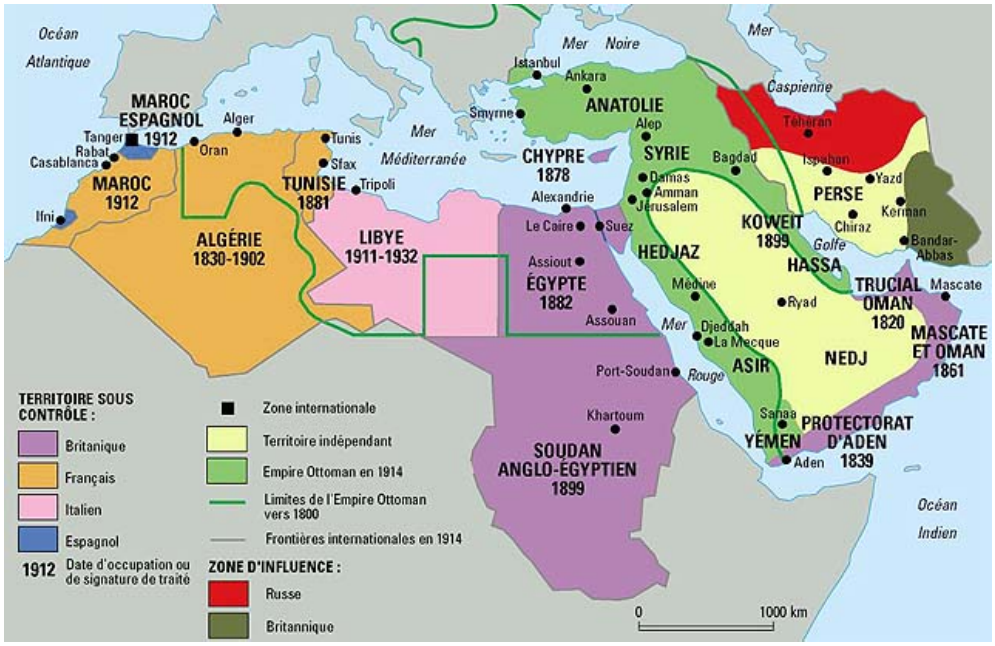


Figure 2.5: A map of the colonial powers that controlled the different regions of the Arab world at the eve of the First World War in 1914. **Source**: "The Middle East in 1914", by Philippe Rekacewicz (open access, August 1992) - https://mondediplo.com/maps/middleeast1914

Fast-forwarding to the 20th century, another set of foreign languages started to impact the varieties of DA, mostly a result of colonization as per Figure 2.5. It could be argued that the impact of these foreign languages depended on the strategies that different colonization powers used in the different regions, and the period for which each country was colonized. Despite being unambiguously vicious in nature, some colonial powers like the French brutally enforced the usage of French in the territories they occupied, in contrast to less enforcement by the British (Walters et al., 2023). This could be why the impact of French on some dialects (e.g., those of Morocco, Algeria, and Tunisia) is arguably larger than the impact of English on other dialects

(e.g., those of Egypt and Sudan). The following quote by John Stuart Mill—an English philosopher—provides an example of the strategies of British colonialism:

> "To suppose that the same international customs, and the same rules of international morality, can obtain between one civilized nation and another, and between civilized nations and barbarians, is a grave error. ... To characterize any conduct whatever towards a barbarous people as a violation of the law of nations, only shows that he who so speaks has never considered the subject". (Mill, 1865, p. 252)

Additionally, the cosmopolitan nature of some cities puts other foreign languages in contact with Arabic, such as Persian, Turkish, Greek, and Italian. For example, *Alexandria*, a coastal city in Egypt on the Mediterranean Sea, used to have a significant population of Greeks and Italians. It is not a surprise to find words in Egyptian Arabic like تورتة *twrtħ* (cake), كابوريا *kAbwryA* (crab),[3] which are quite similar to the respective Italian word *Torta* (cake), and the Greek word καβούρι (crab) respectively, demonstrating the impact of the contact between the different languages.

English and French are common second languages in different parts of the Arab world. Moreover, textbooks and references used in schools and universities could be in English or French. Mass media have also allowed for non-Arabic shows, magazines, and films to be widely broadcast. All of these factors might be why it is also common to find some Arabic speakers code-switching to English and/or French in their conversations (Cotterell et al., 2014; Hamed et al., 2020, 2025), another manifestation of these languages' influence on Arabic speakers.

2.1.3 Other Subvarieties of Dialectal Arabic

Linguists make another disctinctions between *Bedouin*, *Rural*, and *Sedentary* varieties. These varieties tend to be found in most, if not all, Arab countries. For instance, Egyptians are aware of the differences between the *Sedentary* varieties spoken in Cairo (the capital) and the *Rural* varieties spoken in Upper Egypt, known as *Sa'idi Arabic*. In this thesis, I mostly analyze dialectal data from online sources such as Social Media, on which varieties linked to prestige are more prevalent, and other varieties like *Sa'idi Arabic* are not well-represented (Mohamed Eida et al., 2024).

This section provides an overview of the different varieties of Arabic. However, these varieties do not exist in isolation, and the boundaries between the three broad

[3]The transliterations provided in this thesis follow the Habash-Soudi-Buckwalter (HSB) scheme (Habash et al., 2007).

varieties mentioned above are not crystal clear. We will explore the interaction between these varieties within the same Arabic-speaking community in the §2.3, after first explaining the significance of the variation on the different linguistic aspects.

2.2 The Different Manifestations of Variation

This section provides some examples of the differences between the Arabic varieties (mainly between MSA and varieties of DA) in various linguistic aspects. These differences between MSA and DA, exemplified below, and the fact that speakers commonly code-switch between the MSA, DA, and sometimes foreign languages (e.g., English, French), are a major challenge for Arabic NLP systems.

Phonology Different varieties of DA can have different phonology compared to that of MSA. For instance, MSA does not have a voiced labial explosive phoneme /p/, while some dialects adopt this phoneme, specifically for loan words. Variation also exists in the pronunciation of some consonants. A commonly used example is how the MSA consonant ق /q/ is realized as a glottal stop /ʔ/ in Egyptian and Levantine Arabic, and as /g/ in Gulf and Iraqi Arabic (Habash, 2010, p. 30). MSA has three long vowels and three corresponding short ones. Dialects can change or completely drop these short vowels (Habash, 2010, p. 31).

Orthography MSA is the standardized variety of Arabic, written in the Arabic script. Given the lack of standardized orthographies for the varieties of DA, Arabic speakers use their intuition to write dialectal terms based on pronunciation. Consequently, the same term could have multiple orthographic forms. For instance, Habash et al. (2018) identified 27 orthographic forms for writing the phrase *he does not say it* in Egyptian Arabic, as shown in Table 2.1. However, they also showed that the frequency of these forms differs (according to Google Search trends), indicating that some regularity exists in how dialectal words are written. For instance, the most common form مبيقولهاش consists of the following morphemes (م + ب + يقول + ها + ش). The morpheme يقول (says he) uses the grapheme ق, which is realized as a voiceless uvular stop /q/, matching the orthography and phonology of MSA. However, this phoneme is a glottal stop /ʔ/ in Egyptian Arabic, which is generally written as ء. Hence, terms with an MSA origin are sometimes written etymologically, even when the dialectal pronunciation differs from the standard pronunciation (Habash et al., 2012a).

While regional differences in the pronunciation of MSA terms are typically lost in writing due to the standardized orthography, individuals usually use non-standard orthography that matches their regional pronunciations (e.g., *Man* written as راجل *rAjl* instead of the standardized form رجل *rjl*. Notably, short vowels are orthographically realized as optional diacritics. These diacritics are rarely used in writing in DA, and are commonly dropped in non-religious MSA text as well (Habash, 2010, p. 32).

Arabic Orthography	Arabic Transliteration	Frequency
مبيقولهاش	*mbyqwlhAš*	≈ 26,000
ما بيقولهاش	*mA byqwlhAš*	≈ 13,000
مابقلهاش، مبقولهاش، مبقلهاش، ما بقلهاش، مابيقولهاش	*mAbqlhAš, mbqwlhAš, mbqlhAš, mA bqlhAš, mAbyqwlhAš*	≤ 10,000
مابقولهاش، ما بقولهاش، مبيقلهاش، ما بيقلهاش	*mAbqwlhAš, mA bqwlhAš, mbyqlhAš, mA byqlhAš*	≤ 1,000
مبٔلهاش، مابيئولهاش، ما بيئولهاش، مابيؤلهاش	*mbŷlhAš, mAbyŷwlhAš, mA byŷwlhAš, mAbyŵlhAš*	≤ 100
ما بيؤلهاش، مابئلهاش، مبيئولهاش، ما بيئلهاش، مابئولهاش، ما بئلهاش، ما بؤلهاش، مبئولهاش، مبيؤلهاش، مابؤلهاش، مبؤلهاش	*mA byŵlhAš, mAbŷlhAš, mbyŷwlhAš, mA byŷlhAš, mAbŷwlhAš, mA bŷlhAš, mA bŵlhAš, mbŷwlhAš, mbyŵlhAš, mAbŵlhAš, mbŵlhAš*	≤ 10

Table 2.1: The different ways of writing "he does not say it" in Egyptian Arabic with their corresponding frequencies according to Google Search trends. **Source**: (Habash et al., 2018). In reality, some forms are not fully accurate translations of the phrase. For instance, ما بقولهاش and مبقولهاش are actually in the singular first person and not the singular third masculine person. **Note:** The transliterations in this table follow the Habash-Soudi-Buckwalter (HSB) scheme (Habash et al., 2007).

Arabizi (also known as *Arabish* or *Franco Arabic*) is a romanized form of Arabic, where the letters are transliterated using the Latin script. There is not a 1-1 mapping between the Arabic letter and the English ones, with multiple variations of the system adopted by the speakers of different dialects. The lack of a standardized form of words in Latin script, even for MSA terms, makes the phonological features of dialects more noticeable, since the words are written using letters that mimic how they are pronounced.

Morphology Arabic is a morphologically rich language. A single token is typically composed of multiple morphemes. Additionally, clitics (proclitics and enclitics) also exist. Varieties of DA could have morphemes that diverge from the MSA ones. For

instance, the future clitic is س in MSA, but ه or ح in Egyptian Arabic (Darwish et al., 2014).

Lexicon as previously introduced in §2.1.2 and §2.1.2, the varieties of DA were influenced by contact with multiple languages.

Syntax MSA has a free word order, with diacritics on the final letter of the morphemes used to mark the grammatical cases. That said, a VSO is a common word order in MSA. In contrast, varieties of DA tend to drop the diacritics for a simplified phonology. Consequently, they employ a more restrictive word order to avoid syntactic ambiguity, with most of them having a preference for the SVO word order (El-Yasin, 1985) and (Aoun et al., 1994; Shlonsky, 1997 as cited in Zaidan and Callison-Burch, 2014).

Semantics Lexical overlap exists between MSA and DA varieties, and between the different varieties of DA. However, some terms shared between different varieties do not share the same semantic meaning (Aminian et al., 2015). These terms are called *False Friends* or *Faux Amis*. The semantic divergence could be extreme to the extent that the same term/phrase has positive connotations in one variety and negative connotations in another. As an example, يعطيك العافية in Levantine dialects means 'may you have good health'; however, in Moroccan it translates to 'go to hell' (Bergman and Diab, 2022).

Speaking Production While Arabs can understand and read the standard language (MSA and CA to a great extent, as exemplified in Figure 2.1), spontaneously speaking in the standard language is not a natural task for most of them (Chiang et al., 2006; Habash and Rambow, 2006). Varieties of DA are generally used in everyday communications, especially in spontaneous situations.

2.3 Linguistic Theories of Intraspeaker Variation within Arabic

Given the complexity of variation within the same Arabic-speaking community, multiple linguistic theories have been proposed. In this section, I will describe three main theories of intraspeaker variation within Arabic, namely: *Diglossia*, *Diglossic Code-switching*, and *Dialect Levels*.

1) Diglossia is a language state in which two varieties of the language co-exist within the same speaking community, a high variety linked to higher prestige and a low variety perceived to be of lower status (Ferguson, 1959). These two language varieties tend to be used for different functions and in different situations, as indicated in Figure 2.6. Arabic is one of the four languages that Ferguson (1959) identified as examples of languages that exist in a state of diglossia.

	H	L
Sermon in church or mosque	x	
Instructions to servants, waiters, workmen, clerks		x
Personal letter	x	
Speech in parliament, political speech	x	
University lecture	x	
Conversation with family, friends, colleagues		x
News broadcast	x	
Radio "soap opera"		x
Newspaper editorial, news story, caption on picture	x	
Caption on political cartoon		x
Poetry	x	
Folk literature		x

Figure 2.6: The language variety normally used in different possible situations according to Ferguson (1959, p. 329). **Note**: For Arabic, the high variety (*H*) refers to *MSA*, and the low variety (*L*) refers to *the local variety of DA*.

Limitations While *Diglossia* could provide a useful framework for understanding the intraspeaker variation within Arabic—as the co-existence of two varieties (MSA and DA)—it adopts an unrealistic assumption, which I refer to as the *Binarization* limitation. More specifically, diglossia assumes that Arabic speakers either use pure MSA (the high variety) or pure DA (the low variety) and that each variety serves different functions. Even in situations like "sermons in church or mosque" mentioned in Figure 2.6, where MSA is expected to be the only variety used, DA would sometimes be employed for specific cases like rephrasing ideas in a colloquial way to make them easier to understand. Ferguson himself had his doubts that the concept of Diglossia could be simplistic as indicated by the following quote: "*Perhaps the collection of data and more profound study will drastically modify the impressionistic remarks of this paper, but if this is so the paper will have had the virtue of stimulating investigation and thought*" (Ferguson, 1959, p. 340). He also acknowledged the presence of a third variety—an unstable intermediate form of Arabic—that he assumed is a spoken variety used in semiformal or cross-dialectal situations, that relies on a mixture of classical and colloquial vocabulary, and has lots of signs of colloquial morphology and syntax. This third variety could not be explained according to the theory of *Diglossia*.

2) Diglossic Code-switching is another theory proposed to model the intraspeaker variation. Under this framework, Arabic speakers are assumed to code-switch between MSA and DA either intersententially or intrasententially. *Diglossia Code-switching* is the term used to describe this subcategory of code-switching to distinguish it from code-switching between two different languages.

Conforming to this theory, Abdul-Mageed et al. (2020a) noted that some Arabic speakers use more MSA terms in their speech to make their message understandable by a wider audience. A similar remark was made by Holes (1995), as elaborated on in the following quote:

> "They (Arabs) know that MSA is always there as a kind of communally-owned reservoir which they can dip into when they need to – a word here, a borrowed phrase there – in order to ensure that they make themselves understood to Arabs from distant countries or outsiders such as Arabic-speaking foreigners". (Holes, 1995, p. 5)

As a subcategory of code-switching, Bassiouney (2009) argued that *Diglossic Code-switching* should be governed by the same rules and constraints of code-switching between different languages. She studied the applicability of two models of code-switching—the Matrix Language (ML) and 4-M models—to data consisting of thirty hours of mosque sermons, university lectures, and political speeches. In her analysis, she differentiates between MSA morphemes, Egyptian Colloquial Arabic (ECA) morphemes, and Neutral morphemes for the ones she thought belonged to both MSA and ECA. Table 2.2 lists one of the many labeled sentences she provided, which exemplifies how the morphemes are annotated.

<u>*Limitations*</u> Bassiouney (2009) noted that "*Data on code-switching usually comes from oral performances rather than written*". Hence, applying the same analysis to written text has several limitations, which I list below.

First, some phonological differences in speech get normalized in text as mentioned in §2.2. Contrasting Bassiouney's (2009) IPA transcription and my Arabic-script transcription in Table 2.2 better explains this. In this example, the Arabic consonant ق is realized as a glottal stop /ʔ/ in yibʔa (a marker of ECA) and as a uvular stop /q/ in mana:ṭiq (a marker of MSA), as indicated by the IPA transcription. This phonological difference is normalized in Arabic script, where the letter is written as ق in both terms. Since both terms have MSA origins, Arabic speakers tend to etymologically write them in text. On doing so, these terms' labels become ambiguous and can only be inferred in some cases based on the surrounding context. Consequently, many tokens might be labeled as *Neutral*, failing to sufficiently represent the intraspeaker variation. A similar

Bassiouney's (2009) Transcription in IPA

Morphemes	l-mana:ṭiq		haðihi	fi:	ḥima:ya	fi:h	b-yibʔa		yaʕni
	mana:ṭiq	l-	haðihi	fi:	ḥima:ya	fi:h	yibʔa	b-	yaʕni
Labels	MSA	MSA	MSA	ECA or MSA	Neutral	ECA	ECA	ECA	ECA

A Potential Corresponding Transcription in Arabic Script (mine)

Morphemes	المناطق		هذه	في	حماية	فيه	بيبقى		يعني
	مناطق	ال-	هذه	في	حماية	فيه	يبقى	ب-	يعني
Labels	Neutral	Neutral	MSA	ECA or MSA	Neutral	ECA	Neutral	ECA	ECA

Table 2.2: An example of a transcribed sentence analyzed under the diglossic code-switching framework, extracted from Bassiouney (2009, p. 53). The sentence could be translated as: 'That is to say, there was protection in these places'. Some phonological differences between MSA and Egyptian Colloquial Arabic (ECA) are not rendered in Arabic script, converting some terms' labels to Neutral in the absence of the original speech. **Note:** It is unclear why *fi:* is considered an *ECA or MSA* morpheme instead of a *Neutral* one, for the IPA transcription.

observation was reached by Parkinson (1991), who instructed subjects to read through an editorial where MSA and ECA are mixed. The subjects were then asked to mark sections of the editorial as being in Fusha (MSA), being in ECA, or as indistinguishable. He found that: "*subjects label whole sections together as colloquial or fusha based on just a few markers, and that for most the presence of even one or two colloquial markers justifies a colloquiral rating for a whole section, indicating that speakers may feel that mixed texts are at base colloquial with borrowings from fusha rather than the other way around*". This indicates that applying code-switching tagging to text could end up with the same *Binarization limitation* as in *Diglossia*, where long spans of text (potentially whole sentences) are tagged as being in MSA or in DA.

Second, labeling tokens (morphemes) as belonging to MSA, DA, or both (i.e., Neutral) is not a completely objective task. For instance, Altanir (2017) identified more than 1400 terms and phrases used in ECA that either exist in MSA or have MSA origins. Speakers of ECA could perceive terms not commonly used as MSA ones but used colloquially as dialectal terms, when in fact they are also valid in MSA. As an example, both خمر *xmr* and خمرة *xmrħ* are valid MSA terms for *wine*. The first is mentioned in some verses of the Holy Qur'an, while the latter is more colloquially used in Egypt. Hence, Egyptians might link the first to MSA, and the latter to ECA, when in fact the latter is also valid in MSA.

Last, this framework fails to differentiate between colloquial words that have MSA origins from other colloquial words that are not etymologically related to MSA. More elaboration on this limitation will be provided in chapter 3.

3) Dialect Levels provides a more nuanced perspective on the intraspeaker variation of Arabic, in contrast to adopting a binarized framing on the sentence or token/morpheme level. This theory attempts to identify and define multiple levels along a continuum between pure standard and highly colloquial Arabic. For instance, the five following varieties or style levels were distinguished by Blanc (1960) as cited in (Drozdík, 2006):[4] (1) Standard Classical, (2) Modified Classical, (3) Semiliterary or Elevated Colloquial, (4) Koineized Colloquial, and (5) Plain Colloquial. Additionally, Badawi (1973) defined five different levels that exist in Modern Egypt: (1) Heritage Fus-ha (فصحى التراث), (2) Fus-ha of the age (we live in) (فصحى العصر), (3) Dialect of the (well-)educated (عامية المتعلمين), (4) Dialect of the Literate (عامية المتنورين), (5) Dialect of the Illiterate (عامية الأميين).[5] Similar to work on *Diglossic Code-switching*, his analysis also focused on spoken Arabic. He provided examples of how each of these levels is used in radio and TV shows, as summarized in Figure 2.7. Moreover, he described some phonological, morphological, lexical, and syntactic features linked to each level.

Level 1: Recitation of the Koran; dramatic recreation of events in Islamic history

Level 2: Political speech to the nation, read from a prepared text; news bulletin; voice-over commentary on serious documentary

Level 3: Studio discussion on any serious topic, for example, literature, the environment; unprepared interview with government minister, scientist, writer

Level 4: "Vox pop" interviews in the street with ordinary people; "ordinary people" depicted in television/radio plays, serials, soap operas; discussions, interviews on nonserious topics, especially if involving women (e.g., cooking, fashion); game shows; sports commentary

Level 5: Rarely represented except by speech of stereotypical working-class characters (doormen, porters, messengers, cleaners) in comedies and soap operas.

Figure 2.7: Examples of how the five levels identified by Badawi (1973) are used in radio and TV shows as summarized in Holes (1995, p. 282).

The different levels allow for distinguishing between sentences with colloquial

[4] I could not find a way to access (Blanc, 1960) online.

[5] The provided English translations of the levels are incoporated from Holes (1995).

terms that originate from MSA and others that do not. Badawi (1973) proposed that these levels can even be adopted for loan words. Consider the following three words: سندس /snds/ (silk), أيلاينر *ÂylAynr* (eyeliner), and باصة *bASħ* (a pass). The first has a Persian origin, but is mentioned in the Holy Qur'an, and is perceived as a standard term rooted in CA, hence linked to Level 1. The second is only used by Arabs with some knowledge of English, and hence can be linked to Level 4. The last is a term related to football. It is widely used and understood, making it belong to Level 5, despite being derived from the English word (pass), which would be transliterated as باص.

Limitations Similar to *Diglossic Code-switching*, a lot of the phonological features that are linked to each level are lost in text. Another major limitation is that the definition of these levels depends on linking them to different social groups, which could reinforce existing biases that consider some varieties of colloquial Arabic more prestigious than others. Lastly, Badawi (1973) acknowledged that these five levels are not completely separable. Conversely, they exist on a continuum, with the same person able to use different levels in different circumstances:

> " وأن كل فرد من أفراد المجتمع المصري (إلا في الحالات النادرة وبين الأميين فقط وخاصة النساء منهم) يستطيع استخدام أكثر من مستوى. والمثقف الذي أكمل تعليمه الجامعى من شأنه أن يكون قادراً على استخدام المستويات الثلاثة: الثاني والثالث والرابع كل فيما يناسبه. ولما كان الشخص ذاته وحدة متشابكة فإن المستويات التى يستخدمها -- حتى ولو كانت في مناسبات مختلفة -- تبقى على اتصال فني بينها. "
>
> (Badawi, 1973, p. 93)
>
> Translation Every member of the Egyptian society (except in rare cases and only among the illiterate, especially women) can employ more than one level. The educated person who has completed their university education has the ability to use the three levels: the second, the third, and the fourth, each in its suitable situations. Since a person is an interconnected unit, the levels they use—even if on different occasions—remain in technical contact with each other.

In this section, I described three linguistic theories that explain intraspeaker variation within Arabic. Only the first two theories were adopted in NLP models, with the third one considered only in the development of a few datasets. In §3.1 of chapter 3, I will summarize the different attempts to incorporate these theories into NLP models. I will also motivate why operationalizing the *Dialect Levels* theory as a variable that can be automatically estimated is more capable of capturing the intraspeaker variation than the previous operationalizations of the other two theories.

2.4 The Resourcedness Levels of the Arabic Varieties

In their survey, Nigatu et al. (2024) qualitatively analyzed how the Association for Computational Linguistics (ACL) community defines "low-resourcedness", and identified three main aspects, based on which this classification is made, namely: *socio-political* (i.e., economic or political reasons), *resources* (native speakers, online presence, and language experts), and *artifacts* (linguistic knowledge, data, and technological support). Generally speaking, MSA is not considered a low-resource language according to the different factors. Earlier efforts in Arabic NLP focused on MSA, given the availability of digital MSA corpora. However, most of the different varieties of DA are low-resource for one or more reasons, making it harder to develop NLP models for varieties of DA compared to MSA.

Socio-political factors Until recently, all the varieties of DA would be considered "low-resource" mainly due to the lack of economic interest supporting them. In the first decade of the 21st century, some initiatives targeted dialects spoken in regions of conflict, such as Levantine Arabic (Chiang et al., 2006) and Iraqi Arabic (Graff et al., 2006). Multiple projects interested in some Arabic dialects, such as Iraqi Arabic, were funded by non-Arab organizations, like the TRANSTAC (Spoken Language Communication and Translation System for Tactical Use) project funded by DARPA (Defense Advanced Research Projects Agency), with the following goal: "*The primary use case involves US military personnel and foreign language speakers. The military personnel will be trained to use the systems with the assumption that the foreign language users will receive system-provided instruction at the beginning of an interaction.*" (Weiss et al., 2008; Meermeier et al., 2018). Since these projects did not aim to serve the Arabic speakers, the various resources—data and models—developed as part of these projects are not publicly released (Condon et al., 2009, 2010, 2011, 2012). Additionally, many Arab countries (e.g., Palestine) are in an extreme political situation. On the positive side, recent interest backed by state funding is growing for a limited set of Gulf dialects, such as Emirati (Khalifa et al., 2018; Hamed et al., 2024) and Saudi (Alharbi et al., 2024).

Resources In terms of human resources, some Arab countries, such as Qatar and Bahrain, have a small population of a few million native speakers compared to other dialects, of which an even smaller sector has the incentive to participate in tasks such as

data annotation. Moreover, countries like Mauritania might not have enough language experts and NLP researchers who can better serve the dialects of these countries. Lastly, the online presence of the speakers from some countries is limited. For example, it is estimated that only 13.2% of Chad's total population of 20.7 million and 35.7% of Comoros's total population of 875 thousand have access to the internet in 2025.[6] This might explain why the NADI shared tasks organizers reported difficulties in curating enough tweets from Comoros (Abdul-Mageed et al., 2021b, 2022), and unrepresentative data from Djibouti and Somalia (Abdul-Mageed et al., 2022).

Conversely, countries such as Egypt, Morocco, and Tunisia are in a significantly better position in terms of the availability of language experts and researchers, as well as the online presence of their speakers.

Artifacts Linguistic documentation artifacts, such as lexicons and grammars, exist for some Arabic dialects (Holbrook, 1942; Emerson and Ghanim, 1943; Cowell et al., 1964; Aoun et al., 1994; Ennaji et al., 2004). In contrast, we lack basic linguistic documentation of the dialects spoken in countries like Comoros. Ethnologue only lists *Standard Arabic* as one of the languages spoken in Comoros, but there is little evidence on how this Comorian variety of Standard Arabic relates to other varieties in the rest of the Arab world.

2.5 NLP Efforts to Represent the Arabic Varieties

Due to the rise of social media text, handling DA has become increasingly important for Arabic NLP systems. The need to develop models that support a specific non-standardized variety of DA requires finding (selecting) a relatively large amount of pretraining data in this variety. An extra verification step is needed for generative models to ensure that their outputs adhere to the specified variety.

However, defining clear boundaries for a specific variety is challenging. Along a large geographical area where a language is spoken, Chambers and Trudgill (1998) argue for the presence of a "geographical dialect continuum". If one travels along this vast geographical area, one will find some linguistic differences that can distinguish one variety from another. Despite these differences, and given that the degree of linguistic difference varies, varieties spoken in neighboring areas are still mutually intelligible.

[6] Source: `https://datareportal.com/reports/digital-2025-chad`, `https://datareportal.com/reports/digital-2025-comoros`

However, the linguistic differences accumulate, such that the varieties of distant areas can be too different to the extent that they are not mutually intelligible. Holes (1995, p. 3) argues that Arabic dialects could also be thought of as being "*distributed along innumerable sets of intersecting continua*". Assuming continua instead of a single continuum potentially accounts for the fact that inhabited areas along the Arab world are separated by huge deserted areas, which prevent the presence of a single continuum.

To computationally distinguish between the varieties of DA, they are generally grouped into dialect areas. Arabic Dialect Identification (ADI)—a widely studied task in Arabic NLP with multiple datasets created and shared tasks run—could be employed to determine the variety of an Arabic sentence or a speech segment, according to a predefined set of dialect areas. It has so far been modeled as a single-label classification task. While some linguistic features (e.g., morphological, phonological) can be used to differentiate between some varieties of DA, it is hard to find a set of features that can completely split the Arab world into distinguishable dialect areas (Behnstedt and Woidich, 2013). Hence, grouping the dialects based on their geographical distribution is a general strategy that is widely adopted.

2.5.1 Groupings of the Dialectal Arabic Varieties

The **East/West dichotomy** splits the Arabic dialects into eastern dialects spoken in the Gulf and Levant, and western dialects spoken in North African countries, with Egyptian Arabic as an intermediate dialect (Kaye and Rosenhouse, 1997). These could be split further into **macro-regional dialects**. One grouping of the macro-regional dialects is: the Levant (Lebanon, Jordan, Palestine, Syria), Nile Basin (Egypt, Sudan), Gulf (Saudi Arabia, Oman, Qatar, Bahrain, United Arab Emirates, Iraq), Gulf of Aden (Yemen, Djibouti, Somalia), and Maghreb (Morocco, Tunisia, Algeria, Mauritania, Libya).[7] More fine-grained groupings are defined at the **country level**, and can even go to the **province/city levels**. For ADI, much of the work has been done at the sentence or document level, but there has also been work on token-level DI for code-switching, for example, on Egyptian Arabic-MSA tweets (Solorio et al., 2014; Molina et al., 2016) and on Algerian Arabic (Adouane and Dobnik, 2017). Moreover, ADI has progressively developed from being modeled on the regional level to being modeled on more fine-grained levels according to the following timeline:

[7]The geographical-based macro-regional classification is not fully agreed upon within the community, with some slight differences in the classifications used (Habash, 2010; Abdul-Mageed et al., 2020b).

(1) Regional-level Dialects Early efforts in ADI considered distinguishing between a subset of the macro-regional varieties, including MSA as an independent variety/class (Biadsy et al., 2009; Zaidan and Callison-Burch, 2011). Three follow-up papers—in which regional-level dialect labels were collected—introduced a new class (*General*) for the cases where "enough evidence exists that the sentence is not in MSA, but contains no evidence for a specific dialect" (Zbib et al., 2012; Cotterell and Callison-Burch, 2014; Zaidan and Callison-Burch, 2014).

For the latter (AOC dataset; Zaidan and Callison-Burch, 2014), the *General* class represented about 6.3% of the total annotations, demonstrating how the regional dialects are not always fully distinguishable from each other. However, including this additional label introduced multiple complications. First, some annotators used the *General* label when they could not decide the dialect of the sentence, for which they should have chosen a *Not Sure* label. Second, annotators tended to overidentify their dialects in the data, a trend that the dataset creators called **native-dialect bias**. For instance, Levantine annotators selected the *Levant* label more frequently than non-Levantine annotators. However, I think that the **native-dialect bias** is not necessarily an annotation error. Imagine a sentence valid in multiple dialects. While an Arabic speaker could identify if the sentence is valid in their dialect or not, they conversely could not confidently identify if the sentence is valid in any of the other dialects. For these sentences, annotators might tend to select their native dialect instead of the correct *General* label.

The overlap between the regional dialects—indicated by the non-negligible proportion of *General* annotations in the AOC dataset—was ignored on annotating further regional-level datasets, which only used a set of regional dialects and MSA as labels for the sentences (Bouamor et al., 2014; Salama et al., 2014; Huang, 2015; Malmasi et al., 2016; Zampieri et al., 2017, 2018; El-Haj et al., 2018; Alsarsour et al., 2018; Abu Farha et al., 2021). However, some papers still acknowledged the limitation that some sentences are valid in multiple regional dialects (Malmasi et al., 2016; Lulu and Elnagar, 2018; Salloum, 2018; El-Haj, 2020), or valid in both MSA and a regional dialect (El-Haj et al., 2018), especially that some phonological differences are lost in text.

Performance of Regional-level ADI systems On building a regional-level ADI using the AOC dataset, Zaidan and Callison-Burch (2014) did not rely on the majority-vote label based on the three annotations curated for each sentence. Given that the AOC's sentences are online comments to three newspapers' articles, the dialectal sentences

(i.e., the ones whose majority-vote label is not MSA) are assigned a dialect based on where the newspapers are published. More specifically, Egyptian, Gulf, and Levantine are assigned to the comments posted to *Youm7*'s, *AlRiyadh*'s, and *AlGhad* articles, respectively. This once again highlights shortcomings of how the task is formulated and how the dataset was annotated.

Test Set(s) Information and Label Distribution	**Results**
- AOC *: A random 10% of the dataset (>110K samples) *MSA* (>60% of the samples) - *EGY* - *LEV* - *GLF* Zaidan and Callison-Burch (2014)	Acc = 81%†
- AOC: *MSA* (6,355) - *EGY* (1,050) - *LEV* (1,050) - *GLF* (1,050) **- FB test set**: *MSA* (1,363) - *EGY* (800) - *LEV* (123) - *GLF* (96) Huang (2015)	Acc = 87.8%† Acc = 68.2%
VarDial 2016: *MSA* (274) - *EGY* (315) - *LEV* (344) - *GLF* (256) - *NOR* (351) Malmasi et al. (2016)	Acc = 51.2%†
VarDial 2017: *MSA* (262) - *EGY* (302) - *LEV* (334) - *GLF* (250) - *NOR* (344) Zampieri et al. (2017)	$F1_{weighted}$ = 0.763 **Sp**
VarDial 2018 (Broadcast): *MSA* (262) - *EGY* (302) - *LEV* (334) - *GLF* (250) - *NOR* (344) + **VarDial 2018 (YouTube)**: *MSA* (944) - *EGY* (1,143) - *LEV* (1,131) - *GLF* (1,147) - *NOR* (980) Zampieri et al. (2018)	$F1_{macro}$ = 0.589 **Sp**
MADAR (CORPUS-6): *MSA* (2,000) - *BEIRUT* (2,000) - *CAIRO* (2,000) - *DOHA* (2,000) - *TUNIS* (2,000) - *RABAT* (2,000) (Salameh et al., 2018)	Acc. = 93.6%†
Arabic Dialects Dataset: A subset of AOC and a Tunisian Corpus *EGY* (1,741) - *GLF* (1,092) - *LEV* (1,056) - *MSA* (1,600) - *NOR* (1,584) El-Haj et al. (2018)	Acc = 66.12%†
Habibi *: A random 30% of the Habibi dataset (50,550 samples) EGY (27.7%) - LEV (24.1%) - GLF (18.3%) - Sudan (13.0%) - Iraqi (10.5%) - MGH (6.4%) El-Haj (2020)	Acc = 72.6%†

Table 2.3: The performance of regional-level ADI systems introduced in 8 different papers. The result of the best-performing model in each paper is reported. **Note**: *: the exact number of samples in each split is not explicitly reported, and the used data splits could not be found, †: the train/test sets are based on random sampling from the same dataset (i.e., the same data distribution), **Sp**: the models' predictions are also based on additional speech features provided by the shared task organizers.

When I analyzed the performance of Zaidan and Callison-Burch's (2014) model and subsequent regional-level ADI models summarized in Table 2.3, two issues arose, which might have led to inflated models' performances. First, five papers used random train/test splits, a setup that does not evaluate the models' generalization to out-of-

domain data. More specifically, Søgaard et al. (2021) empirically demonstrated that randomly generated train/test splits underestimate the error observed on new samples from the same in-domain distribution, and recommended that multiple independent test sets be used instead to have more realistic performance estimates.

Second, five papers reported accuracy scores on imbalanced test sets, for which macro-averaged F1-scores are more appropriate. Despite these two performance-inflating issues, all the reported scores still indicate that the task is not solved, except for the *MADAR (Corpus-6)* dataset (Salameh et al., 2018), for which I identify two potential reasons. MADAR's authors identified Beirut, Cairo, Doha, Tunis, and Rabat as anchor cities for wider regional dialects. Hence, sentences written in these city-level dialects might have been more distinguishable from each other compared to sentences from other non-anchor cities. Moreover, the dataset was created by translating the same sentences from English or French into MSA in addition to the 5 city dialects. The translators might have tried to include more cues of their dialects in their translations to distinguish them from MSA translations and the other dialects' translations.

This analysis indicates that the regional-level ADI is not fully solved as a single-label classification task. However, the field decided to move to performing ADI at the country level and even attempted to model ADI at the city/province level. It is worth mentioning that if a significant proportion of Arabic sentences are valid in multiple macro-regional dialects, then this overlap will still exist if the dialect labels are assigned on the country level.

(2) Country-level Dialects Grouping dialects into macro-regions can still abstract differences between the various dialects spoken within each region (Shon et al., 2020; Althobaiti, 2020; Messaoudi et al., 2022) (e.g., Egyptian vs Sudanese Arabic as dialects within the Nile Basin (Abdul-Mageed et al., 2018), or the various dialects of the Levant (Abu Kwaik et al., 2018)). Therefore, sets of labels that are more fine-grained than the macro-regional ones were proposed for the task of ADI.

Country-level ADI is the most common setup (Abu Kwaik et al., 2018; Shon et al., 2020; Abdelali et al., 2021), with some datasets targeting both country-level and province/city-level ADI (Abdul-Mageed et al., 2018; Salameh et al., 2018; Bouamor et al., 2019; Abdul-Mageed et al., 2020a, 2021b).

Notably, the Nuanced Arabic Dialect Identification (NADI) shared tasks (Abdul-Mageed et al., 2020a, 2021b, 2022, 2023) used datasets that were built by collecting Arabic tweets authored by users who have been tweeting from the same location for

10 consecutive months. The geolocation of the users is then used as a label for their tweets. NADI's organizers have been improving the quality of the dataset from one year to another, as summarized in Figure 2.8. While the NADI shared tasks have been attracting active participation, the best-performing models in NADI 2022 achieved macro F1 scores of 36.48% and 18.95%, and accuracies of 53.05% and 36.84% on two independent test sets (Abdul-Mageed et al., 2022).[8]

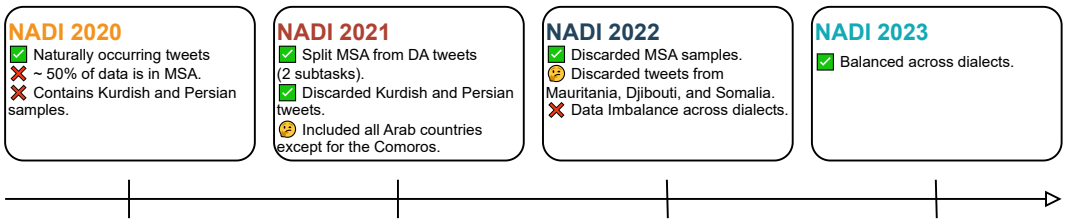


Figure 2.8: The evolution of the NADI datasets used for the shared tasks run between 2020 and 2023.

Intuitively, the number of sentences valid in multiple dialects increases as the set of dialects gets more fine-grained. On analyzing the errors of single-label ADI systems, it is commonly mentioned that the many errors are caused by confusing dialects spoken in neighboring countries, which are mostly part of the same regional dialect (Biadsy et al., 2009; Salameh et al., 2018; Talafha et al., 2019; Samih et al., 2019; Ragab et al., 2019; Přibáň and Taylor, 2019; Ghoul and Lejeune, 2019; Eltanbouly et al., 2019; Abu Kwaik and Saad, 2019; Dhaou and Lejeune, 2020; Talafha et al., 2020; Aloraini et al., 2020; Abdelali et al., 2021; AlKhamissi et al., 2021; El Mekki et al., 2021; Jamal et al., 2022; Khered et al., 2022; Attieh and Hassan, 2022). All the different methods previously proposed for curating single-label ADI datasets share this issue—refer to §5.2 for a detailed description of these methods and their corresponding datasets.

2.5.2 Calls for Improvement

As a way to achieve a less harsh evaluation of single-label ADI models, Alsudais et al. (2022) relied on the concept that the dialects spoken in geographically proximate areas are not too different, and proposed not to penalize the model if it predicts a dialect (on the city level) that is within a 'tolerance distance' from that of the gold-standard city-level dialect.

[8] Surprisingly, the best-performing model in NADI 2023 achieved a macro-averaged F1 score of 87.27, but the shared task's overview paper does not provide explanations for this huge improvement, nor did the shared task's main organizer when contacted personally.

Based on limited qualitative analysis, a few calls recommended that ADI should be framed as a multi-label classification task instead of a single-label one (Kchaou et al., 2019; Touileb, 2020). However, these recommendations did not get much notice from the community. It is conceivable that the difficulty of creating multi-label ADI datasets is a major obstacle. Another reason might be that the community is underestimating the pervasiveness of these multi-label samples. More specifically, a common belief is that most multi-label samples are very short (Alorifi, 2008; El-Haj et al., 2018; Alsarsour et al., 2018; Abu Kwaik and Saad, 2019; Althobaiti, 2022). Since most NLP models would struggle with these short sentences, widely holding this belief might explain why ADI continued to be modeled as a single-label classification task.

2.6 Research Gaps

The previous section explained how single-label ADI is a popular task in Arabic NLP, where intraspeaker variation is modeled by linking each sentence to a dialect, according to a predefined taxonomy of dialects. This framing has a *Disjointedness* limitation, where the predefined set of dialects is supposed to be disjoint. It was qualitatively found that some sentences could be valid in multiple neighboring dialects. However, the lack of quantitative analysis of this overlap could be the reason why this limitation has been ignored for a long time. I try to tackle this in (***RQ4***): *How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?*

To model the intraspeaker variation, most single-label ADI systems have MSA as an independent dialect, conforming to the theory of *Diglossia*. This ignores any levels of variation between pure MSA and pure colloquial sentences, which I refer to as the *Binarization* assumption. The linguistic theory of *Dialect Levels* has been proposed to provide a more realistic representation of the intraspeaker variation. (***RQ1***) attempts to adopt this theory as an NLP tool, investigating *How can the concept of Dialect Levels be operationalized in a way that can be effectively estimated?*

The following chapters investigate these research questions in detail, to tackle the two main aforementioned limitations: *Binarization* and *Disjointedness*. Subsequent questions related to each limitation are investigated based on the findings of (***RQ1***) and (***RQ4***). Lastly, I showcase how alleviating these limitations could be useful for multiple applications, such as quantitatively analyzing political speeches of Arab presidents and having a more efficient pipeline for routing dialectal samples to Arabic-speaking annotators.

Chapter 3

Arabic Level of Dialectness

This chapter presents how a more nuanced computational modeling of intraspeaker variation in Arabic could be achieved. As explained in §2.2, non-negligible differences exist between MSA and the varieties of DA. Additionally, Arabic speakers commonly code-switch between the two. Both these factors form a major challenge for Arabic NLP systems. As a result, many systems have been designed to perform Dialect Identification (DI), often on the sentence level (e.g., Zaidan and Callison-Burch, 2011; Elfardy and Diab, 2013; Salameh et al., 2018), but also on the token level as a way of detecting code-switching points (Solorio et al., 2014; Molina et al., 2016). Both formulations take a binary view of the problem (a sentence or token is either MSA or DA), and assume all the features of DA have the same impact on the perceived "dialectness" of a sentence. These formulations follow the theory of *Diglossia*. In contrast, I argue that Arabic speakers perceive a spectrum of dialectness (**Hypothesis 1**), as illustrated in Table 3.1. This formulation is closely tied to the *Dialect Levels* theory (confer §2.3 for a more detailed discussion of the different linguistic theories).

This chapter visits (***RQ1***): ***How can the concept of Dialect Levels be operationalized in a way that can be effectively estimated?*** Throughout the thesis, I show that the level of dialectness is an important but overlooked aspect of Arabic text, which is complementary to and more nuanced than dialect identification. To support this claim and promote further research in the area, I:

1. Define the ***Arabic Level of Dialectness (ALDi)*** as a continuous linguistic variable that quantifies the dialectness of a sentence (or sentence-like unit) and can enrich the analysis of Arabic text.

2. Release *AOC-ALDi*[1], a dataset of 127,835 Arabic comments with their ALDi labels, which is derived from the Arabic Online Commentary dataset (Zaidan and Callison-Burch, 2011, 2014). I provide the first detailed analysis of the *level of dialectness* labels and form canonical splits for the *AOC-ALDi* dataset.
3. Present an effective model for estimating the ALDi of sentences, that can generalize to corpora of other genres and dialects.[2] This model is the first of its kind that automatically quantifies the intraspeaker variation within Arabic in a more nuanced way beyond the binary distinction between MSA and DA.

Level of Dialectness	Egyptian	Levantine
MSA	أسعدنا الرجل	أسعدنا الرجل
Low	الراجل أسعدنا	الزلمة أسعدنا
Medium	الراجل بسطنا	الزلمة بسطنا
High	الراجل شهيصنا	الزلمة نغنجنا

Table 3.1: Example sentence meaning *the man cheered us* written with different levels of dialectness in two Arabic dialects. Words with DA features are underlined. The dialectal sentences use their preferred SVO word order, contrasted by the VOS order for MSA. The low dialectness example also shows a lexical dialectal feature for the word *the man* (MSA الرجل - Alrjl): the Egyptian word (الراجل - AlrAjl) differs from MSA in a single character, while the equivalent Levantine word (الزلمة - Alzlmħ) has a different origin. Both dialects allow different variants for the verb: one variant (بسطنا - bsTnA), used in both dialects, shares a root with the MSA variant, while the more dialectal variants (شهيصنا - šhySnA in Egyptian and نغنجنا - nɣnjnA in Levantine) do not.

The work presented herein was reported in the following paper:

- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. *ALDi: Quantifying the Arabic Level of Dialectness of Text.* In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10597–10611, Singapore.

[1]The code and data files can be accessed through: `https://github.com/AMR-KELEG/ALDi`
[2]A live demo for ALDi estimation: `https://huggingface.co/spaces/AMR-KELEG/ALDi`

3.1 Previous Attempts to Computationally Model Intraspeaker Variation

§2.3 presented three different linguistic theories—*Diglossia, Code-switching*, and *Dialect Levels*—that were proposed to model the intraspeaker variation in Arabic, and discussed the limitations of each of them. NLP researchers attempted to incorporate these theories in their models and tools, as I elaborate in this section.

(1) Diglossia is by far the most widely adopted theory of the three in NLP. In building sentence-level dialect identification datasets, MSA is generally considered an independent class, with different varieties of DA (at different levels of granularity) representing the other classes. ADI has attracted considerable research attention, from organizing multiple shared tasks (Zampieri et al. 2014; Bouamor et al. 2019; Abdul-Mageed et al. 2020a, 2021b, 2022, 2023) to building datasets (Zaidan and Callison-Burch, 2011; Bouamor et al., 2014; Salama et al., 2014; Bouamor et al., 2018; Alsarsour et al., 2018; Zaghouani and Charfi, 2018; El-Haj, 2020; Abdelali et al., 2021; Althobaiti, 2022).

Limitations Clearly, ADI models take a binary view by which a sentence is either in pure MSA or a variety of DA, failing to differentiate between the different levels of dialectness within the non-MSA sentences.

(2) Code-switching was considered by a couple of computational efforts which modeled code-switching between MSA and DA as a token-level language (variety) identification (Solorio et al., 2014; Molina et al., 2016). Both efforts annotated datasets that were used to organize two shared tasks for multiple language/variety pairs. Solorio et al. (2014) included the following four language pairs: Mandarin-English, Nepali-English, Spanish-English, and MSA-Egyptian Arabic (EGY), while Molina et al. (2016) focused on the last two pairs only.

For the MSA-EGY pair, each token was tagged as being in *lang1 (MSA)* or *lang2 (EGY)*. Additional labels were used for different tags, such as *named entities*, which are not expected to be linked to a specific variety, and *others* for non-textual tokens (e.g., punctuation marks and emoticons). Neither of the two papers explicitly mentions how the annotation was done for the terms valid in both MSA and EGY. However, the provided description of the additional label '*ambiguous*' hints that context was used while assigning the tags for tokens shared between MSA and EGY, since the *ambiguous* label was used for *cases where a term, belonging to both languages (varieties), appears in a context that does not indicate one language over the other*.

In addition to evaluating the token-level performance of the models' predictions submitted by the participating teams, the organizers evaluated the models' ability to distinguish between monolingual sentences (i.e., ones in a single language/variety of each pair) and code-switched sentences (i.e., ones having at least one token tagged as being in *lang1* and another as in *lang2*). This additional evaluation step is beneficial for the MSA-EGY, as it shows the models' ability to distinguish between three classes of sentences: pure MSA ones (monolingual in this setup with none of the tokens tagged as EGY), ones with a fair amount of code-switching between MSA and EGY, and pure DA ones (monolingual in this setup with non of the tokens tagged as MSA).

Limitations From a dataset annotation perspective, we already know that words/-tokens could be valid in both varieties (MSA and EGY in this case), especially in the absence of a speech signal corresponding to the annotated text (Parkinson, 1991) (refer to §2.3 for a more detailed discussion).

Dataset	MAN-EN	NEP-EN	SPA-EN	MSA-EGY
	$F1_{Macro}$	**$F1_{Macro}$**	**$F1_{Macro}$**	**$F1_{Macro}$**
Test-1	0.894	0.977	0.822	0.095
Test-2	-	-	-	0.417
Suprise	-	0.702	0.753	0.277

(a) $F1_{Macro}$ for the four language pairs of (Solorio et al., 2014).

Dataset	SPA-EN			MSA-EGY		
	$F1_{Mono}$	$F1_{CS}$	**$F1_{Macro}$**	$F1_{Mono}$	$F1_{CS}$	**$F1_{Macro}$**
Test	0.89	0.9	0.895	0.93	0.5	0.715

(b) $F1_{Macro}$ for the two language pairs of (Molina et al., 2016). The reported $F1_{Mono}$ and $F1_{CS}$ correspond to the F1 scores computed when each of the *monolingual* label and the *code-switched* label is considered as the positive class, respectively.

Table 3.2: The macro-averaged F1 scores for the sentence-level binary identification of monolingual (i.e., single language/variety) against code-switched sentences for the best-performing systems of both shared tasks.

From a modeling perspective, the results summarized in Table 3.2 show a significant gap in the ability to distinguish between monolingual and code-switched sentences for the MSA-EGY pair compared to the other pairs. Referring to Table 3.2b, consider the individual F1 scores when the positive class is the monolingual label ($F1_{Mono}$) and the

code-switched label ($F1_{CS}$), respectively. We can infer that the top-performing model tends to predict a single variety for most sentences, as indicated by achieving an $F1_{mono}$ of 0.93 for monolingual sentences, but a much lower $F1_{CS}$ of 0.5 for code-switched sentences.

From a framing perspective, both sentence-level and token-level DI methods fail to distinguish between sentences having the same number of dialectal cues, yet different levels of dialectness. As per Table 3.1, each of the sentences الزلمة بسطنا (Alzlmħ bsTnA) and الزلمة نغنجنا (Alzlmħ nɣnjnA) has two lexical cues of dialectness, yet the latter sentence is perceived as being more colloquial than the former.

(3) Dialect Levels Earlier initiatives recognized the presence of such a spectrum (Habash et al., 2008; Zaidan and Callison-Burch, 2011); however, the datasets that were developed are either skewed toward more standardized documents with limited code-switching or lack information about the distribution and the quality of the levels of dialectness labels, as elaborated next. Consequently, the *Level of Dialectness* has not yet been adopted as a linguistic variable for automatically analyzing Arabic text, despite its potential usefulness for NLP applications.

Habash et al. (2008) proposed a word-level annotation scheme consisting of four levels: (L0) Pure MSA, (L1) MSA with non-standard orthography, (L2) MSA with dialect morphology, and (L3) Dialectal lexeme. Annotators also labeled the full sentences' level of dialectness into five levels based on the word-level annotations according to the description in Table 3.3.

Although the inter-annotator agreement was relatively good (less so for the sentence level), only a small corpus was annotated (19k words). Moreover, the corpus has sentences that are mostly in MSA with limited code-switching. A later work piloted a simplified version of the scheme—where the different word levels are mapped to MSA, DA, or both—on another corpus of 30k words (Elfardy and Diab, 2012), a setup that is close to the code-switching framing. Neither corpus is publicly released. Additionally, these two token-level annotation approaches have similar limitations to the aforementioned limitations of the code-switching setup.

Conversely, Zaidan and Callison-Burch (2011) collected *sentence-level* dialectness annotations in the Arabic Online Commentary dataset. Although the dataset has been released, there has been no published description or analysis of these annotations that I know of, and (perhaps for this reason) no follow-up work using them.[3] The following

[3]This contrasts with the *Dialect* annotations for the same corpus, which were analyzed in Zaidan and Callison-Burch (2014), and have been widely used in the ADI task.

section aims to remedy this and introduces the Arabic Level of Dialectness (ALDi) variable.

Segment Level 0 - perfect MSA
Segment Level 1 - imperfect MSA The source is trying to produce MSA, but some dialectal phenomena are sneaking in. *Criteria*: A segment can not be in this level if the number of words of word-level 3 is larger than a threshold. The authors were to decide on this threshold later.
Segment Level 2 - Arabic with full dialect switching It is not clear whether the source is aiming for writing in MSA or dialect. *Criteria*: A segment can not be in this level if all the words are in word-level 2 or word-level 3.
Segment Level 3 - Dialect with MSA incursions The source is producing dialectal Arabic, but uses words clearly borrowed from MSA. *Criteria*: A segment can not be in this level if all the words are in word-level 2 or word-level 3.
Segment Level 4 - pure Dialect *Criteria*: A segment in this level at least has one word in word-level 2 or word-level 3.

Table 3.3: The five segment levels introduced by Habash et al. (2008), with their provided description and criteria. **Note:** the authors mentioned that the term *segment* is used to refer to sentences or utterances.

3.2 The Arabic Level of Dialectness (ALDi)

I define the *Arabic Level of Dialectness (ALDi)* of a sentence as the **extent by which the sentence diverges from the standard language**, which can be based on any of the cues described above. This definition is consistent with the crowdsourced annotation of the Arabic Online Commentary (AOC) dataset (Zaidan and Callison-Burch, 2011), where annotators labeled the *dialect* and *level of dialectness* of user comments on articles from Arabic newspapers.

AOC's Annotation Guidelines AOC's creators provided the annotators with minimal guidelines for determining the *dialect* and *dialect level* of the comments. For the latter,

they only instructed the annotators to answer the following: "Tell us how much dialect (عامية) is in the sentence" as indicated by the screenshot of their annotation interface shown in Figure 3.1.[4] They employed the following four labels for the *dialect level* annotations: {*No dialect, A bit of dialect, Mixed, Mostly Dialect*}, without defining each of these levels, allowing the annotators to gave their own interpretations.

The annotation guidelines could be interpreted as measuring the percentage of DA in a sentence, which is not too different from the aforementioned code-switching framing. However, AOC's guidelines allow for assigning a high level of dialectness to a sentence having a highly colloquial word, even if the remaining words are perceived to be less colloquial or in MSA, which is not the case for previous guidelines. Moreover, the guidelines do not explicitly rely on sociolinguistic factors in defining the different levels, which is a limitation of Badawi's (1973) definitions of the different *Dialect Levels*.

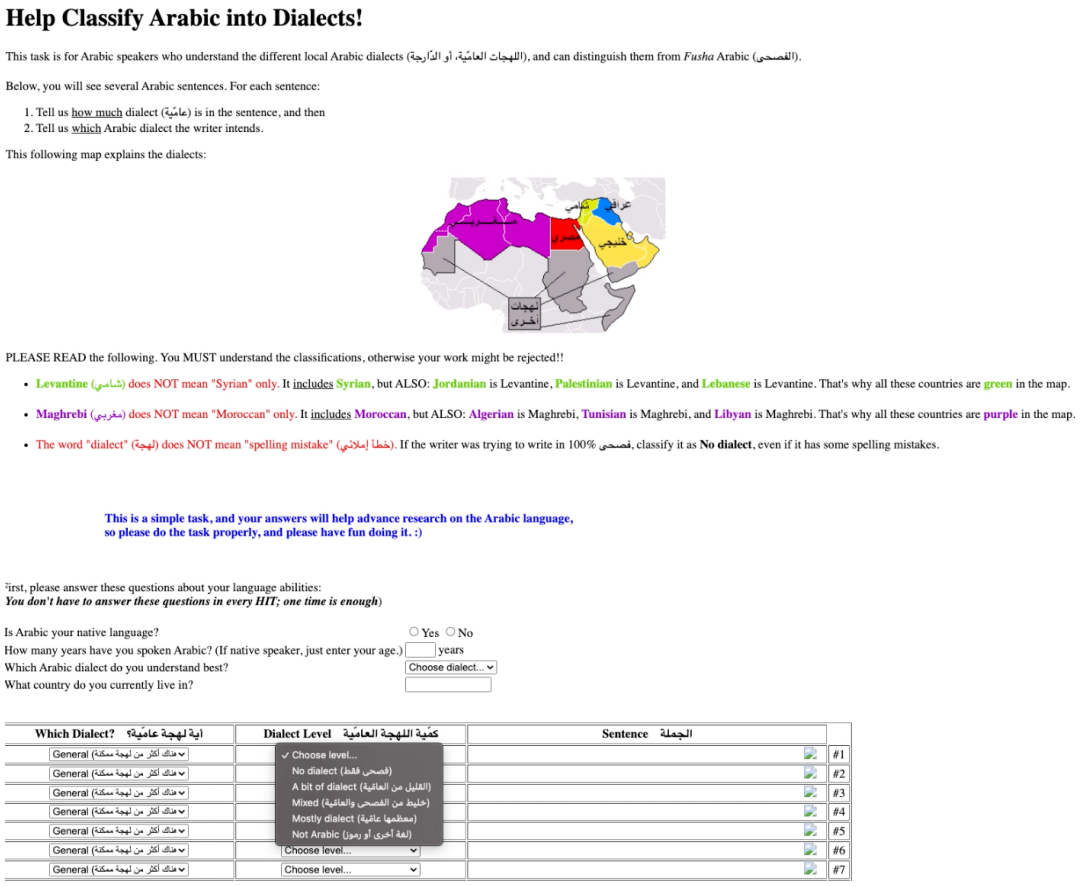
The original paper in which the AOC dataset was introduced and subsequent work only used the dialect labels. Given that the *dialect level* annotations were ignored, the following subsection will provide the first in-depth analysis of these annotations.

3.2.1 Analyzing the AOC Dataset

The AOC dataset was created by scraping user comments on articles from three newspapers, published in Egypt (*Youm7* اليوم السابع), Jordan (*AlGhad* الغد), and Saudi Arabia (*AlRiyadh* الرياض); thus, expecting the majority of each source's comments to be in Egyptian (EGY), Levantine (LEV), and Gulf (GLF) dialects, respectively. Each comment is labeled for its *Dialect Level* (*MSA, little, mixed, mostly dialectal, not Arabic*). For comments labeled as Non-MSA, the annotators also chose the *dialect* in which the text is written: EGY, LEV, GLF, Maghrebi (MAG), Iraqi (IRQ), General (GEN: used when the text is DA, but could belong to multiple dialects), Unfamiliar, and Other.

Each row of the released AOC dataset consists of 12 different sentences representing a Human Intelligence Task (HIT) on Amazon Mechanical Turk, with annotations provided by the same human judge. A HIT has 10 comments in addition to 2 control sentences sampled from the articles' bodies, which are expected to be mostly written in MSA. As part of each HIT, annotators provided some personal information such as their place of residence, whether they are native Arabic speakers, and the Arabic dialect they understand the most. Table 3.4 shows the number of annotations collected for sentences from each source.

[4]The annotation webpage can be accessed through https://www.cs.jhu.edu/data-archive/RCLMT-2011/html/dialect_classification.shtml.



Help Classify Arabic into Dialects!

This task is for Arabic speakers who understand the different local Arabic dialects (اللهجات العامّيّة، أو الدّارجة), and can distinguish them from *Fusha* Arabic (الفصحى).

Below, you will see several Arabic sentences. For each sentence:

1. Tell us how much dialect (عامّيّة) is in the sentence, and then
2. Tell us which Arabic dialect the writer intends.

This following map explains the dialects:

PLEASE READ the following. You MUST understand the classifications, otherwise your work might be rejected!!

- **Levantine** (شامي) does NOT mean "Syrian" only. It includes **Syrian**, but ALSO: **Jordanian** is Levantine, **Palestinian** is Levantine, and **Lebanese** is Levantine. That's why all these countries are **green** in the map.
- **Maghrebi** (مغربي) does NOT mean "Moroccan" only. It includes **Moroccan**, but ALSO: **Algerian** is Maghrebi, **Tunisian** is Maghrebi, and **Libyan** is Maghrebi. That's why all these countries are **purple** in the map.
- The word "dialect" (لهجة) does NOT mean "spelling mistake" (خطأ إملائي). If the writer was trying to write in 100% فصحى, classify it as **No dialect**, even if it has some spelling mistakes.

This is a simple task, and your answers will help advance research on the Arabic language, so please do the task properly, and please have fun doing it. :)

First, please answer these questions about your language abilities:
You don't have to answer these questions in every HIT; one time is enough)

Is Arabic your native language? ○ Yes ○ No
How many years have you spoken Arabic? (If native speaker, just enter your age.) years
Which Arabic dialect do you understand best? Choose dialect...
What country do you currently live in?

Which Dialect? أية لهجة عامّيّة؟	Dialect Level كمّية اللهجة العامّية	Sentence الجملة	
General (هناك أكثر من لهجة ممكنة)	✓ 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)		#1
General (هناك أكثر من لهجة ممكنة)			#2
General (هناك أكثر من لهجة ممكنة)			#3
General (هناك أكثر من لهجة ممكنة)			#4
General (هناك أكثر من لهجة ممكنة)			#5
General (هناك أكثر من لهجة ممكنة)	Choose level...		#6
General (هناك أكثر من لهجة ممكنة)	Choose level...		#7

Figure 3.1: A screenshot of the annotation interface used for the AOC dataset (Zaidan and Callison-Burch, 2011).

Type	AlGhad	AlRiyadh	Youm7
Comment	94,236	156,345	80,349
Control	48,210	8,925	9,051
All	142,446	165,270	89,400

Table 3.4: Statistics of the AOC dataset, showing the number of annotations of each type from each newspaper source. Each sentence has three independent annotations.

Table 3.5 shows the distribution of Level of Dialectness annotations in AOC. As expected, the control sentences are nearly all (94%) annotated as MSA. MSA is also the most common label for the scraped comments (57% of their annotations), followed by the mostly dialectal label (23%), little dialectal (11%), and mixed (6.5%).

Type	MSA	Little	Mixed	Most	Not Arabic	Missing
Comment	189,020 (57.12%)	36,930 (11.16%)	21,622 (6.53%)	76,284 (23.05%)	5,421 (1.64%)	1,653 (0.5%)
Control	62,456 (94.36%)	1,060 (1.6%)	436 (0.66%)	754 (1.14%)	1,165 (1.76%)	315 (0.48%)
All	251,476 (63.33%)	37,990 (9.57%)	22,058 (5.55%)	77,038 (19.4%)	6,586 (1.66%)	1,968 (0.5%)

Table 3.5: The distribution of AOC's *Level of Dialectness* annotations. Each sentence has three independent annotations. *Control* are sentences extracted from the article body, most likely MSA, to check the quality of the annotations.

Figure 3.2 shows the distribution of dialectness labels split out by dialect (sentences labeled as MSA are not shown). We can see that the proportions of different levels of dialectness for the LEV, GLF, and EGY dialects are similar, even though the total number of annotations per source (Table 3.4) is more skewed. This is likely due to the fact, noted by Zaidan and Callison-Burch (2014), that the highest proportion of MSA annotations is for AlGhad's comments, followed by AlRiyadh and then Youm7. Figure 3.2 also shows that the distribution of dialectness levels is similar for the LEV, GLF, and EGY dialects, whereas the GEN dialect label has a higher proportion of *little* dialectness. This makes sense, since for sentences with few cues of dialectness, the level of dialectness would be low, and it would be hard to assign these sentences to a specific dialect.

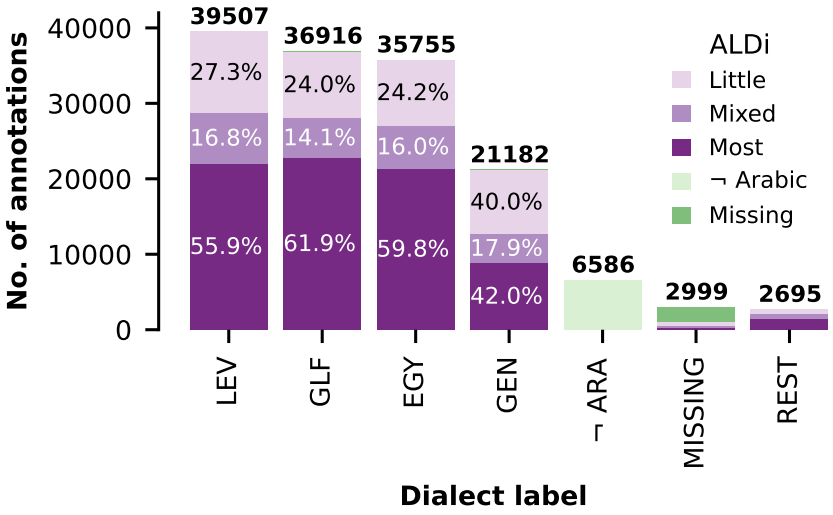


Figure 3.2: The distribution of the annotations for the dialect and the level of dialectness in AOC. Note that each comment has three different annotations. 251,476 MSA annotations are not shown in the Figure. The *General* dialect label is used when a sentence is natural in multiple varieties of DA. The *REST* bar represents the (Maghrebi, Iraqi, Unfamiliar, and Other) labels.

3.2.2 From AOC to AOC-ALDi

In order to transform the AOC level of dialectness annotations into numeric ALDi scores, I applied the following steps:

Step #1 - HIT to annotation rows: I split each row (HIT) of the AOC dataset into 12 annotation rows, one for each sentence of the HIT, with the annotator's information shared across them.

Step #2 - Grouping identical comments: Comments on the same article can sometimes be identical. I decided to group identical comments on the same article together. Out of 129,873 grouped comments, only 1,377 comments have more than three annotations. I discarded 2,038 grouped comments for which at least $\frac{2}{3}$ of the dialectness level annotations are either *Missing* or *Not Arabic*,[5] leaving a total of 127,835 comments with at least three annotations each. The average length of these comments is 20 words.

I measured inter-annotator agreement on the *level of dialectness* annotations for the 124,257 comments which have 3 annotations that are not *Not Arabic* or *Missing*. The Fleiss' Kappa (κ) is 0.44 (Fleiss, 1971), while Krippendorff's Alpha (interval method) (α) is 0.63 (Krippendorff, 2004). Both metrics are corrected for chance

[5]The main categories of these discarded comments are discussed in Appendix §A.

agreement and disagreement, respectively. κ considers the labels as categorical, while α penalizes disagreements according the differences between their values. Although these agreement levels are considered only moderate, my experiments demonstrate that the corpus can nevertheless be useful.

Step #3 - Label aggregation (Operationalization of ALDi): Multiple human annotations for the level of dialectness were aggregated into a single label. I transformed the ordinal labels (MSA, Little, Mixed, Mostly) into the numeric values (0, $\frac{1}{3}$, $\frac{2}{3}$, 1), then took the algebraic mean of these as the gold standard label, which has the range $[0, 1]$.[6] The distribution of the aggregated scores across four intervals is shown in Figure 3.3.

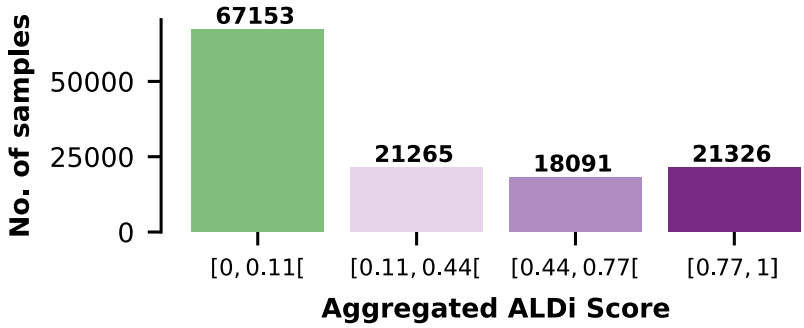


Figure 3.3: AOC-ALDi's distribution of ALDi scores.

Step #4 - Splits creation: To build reliable splits of AOC, I made sure comments to the same document are in the same split. For each source, I group sentences belonging to the same article together, shuffle these groups, and then assign the first 80% of the comments to the training split, the following 10% to the development split, and the remaining 10% to the test split. This way, the dev and test sets evaluate whether a model generalizes to comments from articles not seen in training. The distribution of the sources across AOC-ALDi's splits is in Table 3.6.

Qualitative Analysis Table 3.7 shows three example sentences from the AOC-ALDi dataset with their corresponding annotations, where all annotators labeled the dialect as either MSA, EGY, or GEN. The first sentence begins with an English loanword. The rest of the sentence has MSA terms that will not sound natural if pronounced according to the phonetic rules of a variety of DA. Unsurprisingly, two annotators considered the sentence to be in MSA, while the third might have perceived the presence of the loanword as a sign of dialectness, thus marking the sentence as *little* dialectal. The second example shows code-switching between MSA and Egyptian DA, but an

[6] AOC-ALDi also includes the original separate labels.

Split	AlGhad		AlRiyadh		Youm7	
	Cmnt	Cntrl	Cmnt	Cntrl	Cmnt	Cntrl
Train (80%)	24,039	12,613	41,479	2,335	20,041	2,379
Dev (10%)	3,107	1,513	4,567	275	2,475	323
Test (10%)	2,945	1,587	5,012	360	2,514	271

Table 3.6: The number of grouped comments in AOC-ALDi's splits. 127,835 comments of 20 words on average are distributed across all splits.

Egyptian can still naturally pronounce the MSA portion, abiding by the phonetic rules of Egyptian Arabic. This might be the reason why one of the annotators labeled the sentence as mostly dialectal (see Parkinson (1991), who observed the same relation between pronunciation and perceived levels of dialectness). For the third example, all the tokens except for the first one show dialectal features, which made it easy for the three annotators to classify it as *mostly* dialectal.

ALDi	**English translation (mine)**	**Comment**
$\overline{0, 0, \frac{1}{3}}$	Bravo to the wonderful Minister, who proved	برافو للسيد الوزير الرائع الذي
$\approx$ 0.11	that he is responsible, feeling the importance of	اثبت انه مسئول يشعر بمدى أهميه
	his responsibility for the first time in a long time	مسئوليته لاول مره منذ زمن
	in the history of Egyptian education.	بعيد في تاريخ التعليم المصري
$\overline{\frac{1}{3}, \frac{1}{3}, 1}$	We start with the right task of developing	نبتدى بقى الشغل الصح فى تطوير
≈ 0.56	schools and providing observers over them	المدارس وتوفير المراقبين عليها
$\overline{1, 1, 1}$	Honestly, a serious minister	وزير جدع بصراحة ....
≈ 1.00	I hope he stays like this all the time	ياريت يفضل كدا على طول

Table 3.7: Sample comments to the same article with their level of dialectness labels (3 annotations for each comment with their $\overline{mean}$ as the ALDi score). The labels are *MSA* (0), *Little* ($\frac{1}{3}$), *Mixed* ($\frac{2}{3}$), *Most* (1). DA segments are underlined. Loanwords are double-underlined.

3.3 The ALDi Estimation Task

In this section, I show that a model trained to predict ALDi is competitive with a DI system in discriminating between dialectal and MSA sentences (including dialects barely represented in AOC-ALDi), providing more nuanced dialectness scores. I then consider several specific features of Egyptian Arabic, and again show that the ALDi regression model is more sensitive to these than the baseline approaches.

3.3.1 Models

The primary model I use to predict ALDi is a BERT-based regression model, known as the *Sentence ALDi* model. Using the training split of *AOC-ALDi*, I fine-tune a regression head on top of MarBERT, an Arabic BERT model (Abdul-Mageed et al., 2021a), and clip the output to the range $[0, 1]$. To measure the consistency of the model's performance, I repeat the fine-tuning process three times using 30, 42, and 50 as the random seeds, and report averaged evaluation scores for the model (similarly for Baseline #3). I compare this model to three baselines, which use existing Arabic resources and are not trained on AOC-ALDi.

Baseline #1 - Proportion of tokens not found in an MSA lexicon: The presence of dialectal lexical terms is one of the main signals that humans use to determine dialectal text. Sajjad et al. (2020) built an MSA lexicon from multiple MSA corpora. They then computed the percentage of tokens within a sentence not found in the MSA lexicon as a proxy for sentence-level dialectness. I replicate this method using the tokens occurring more than once in the Arabic version of the United Nations Proceedings corpus (Ziemski et al., 2016) as the source for the MSA lexicon.

Baseline #2 - Sentence-Level DI: I use an off-the-shelf DI model implemented in (Obeid et al., 2020) based on (Salameh et al., 2018). The model is based on Naive Bayes, trained on the MADAR corpus (Bouamor et al., 2018), and uses character and word *n*-grams to classify a sentence into six varieties of DA in addition to MSA. A sentence is assigned an ALDi score of 0 if it is classified as MSA and a score of 1 otherwise.

Baseline #3 - Token-level DI: Molina et al. (2016) created a token-level DI dataset *(MSA-EGY token DI)*, in which tokens of tweets were manually tagged as MSA, EGY, Named-Entity, ambiguous, mixed, or other. I use this dataset to fine-tune a layer on top of MarBERT to tag tokens of a sentence. The tag of the first subword for each token is used as the tag for the whole token, as done in (Devlin et al., 2019). I use token-level

tags to compute the Code-Mixing Index (CMI; Das and Gambäck 2014) as a proxy for ALDi: $CMI = \frac{N_{EGY\ tokens}}{N_{EGY\ tokens}+N_{MSA\ tokens}}$ (set to 0 if none of the tokens are tagged as MSA or EGY).

3.3.2 Evaluation

Intrinsic AOC-ALDi evaluation Treating the aggregated human-assigned scores of AOC-ALDi's test split as the gold standard, I measure how the models' ALDi predictions deviate from the gold standard ones using Root Mean-Square Error (RMSE). As expected, since it is the only model trained on AOC-ALDi, the *Sentence ALDi* model achieves the least RMSE of 0.18 on the AOC-ALDi test split, as indicated in Table 3.8. The two other models that can produce continuous scores at the sentence level, *MSA Lexicon* and *Token DI*, achieve similar RMSE, and are both better than the binary *Sentence DI* model despite more limited exposure to the dialects in this corpus (recall that *Token DI* has only been trained on EGY and MSA, and *MSA Lexicon* has no explicit DA training). All models perform worse on the comments than the controls.

Model	$RMSE(\downarrow)$		
	Control	**Comment**	**All**
	N=2,127	N=10,644	N=12,771
MSA Lexicon	0.13	0.36	0.34
Sentence DI	0.23	0.53	0.49
Token DI	0.11*	0.33*	0.30*
Sentence ALDi	**0.07***	**0.19***	**0.18***

Table 3.8: Models' RMSE on AOC-ALDi's test split. *: Average values across three fine-tuned models with different random seeds. The corresponding standard deviations are 0.015 or less.

Disentangling Parallel MSA/DA Sentences For a model estimating ALDi, a minimal requirement is to assign a higher score to a DA sentence than that assigned to its corresponding MSA translation.

I utilize two parallel corpora of different genres and dialects to test this requirement. First, I use a parallel corpus of 8,219 verses (sentences) from the **Bible**, provided by Sajjad et al. (2020), which includes versions in MSA, Tunisian, and Moroccan Arabic. I also use **DIAL2MSA**, which is a dataset of dialectal Arabic tweets with parallel MSA

translations (Mubarak, 2018). Five MSA translations were crowdsourced for 12,000 tweets, which had distinctive lexical features of Egyptian and Maghrebi Arabic. Each translation was then validated by three judges. For my analysis, I discard samples having a non-perfect validation confidence score and ones that still have a distinctive dialectal lexical term in their MSA translations.

The distribution of the ALDi scores in Figure 3.4 reveals that *MSA Lexicon* does not discriminate strongly between MSA and DA, while *Token DI* mostly assigns scores of 0 or 1 (acting like *Sentence DI*), despite the possibility to do otherwise. The *Sentence ALDi* model provides more nuanced scores while also showing strong discrimination between MSA and DA, even for DA varieties that are barely present in AOC-ALDi (TUN, MOR, MGR; note that *Token DI* also has not seen these). Additionally, the *Sentence ALDi* model yields slightly lower scores for the DA versions of the Bible than for the DA tweets, hinting that the informal genre of tweets may be an indicator of stronger dialectness levels.

It is conceivable that the *Sentence ALDi* model has generalized to identify divergence from MSA, irrespective of the underlying dialect. This generalization could have been aided by the underlying BERT-based model's pretraining data (Abdul-Mageed et al., 2021a), which consisted of 1 billion Arabic tweets that are expected to cover some of the different Arabic varieties. However, the model's pretraining data is not public, and the distribution of the varieties included in the model's pretraining data is neither provided nor easy to estimate, making it difficult to further study the impact of the pretraining data on the *Sentence ALDi* model's generalization.

To further compare the different methods, I computed D', a measure of discrimination, for all models on each pair of parallel corpora, with the results shown in Table 3.9. For a given corpus of parallel MSA and DA sentences, each model yields two distributions of estimated ALDi scores. The mean and variance of the two distributions are (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively. D' is then computed as:

$$D' = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{(\sigma_1^2 + \sigma_2^2)}{2}}} \tag{3.1}$$

On the DIAL2MSA corpora, which are likely more similar in style to AOC-ALDi, my model performs about as well as *Token DI*, the other BERT-based model (which, like mine, has not seen MGR in training), while also providing a wider range of scores (as shown in Figure 3.4). *Token DI* does somewhat better than my model on the Bible corpora, but again by making nearly binary judgments for each sentence.

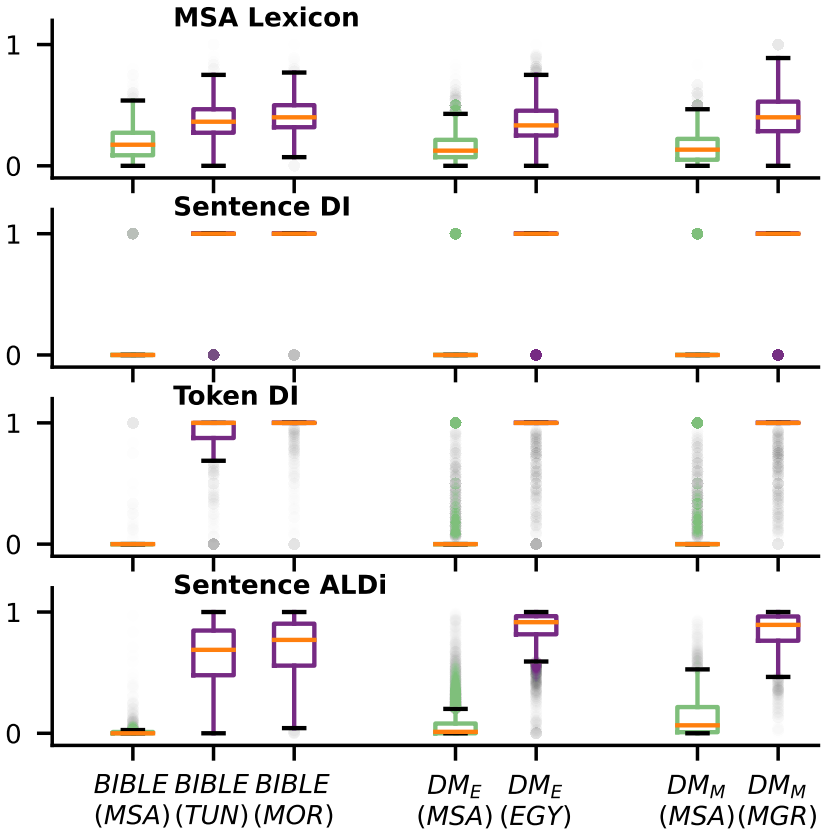


Figure 3.4: The distribution of the ALDi scores assigned by the four models to sentences of the Bible and DIAL2MSA corpora. Each column (across the four plots) represents the same set of sentences as scored by the four different models, and the columns are grouped by corpus to compare the different dialectal versions of that corpus. For each plot, the orange line shows the median score, the box represents the interquartile range (IQR) $[Q1, Q3]$ of the scores, the whiskers represent $\pm 1.5 * \Delta(IQR)$ beyond Q1 and Q3, and the dots represent outliers beyond this. **Note$_1$**: $\Delta(IQR) = Q3 - Q1$. **Note$_2$**: The boxplots for the *Token DI* and *Sentence ALDi* models are not significantly different across the multiple fine-tuning runs of different random seeds.

Model	Bible		DIAL2MSA	
	MSA / TUN	MSA / MOR	MSA / EGY	MSA / MGR
MSA Lexicon	1.28	1.55	1.48	1.73
Sentence DI	2.65	3.89	2.17	2.76
Token DI*	**3.81 ± 0.26**	**5.56 ± 0.34**	**5.83 ± 0.13**	3.93 ± 0.03
Sentence ALDi*	3.35 ± 0.09	3.89 ± 0.25	5.16 ± 0.13	**4.15 ± 0.1**

Table 3.9: The $D'(\uparrow)$ scores for the parallel MSA/DA corpora. **TUN**: Tunisian Arabic, **MOR**: Moroccan Arabic, **EGY**: Egyptian Arabic, **MGR**: Maghrebi Arabic. *: D' scores averaged across three fine-tuned models with different random seeds (30, 42, 50).

3.3.3 Analysis - Minimal Contrastive Pairs

Inspired by Demszky et al. (2021)'s corpus of minimal contrastive pairs for 18 distinctive features of Indian English, I built contrastive pairs of MSA and Egyptian Arabic (EGY) variants of a single sentence. I investigate five features of EGY that were previously recognized by Darwish et al. (2014). For each sentence, I generate versions with different gender markings (masculine and feminine) and word orders (SVO and VSO). While MSA allows for both word orders, it favors VSO (El-Yasin, 1985), while EGY favors SVO (Gamal-Eldin, 1968 as cited in Holes, 2013; Zaidan and Callison-Burch, 2014). In Table 3.10, I display the ALDi scores assigned by the different models to the contrastive pairs.

The *MSA Lexicon* model considers all dialectal features to have the same impact in assigning a non-zero ALDi score (i.e., $\frac{1}{3} \approx 0.33$ or $\frac{1}{2} \approx 0.5$) to the DA sentences. As implied by my previous experiment, the *Token DI* model acts as a sentence-level DI model, tagging all the tokens as dialectal if only one token shows a distinctive dialectal feature. This behavior might be an artifact of the model's fine-tuning dataset, where annotators were asked to use the surrounding context to determine an ambiguous token's language (EGY or MSA).

Conversely, the *Sentence ALDi* model provides a more nuanced distinction between the different features. The negation form (F4, F5) used in EGY seems to cause the model to categorically consider the sentence as highly colloquial. Less salient features such as the (F1) present progressive proclitic بـ increase the ALDi level of the sentence, but to a lesser extent than the negation feature. We also see that the model assigns higher ALDi scores to SVO sentences than to VSO, suggesting that the model may have

learned the common word order in EGY. Finally, feminine-marked sentences tend to get higher scores than their masculine-marked counterparts, which may be indicative of a gender bias in the training data and resulting model—if feminine marking is less common, it may also be seen as less standard language and interpreted as non-MSA.

Feature name	MSA_f	EGY_f	Word order	MSA LEX.		SEN. DI		TOK. DI *		SEN. ALDi *	
				MSA	EGY	MSA	EGY	MSA	EGY	MSA	EGY
F1) Present progressive	تقول البنت الحقيقة	بتقول البنت الحقيقة	VSO	0.0	0.33	0.0	0.0 / 1.0	1.0	1.0	0.1 / 0.12	0.86 / 0.56
En: The girl is saying the truth	البنت تقول الحقيقة	البنت بتقول الحقيقة	SVO	0.0	0.33	0.0	1.0	1.0	1.0	0.23 / 0.26	0.83 / 0.62
F2) Future Morpheme	ستقول البنت الحقيقة	هتقول البنت الحقيقة	VSO	0.0	0.33	0.0	0.0	0.0	1.0	0.0 / 0.07	0.76 / 0.9
En: The girl will say the truth	البنت ستقول الحقيقة	البنت هتقول الحقيقة	SVO	0.0	0.33	0.0	1.0 / 0.0	0.11 / 0.0	1.0	0.02 / 0.09	0.79 / 0.89
F3) Passive formation	قيلت الحقيقة	اتقالت الحقيقة	VO	0.0	0.5	0.0	0.0	0.0	1.0	0.05	0.36
En: The truth was said	الحقيقة قيلت	الحقيقة اتقالت	OV	0.0	0.5	0.0	1.0	0.0	1.0	0.11	0.36
F4) Negation	لا تقول البنت الحقيقة	مبتقولش البنت الحقيقة	VSO	0.0	0.33	0.0	1.0	0.17 / 0.0	1.0	0.08 / 0.11	0.95 / 0.91
En: The girl is not saying the truth	البنت لا تقول الحقيقة	البنت مبتقولش الحقيقة	SVO	0.0	0.33	0.0	1.0	0.25 / 0.33	1.0	0.09 / 0.11	0.91 / 0.9
F5) Negated imperative **En**: Do not say the truth	لا تقولي الحقيقة	ماتقوليش الحقيقة	VSO	0.0 / 0.33	0.5	0.0	1.0	0.0	1.0	0.0 / 0.13	0.84 / 0.91

Table 3.10: The ALDi scores assigned to contrastive MSA and Egyptian Arabic sentences. Only the feminine-marked version of the sentence is shown, and tokens with dialectal features are underlined. A single score is reported if a model assigns the same score to the masculine and feminine versions of a sentence; otherwise, the scores for masculine/feminine are shown. I tested VSO (favored in MSA) and SVO (favored in EGY) word orders. **Note:** Scores $\in [0, 0.11]$ are encoded in green, while ones $\in]0.11, 1]$ have a shade of purple.

*: The scores for these models are averaged across three fine-tuned models with different random seeds.

3.3.4 Analysis - How Do ALDi's Guidelines Compare to Habash et al.'s (2008) Guidelines?

The *Sentence-ALDi* model relies on the ALDi ratings derived from the AOC dataset. These ratings are assigned on the sentence level with minimal explanation of the different dialect levels. It would be useful to investigate how these sentence-level ratings compare to the levels proposed by Habash et al. (2008), where manually set rules are used to derive sentence-level dialectness labels from token-level ones. However, this necessitates annotating the same set of sentences according to the two guidelines.

ZAEBUC is a rich corpus of carefully transcribed Zoom meetings (20 hours in total) involving two students role-playing brainstorming ideas for a specific topic and discussing them with an interlocutor (Hamed et al., 2024). The utterances of the phase in which the two students discuss their ideas with the interlocutor have *dialect levels* labels manually assigned following Habash et al.'s (2008) descriptions of the segment levels. Yet, the guidelines were simplified by removing the word-level annotations criteria, as elaborated in Figure 3.5. Moreover, the provided annotations were not merely based on the transcripts, as the annotators were also asked to listen to the utterances.

L0 denotes perfect MSA.

L1 denotes imperfect MSA. This includes utterances with non-standard forms, such as syntax or morphology that is inclined towards dialects; however, it does not include any strong dialectal markers.

L2 denotes MSA-dialectal code-switching. This includes utterances having strong dialectal markers where the contribution of dialects is nearly equal to or less than MSA.

L3 denotes dialect with MSA incursions. The utterance is mostly dialectal, with some embedded MSA words.

L4 denotes pure dialect.

Figure 3.5: The annotation guidelines used to annotate the ZAEBUC corpus, which are adopted from Habash et al.'s (2008) guidelines.

Instead of annotating ZAEBUC's utterances according to Zaidan and Callison-Burch's (2011) guidelines, I use the *Sentence-ALDi* model as a proxy. More specifically, an ALDi score is estimated for each transcribed utterance. To this end, I preprocessed the transcribed utterances to discard the different tags that were used to indicate speech-

related features such as hesitation marks and non-verbal cues (e.g., pauses, gasps, and laughs). Additionally, I replaced the placeholder tags used to refer to the interlocutors' names with Arab female names (e.g., replacing <*speaker_1*> with حنين).

Although the *Sentence-ALDi* only relies on the textual transcript of the utterances, Figure 3.6 shows that the automatically estimated ALDi scores are correlated with the manually assigned *dialect levels*, even when the latter also relied on the speech signal. In fact, Spearman's correlation coefficient between the ordinal dialect levels and the continuous automatically estimated ones is *0.806*. This indicates that both guidelines are not fundamentally different.

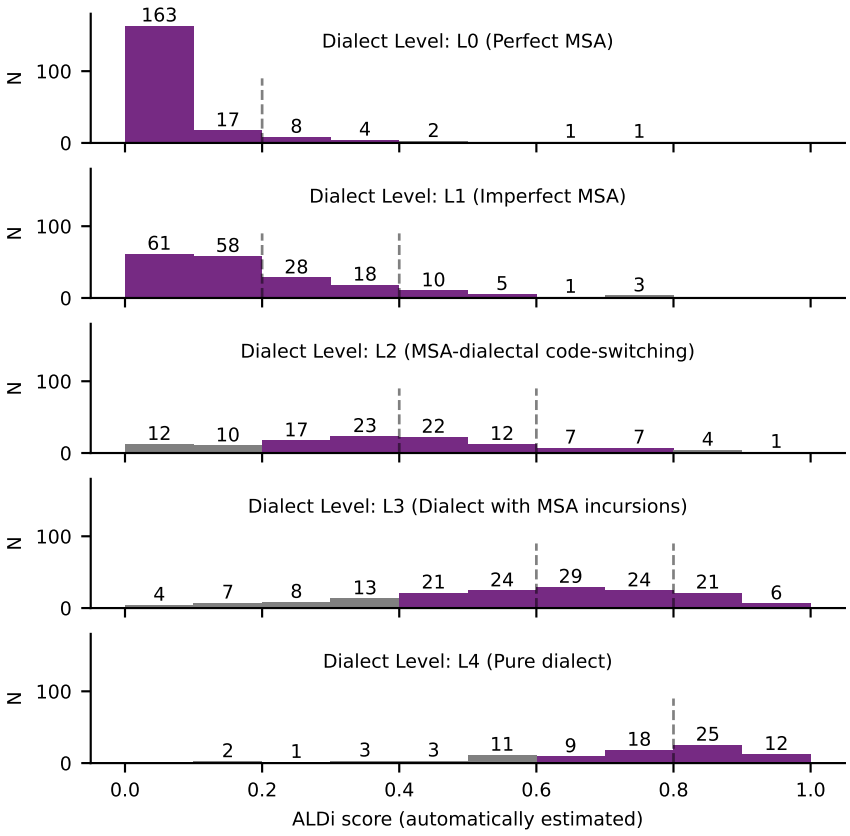


Figure 3.6: The histograms of the automatically estimated ALDi scores for the transcribed sentences of the ZAEBUC corpus (Hamed et al., 2024), split according to the utterances' manually assigned dialect levels (L0, L1, L2, L3, L4), according to a simplified version of Habash et al.'s (2008) guidelines. The dotted vertical lines show the boundaries of the expected range of ALDi scores for each dialect level, if the full ALDi range [0, 1] is uniformly split into five subranges, with each subrange mapped to a respective dialect level. **Note:** A Spearman's correlation coefficient of *0.806* exists between the ordinal dialect levels and the continuous automatically estimated ones.

Upon analyzing a sample of utterances where the automatically estimated ALDi scores diverge from the range of scores expected for the manually assigned dialect levels, I identified three main scenarios. Utterances corresponding to the different scenarios are provided in Table 3.11. The first scenario is for neutral utterances (U0, U1, and U5), where the whole utterance is valid in both MSA and DA. Without access to the speech signal, both scores that are close to zero (i.e., MSA) or indicate the sentences are not highly colloquial (i.e., ones not close to 1) are acceptable. The second is for utterances (U3, U4, and U5) with words/segments that do not come from MSA origins (triple underlined in the Table) and are highly colloquial words/segments. The ALDi scores for the first two utterances having two words with non-MSA origins are high (0.75 and 0.78, respectively), the model assigned a lower score of 0.39 to U9, which only had a single word of this sort (بس). The third and last scenario is for utterances (U2, U6, U7, U8, U10, and U11), where all the dialectal words/segments have MSA origins (double underlined), and hence do not strongly diverge from it. The estimated ALDi scores of these utterances seem to be related to the proportion of the dialectal and neutral segments in the sentence.

Hence, the *Sentence-ALDi* model is a handy tool for automatically estimating the dialect level of Arabic sentences. Analyzing sentences where the model's predictions diverge from the manually assigned dialect levels reveals that the absence of the corresponding speech signal is a significant reason, given that MSA and varieties of DA exhibit a non-negligible lexical overlap.

3.4 Reflection - Level of Dialectness and Formality

Formality is a concept that has been studied, yet it does not generally have an agreed-upon definition (Heylighen and Dewaele, 1999; Lahiri, 2016; Pavlick and Tetreault, 2016; Rao and Tetreault, 2018). Heylighen and Dewaele (1999) define formality as the avoidance of ambiguity by minimizing the context-dependence, and the fuzziness of the used expressions. Later operationalizations recognize factors such as slang words and grammatical inaccuracies have on the people's perception of formality (Mosquera and Moreda, 2012; Peterson et al., 2011) as cited in (Pavlick and Tetreault, 2016).

Arabic speakers indeed tend to use MSA in formal situations, and their regional dialects in informal ones. However, an Arabic speaker can still use MSA and speak informally, or use their dialect and speak formally. The case studies in §4.1 of chapter 4 show how Arab presidents use sentences of different levels of dialectness in their

Level	Preprocessed Utterance	ALDi	ID
L0	تشرفنا يا حنين أهلا وسهلا.	0.49	U0
	Okay. Thank you. مع السلامة.	0.65	U1
L1	ممتاز. شكرا يا حنين. فاطمة تحبي تضيفي شئ؟	0.65	U2
	السنع بلهجتنا العامية يعني نفس ال .. كيف ممكن أشرح السنع يعني نفس نفس الأدب أو الإحترام.	0.75	U3
	طيب أنا أرى هنا يا حنين تقريبا على الprofile بتاعك أو كده أرى رسمة مكتوب عليها حنين رسمة جميلة لبنت جميلة ، هل هذه من تصميمك؟	0.78	U4
L2	مع ألف سلامة، مع السلامة.	0.01	U5
	يعني ممكن تسعين بالمية أو خمسة وثمان خمسة وثمانين بالمية من محصول النفط في الإمارات يأتي من أبوظبي وليس من دبي وهذا وهذا دليل أن أن الإمارات ودبي نجحوا ليس بسبب النفط فقط.	0.04	U6
L3	صحيح، خصوصا إن فيه كثير من المصريين الآن يعيشون في دول الخليج.	0.01	U7
	يعني يكون الباب طويل ليعرض المانيكانات بشكل كامل	0.13	U8
	فيعني الذوق العام الحينه يميل كله حق الأشياء البسيطة، الموسيقى اللي تكون تجذب الإنتباه بس تكون هادية ما تكون موسيقى صاخبة	0.39	U9
L4	ي .. لا هم عندهم يعني عندهم مصممين مواطنين.	0.45	U10
	لأنه إذا إذا المنتشرة .. مثلا الحين مثلا مثلا Zara تاخذ من تصاميم شركات ثانية.	0.57	U11

Table 3.11: Sampled utterances from ZAEBUC with automatically estimated ALDi scores that diverge from the range of expected ALDi scores of their respective manually assigned dialect levels. MSA or non-Arabic segments are not underlined. A single underline is used for Neutral segments (i.e., ones valid in MSA and DA), double-underlined segments are dialectal segments of MSA origins, and triple-underlined segments are dialectal ones that highly diverge from MSA. **Note:** *L0* Perfect MSA, *L1* Imperfect MSA, *L2* Mixed MSA and Dialect, *L3* Dialect with MSA Incursions, *L4* Pure Dialect.

political speeches. While these speeches would all be considered to be formal, different levels of dialectness are employed to sound authoritative (using MSA) or seek sympathy (using a regional dialect). Therefore, I believe the level of dialectness and formality are related yet not interchangeable.

3.5 Summary

The linguistic theory of *Dialect levels* provides a more realistic explanation of the intraspeaker variation within Arabic-speaking communities, in contrast to the theory of Diglossia, which adopts a simplistic binarized view of such variation. To investigate *RQ1*, this chapter describes the first attempt to operationalize the concept of *Dialect levels* as a quantifiable variable that could be automatically estimated. To this end, I presented *Arabic Level of Dialectness (ALDi)*, a linguistic variable that quantifies the level of dialectness of an Arabic sentence. I released AOC-ALDi, a dataset of Arabic comments annotated with their ALDi scores. A BERT-based regression model fine-tuned on AOC-ALDi showed superior performance compared to existing baselines that are based on lexicons and DI models. My analysis shows that the model generalizes to various Arabic dialects. In addition, the model provides a nuanced distinction of dialectal features, which token and sentence DI models can not perform.

The AOC-ALDi dataset is a rich dataset of 127,835 Arabic sentences with *dialect* and *dialect level* annotations. However, it could still have some limitations. First, most of the sentences are online comments on news articles, which is a specific genre of text. Although my experiments show robustness across multiple text genres, it would be useful to have a dataset (even just for intrinsic testing) that comes from other sources, such as social media. Moreover, the gold-standard ALDi scores in my AOC-ALDi dataset are based on normalizing the level of dialectness annotations of the AOC dataset, which were collected by randomly routing the samples to three annotators from a pool of different Arabic speakers. As raised in *RQ3*, it is conceivable that a speaker's native dialect influences what they perceive as MSA and what they perceive as DA, hence, impacting their ALDi ratings. Lastly, the annotation guidelines are minimal. These limitations will be later addressed in chapter 6.

Chapter 4

Applications of Arabic Level of Dialectness

This chapter presents two applications where ALDi provides a more nuanced tool than only relying on DI, to address (***RQ2***): ***What are some applications of automatically estimating Dialect Levels, in text analysis and data annotation?*** The first illustrates how ALDi can reveal Arabic speakers' stylistic choices in different situations, a useful property for sociolinguistic analyses, which existing DI systems fail to detect. The second analyzes how ALDi can be a useful tool when annotating Arabic datasets. More specifically, it investigates the impact of randomly routing samples to annotators on the annotation agreement, and how ALDi can be used to improve the annotation process.

The work presented herein was reported in the following papers:

- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. *ALDi: Quantifying the Arabic Level of Dialectness of Text.* In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10597–10611, Singapore.
- Amr Keleg, Walid Magdy, and Sharon Goldwater. 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), pages 766–777, Bangkok, Thailand. Received an Outstanding Paper Award.

4.1 Automatically Analyzing Intraspeaker Variation

The same speaker can adapt different styles according to various social and linguistic factors (Kiesling, 2011). In this section, I present two case studies that analyze the transcribed speeches of three different Arab presidents. I highlight how quantitatively estimating ALDi can help in revealing different speaking styles, demonstrating ALDi's utility for sociolinguistic studies.

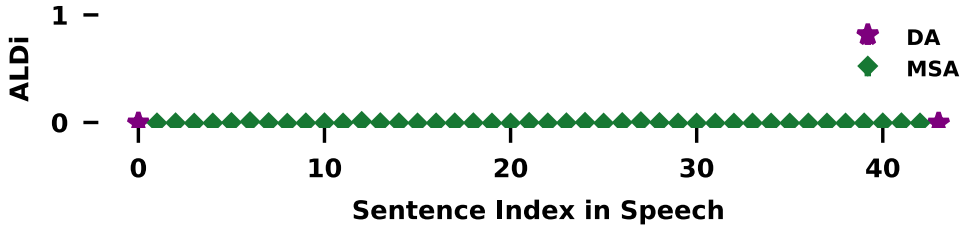
4.1.1 Presidential Speeches during the Arab Spring

Lahlali (2011) qualitatively analyzed the usage of MSA and DA (Tunisian Arabic and Egyptian Arabic) in the last three speeches of Ben-Ali and Mubarak, the former Tunisian and Egyptian presidents, during the period of the Tunisian and Egyptian revolutions in 2011. Mubarak consistently used MSA for his speeches to showcase authority and power. Ben-Ali used MSA for his first two speeches. For his last speech, he explicitly said: "نكلمكم لغة كل التونسيين والتونسيات" - "I talk to you in the language of all the Tunisians", apparently using his choice of dialect as a way to identify himself with a particular group (cf. Shoemark et al. 2017; McNeil 2022). Using different language levels or code-switching to non-standard varieties has also been previously noted as a strategy used by Gamal Abdul Nasser (a former president of Egypt) (Holes, 1995), and Boyko Borisov (a former prime minister of Bulgaria) (Kementchedjhieva, 2016).

I quantitatively replicate the analysis by visualizing the ALDi scores of the transcribed speeches. I scraped the speeches from online websites[1] and used the HTML line breaks <*br*> to segment them into sentences. For each sentence, I employ the *Sentence ALDi* and *Sentence DI* models—introduced in §3.3.1 of chapter 3—to estimate its ALDi score and classify it as DA or MSA, respectively.

Figure 4.1a shows that the *Sentence ALDi* correctly assigns almost-zero ALDi scores for the sentences of Ben-Ali's speech on the 10th of January, while the DI model makes a couple of errors (and similarly for Mubarak's speeches, shown in Figure 4.1c and Figure 4.1d). Both models identify the shift to DA in the second speech (Figure 4.1b), with more sentences identified as DA by the DI model, and many with moderate ALDi scores. Given the nature of the speech, Ben-Ali still used formal terms while speaking in Tunisian Arabic, which is likely the reason for the intermediate ALDi scores.

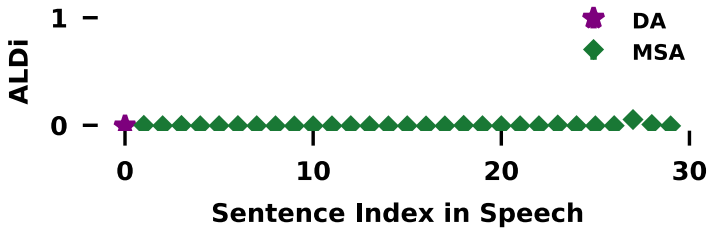
[1] `www.babnet.net` and `egypt-blew.blogspot.com`



(a) Ben-Ali's first speech (10/1/2011).

(b) Ben-Ali's second speech (13/1/2011).

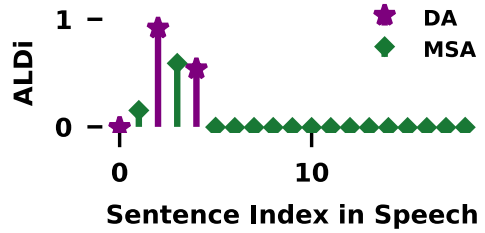
(c) Mubarak's first speech (1/2/2011).



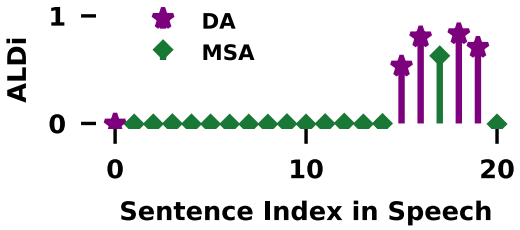
(d) Mubarak's second speech (10/2/2011).

Figure 4.1: The ALDi scores assigned to sentences of transcribed political speeches. Subfigures 4.1a and 4.1b represent two speeches of the former Tunisian president Ben-Ali during the Tunisian Revolution. Subfigures 4.1c and 4.1d represent two speeches of the former Egyptian president Mubarak during the Egyptian Revolution.

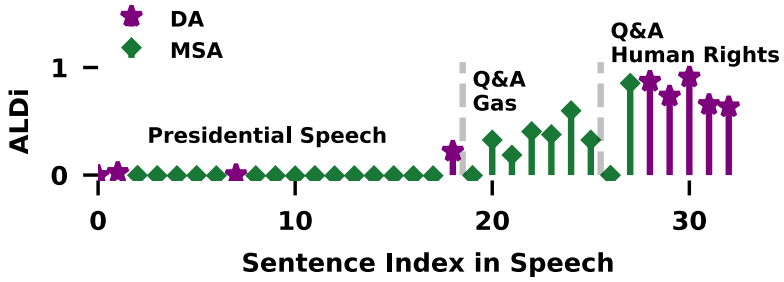
Note: The MSA/DA labels were automatically predicted by the *Sentence DI* model (refer to §3.3.1 of chapter 3).



(a) El-Sisi's speech (7/12/2017).



(b) El-Sisi's speech (10/10/2021).



(c) El-Sisi's speech (22/7/2022).

Figure 4.2: The ALDi scores assigned to sentences of transcribed political speeches of the current Egyptian president El-Sisi. **Note:** The MSA/DA labels were automatically predicted by the *Sentence DI* model (refer to §3.3.1 of chapter 3).

4.1.2 El-Sisi's Speeches

I also applied ALDi to 659 speeches of the current Egyptian president El-Sisi, scraped from `almanassa.com`. The transcripts are not limited to the edited presidential speech, but also include greetings, introductory comments, interventions by the audience, and signs of disfluency or hesitation. The site's editors segmented each speech into coherent sentences, embedded in <*p*> HTML tags, that I adopt as units of analysis.

While most of these speeches are conducted in MSA, multiple cases of code-switching between MSA and Egyptian Arabic occur. In Figure 4.2a and 4.2b, El-Sisi used MSA when reading edited speeches, and Egyptian Arabic with high ALDi scores when spontaneously addressing the audience before or after the edited speech.

Interestingly, Figure 4.2c shows three different ALDi levels as part of the same speech. El-Sisi used MSA for reading the edited speech directed to the press, discussing topics such as Egyptian-German diplomatic relations, climate change, and economic hardships. He then reacted spontaneously to two questions from the press. He attempted to answer the first question—related to gas prices—in MSA. However, the sentences show code-switching between MSA and Egyptian Arabic, indicated by intermediate ALDi scores (though the DI system does not identify these). For the second question about human rights in Egypt, El-Sisi uses sentences that are more colloquial and less formal, inviting the journalist to visit Egypt in order to make a fair assessment of the situation. This is indicated by even higher ALDi scores. Samples of each segment are in Table 4.1.

This speech is a clear example of how an Arabic speaker can adapt different levels of dialectness in their speech and indicates the ability of ALDi to reveal such differences.

Segment	Sentence with English translation (mine)	ALDi
Main Speech	واتفقنا على أن الوضع الحالي يفرض على كافة الفاعلين الدوليين التحلي بالمسؤولية لإيجاد وآليات عملية تخفف من تداعيات الأزمة على الدول الأكثر تضررًا. And we agreed that the current circumstances endeavor all actors to bear their responsibilities by finding practical solutions and mechanisms to mitigate the impact of the crisis on the most affected countries.	0
Q&A (Gas Prices)	وأنا عايز أقول إن إحنا تحدث، يعني أنا تحدثت في هذا الأمر إن المطلوب التنسيق والتعاون بين في ما يخص هذا الملف أثناء حديثي أو خطابي في مؤتمر جدة عن موضوع الطاقة تحديدًا. I spoke about this matter and that coordination and cooperation are required between all countries of the world regarding this topic during my talk or speech at the Jeddah conference, specifically on the issue of energy.	0.41
Q&A (Human Rights)	وإحنا مش مهتمين بيه عشان أنتوا بتسألوا عنه.. مهم قوي إن إنتوا تعرفوا كدا.. إحنا مهتمين بيه عشان إحنا بنحترم شعوبنا، وبنحبها، ومش كلام، إحنا بنحترم شعوبنا زي ما إنتوا ما بتحترموا شعوبكم.. وبالتالي إحنا مش مهتمين عشان أنتوا بتسألونا عليه.. لأ.. ده مسؤوليتنا الأخلاقية والتاريخية والإنسانية تجاه شعوبنا. دي نقطة. And we are not interested in it because you ask about it.. It is very important that you know this.. We are interested in it because we respect our people, and we love them, and these are not just words, we respect our people just as you respect your people.. and therefore we are not interested because you ask us about it. .. No.. This is our moral, historical, and humanitarian responsibility toward our people. This is one point.	0.75

Table 4.1: Three sentences of different estimated ALDi scores sampled from three segments of El-Sisi's speech on the 22nd of July 2022 shown in Figure 4.2c.

4.2 Improved Annotation Guidelines for Arabic Datasets

While MSA can be largely understood by most Arabic speakers, the different varieties of DA are not always fully mutually intelligible. Despite this mutual unintelligibility, a common practice when developing datasets for multi-dialect Arabic NLP is to randomly recruit annotators without regard to their dialect. However, routing dialectal content to speakers of a different dialect for annotation or moderation can present real problems. For example, it has been shown to contribute to unjust online content moderation of DA (Business for Social Responsibility, 2022), and racially biased toxicity annotation in American English varieties (Sap et al., 2022). Two recent studies of multi-dialect DA annotation showed that for annotating hate speech or sarcasm, respectively, annotators were more lenient (for hate speech) and more accurate (for sarcasm) when annotating sentences in their native dialect (Bergman and Diab, 2022; Abu Farha and Magdy, 2022). The authors of both studies made the same recommendation for creating new Arabic datasets, namely to first identify the dialect of each sample and then route it to appropriate annotators.

This recommendation is theoretically appealing, but presents practical difficulties as automatic dialect identification is challenging, and existing systems assume a single correct label when in fact some texts can be natural in different dialects, as I will show in chapter 5 and chapter 6. Moreover, the representation of native speakers of the different Arabic dialects on crowdsourcing sites might be skewed (Mubarak and Darwish, 2016). Hence, recruiting native speakers of some Arabic dialects might be challenging, given the tough conditions of the countries in which these dialects are spoken.

In this section, I address these challenges by building on the ALDi variable. I hypothesize that as sentences with low ALDi scores do not diverge much from MSA, they can still be understood and accurately annotated by most Arabic speakers, while this will be less true for sentences with high ALDi scores. If my hypothesis holds, then annotation can be made more efficient while maintaining accuracy, by routing samples with low ALDi scores to speakers of any dialect. Only high-ALDi samples need to be routed to native speakers of the appropriate dialect.

4.2.1 Methodology

Data I study the impact of ALDi scores on the annotators' agreement for publicly released Arabic datasets. I analyze datasets satisfying the following criteria:

- **Language**: Mixture of MSA and DA.

- **Variation**: Targeting multiple varieties of DA.
- **Annotators**: Speakers of different varieties of DA that are randomly assigned to the samples.
- **Tasks Setup**: Sentence-level classification.
- **Released Labels**: Individual annotator labels or the percentage of annotators agreeing on the majority-vote label.[2]

I searched for datasets on Masader, a community-curated catalog of Arabic datasets (Alyafeai et al., 2021; Altaher et al., 2022). Each dataset on Masader has a metadata field for the varieties of Arabic included. I discarded the datasets that only included MSA samples, and manually inspected the remaining 151. After identifying 28 potential datasets that satisfy the criteria above, I contacted the authors of the datasets that do not have the individual annotations publicly released. Eventually, I had 15 datasets to analyze, listed in Table 4.2, covering: Offensive Text Classification, Hate Speech Detection, Sarcasm Detection, Sentiment Analysis, Speech Act Detection, Stance Detection, and Dialect Identification.

Analysis For each dataset, I compute the Arabic Level of Dialectness (ALDi) score for each annotated sample (sentence) using the Sentence-ALDi model (chapter 3), which returns a score from 0 (MSA/non-dialectal) to 1 (strongly colloquial). To investigate the effect of ALDi on annotator agreement, I bin the samples by their ALDi score into 10 bins of width 0.1. I compute *% full agree*, the percentage of samples in that bin for which all the annotators agreed on a single label. I employ Pearson's correlation coefficient to analyze the relation between ALDi (represented by each bin's midpoint ALDi score) and *% full agree*, and also report the slope of the best-fitting line as a measure of the effect size.[3] As aforementioned, my initial hypothesis is that *% full agree* negatively correlates with ALDi scores.

[2]For some datasets, the percentage of annotators agreeing on the majority vote is weighted by their performance on the annotation quality-assurance test samples. This distinction is irrelevant to this study, where I only consider whether all annotators agreed or not.

[3]The exact values of the slopes and correlation coefficients depend on the number of bins. However, I got similar qualitative results on using 4 or 20 equal-width bins. 10 bins are enough to check if trends are non-linear while keeping a reasonable number of samples in the smallest bins. I also fitted logistic regression (*logreg*) models using ALDi as a continuous variable and a binary outcome *Full Agreement (Yes/No)* for each sample. Both analysis tools reveal similar patterns (See §4.2.5), but the binning method provides useful additional visualization.

Dataset	Task (# labels)	%ALDi <0.1	Description
Deleted Comments Dataset (DCD) (Mubarak et al., 2017)	Offensive (3)	62.57%	About 32K deleted comments from `aljazeera.com`. Confidence scores for the majority vote of 3 annotations are provided.
MPOLD (Chowdhury et al., 2020)	Offensive (2)	27.82%	4000 sentences interacting with news sources, sampled from Twitter, Facebook, and YouTube, annotated three times.
YouTube Cyberbullying (YTCB) (Alakrot et al., 2018)	Offensive (2)	10.24%	15,050 comments and replies to 9 YouTube videos labeled by 3 annotators (Iraqi, Egyptian, Libyan).
ASAD (Alharbi et al., 2021)	Sentiment (3)	35.63%	95,000 tweets with a skewed representation toward the Gulf area and Egypt.
ArSAS (Elmadany et al., 2018)	Sentiment (4) Speech Act (6)	57.45%	21,064 tweets related to a pre-specified set of entities or events, with confidence scores for the majority votes across three annotations per sample.
ArSarcasm-v1 (Abu Farha and Magdy, 2020)	Dialect (5) Sarcasm (2) Sentiment (4)	57.44%	10,547 tweets, sampled from two different Sentiment Analysis datasets: ATSD (Nabil et al., 2015), SemEval2017 (Rosenthal et al., 2017), reannotated for Sentiment, Dialect, and Sarcasm.
Mawqif (Alturayeif et al., 2022)	Sarcasm (2) Sentiment (3) Stance (3)	58.04% 58.04% 57.99%	4,121 tweets about "COVID-19 vaccine", "digital transformation", or "women empowerment" annotated separately for stance and sentiment/sarcasm till the label confidence reaches 0.7 (min. 3 annotators) or 7 annotators label the sample.
iSarcasm's test set (Abu Farha et al., 2022)	Dialect (5) Sarcasm (2)	30.5%	200 sarcastic sentences provided by crowdsourced authors and 1200 non-sarcastic tweets from ArSarcasm-v2 (Abu Farha et al., 2021) reannotated 5 times.
DART (Alsarsour et al., 2018)	Dialect (5)	0.8%	24,279 tweets with distinctive dialectal terms annotated three times for the dialectal region. Samples of complete disagreement are not in the released dataset.

Table 4.2: The datasets included in my study. All datasets have three annotations per sample, except for iSarcasm (5 annotations/sample) and Mawqif (3 or more annotations/sample). For the labels used in each dataset and the proportion of each label, see Table B2 in the Appendix. For some datasets, there is a discrepancy between the number of samples listed in the paper and the raw data files with individual labels (See §B of the Appendix).

4.2.2 Results and Discussion

I use scatter plots to visualize the relation between *% full agree* and ALDi on the studied datasets, as shown in Figure 4.3. Additionally, the histograms of samples across the different bins indicate the dialectal content within the dataset. As per Table 4.2, 6 datasets out of the 15 have more than 50% of the samples with ALDi scores less than 0.1, which are expected to be written in MSA. However, I found that the overall trends depicted in Figure 4.3 will not be affected if I discard these samples with low ALDi scores and only focus on the rest.

For non-DI tasks, ALDi negatively correlates with agreement. Inspecting the trends depicted in Figure 4.3, strong negative Pearson's correlation coefficients exist for 8 out of the 12 datasets for the non-DI tasks (sentiment analysis, sarcasm, hate speech, and stance detection). Both the trends (quantified by the slope m) and the correlation coefficients for most of the tasks indicate that the percentage of samples for which all the annotators assign the same label decreases as the ALDi scores increase, often by a large margin.[4] I notice different trends for DI that I will elaborate on below.

For DI, agreement is lowest for mid-range ALDi scores (if MSA is a possible label) or low ALDi scores (if it is not). By definition, MSA sentences have an ALDi of 0, and normally, the ALDi estimation model assigns them very low scores.

For the ArSarcasm-v1 and iSarcasm datasets, the set of labels for the DI task includes MSA (i.e., some sentences in these datasets are not dialectal). For both datasets, one notices high percentages of agreement scores for the bin having ALDi scores $\in [0, 0.1]$ (generally agreeing that the label is MSA). The percentages decrease for the few succeeding bins, before rising again for the bins with high ALDi scores. Sentences of high ALDi scores (e.g., $\in [0.8, 1]$) are expected to have multiple dialectal cues, which increases the chance of attributing them to a single dialect. For sentences of intermediate ALDi scores, annotators can agree that a sentence is not in MSA. However, they would struggle to determine the dialect(s) of the sentence, which is reflected in having lower percentages of full agreement for these bins.

The authors of the DART dataset do not include MSA in DART's label set since they curated sentences with distinctive dialectal terms. This explains the low percentage of full agreement for the bin of ALDi scores $\in [0, 0.1]$, unlike the other two DI datasets.

[4] Refer to §4.2.4 for a possible explanation of the unexpected trends of the ArSAS dataset.

However, the pattern of having higher full agreement percentages for bins with higher ALDi scores still holds.

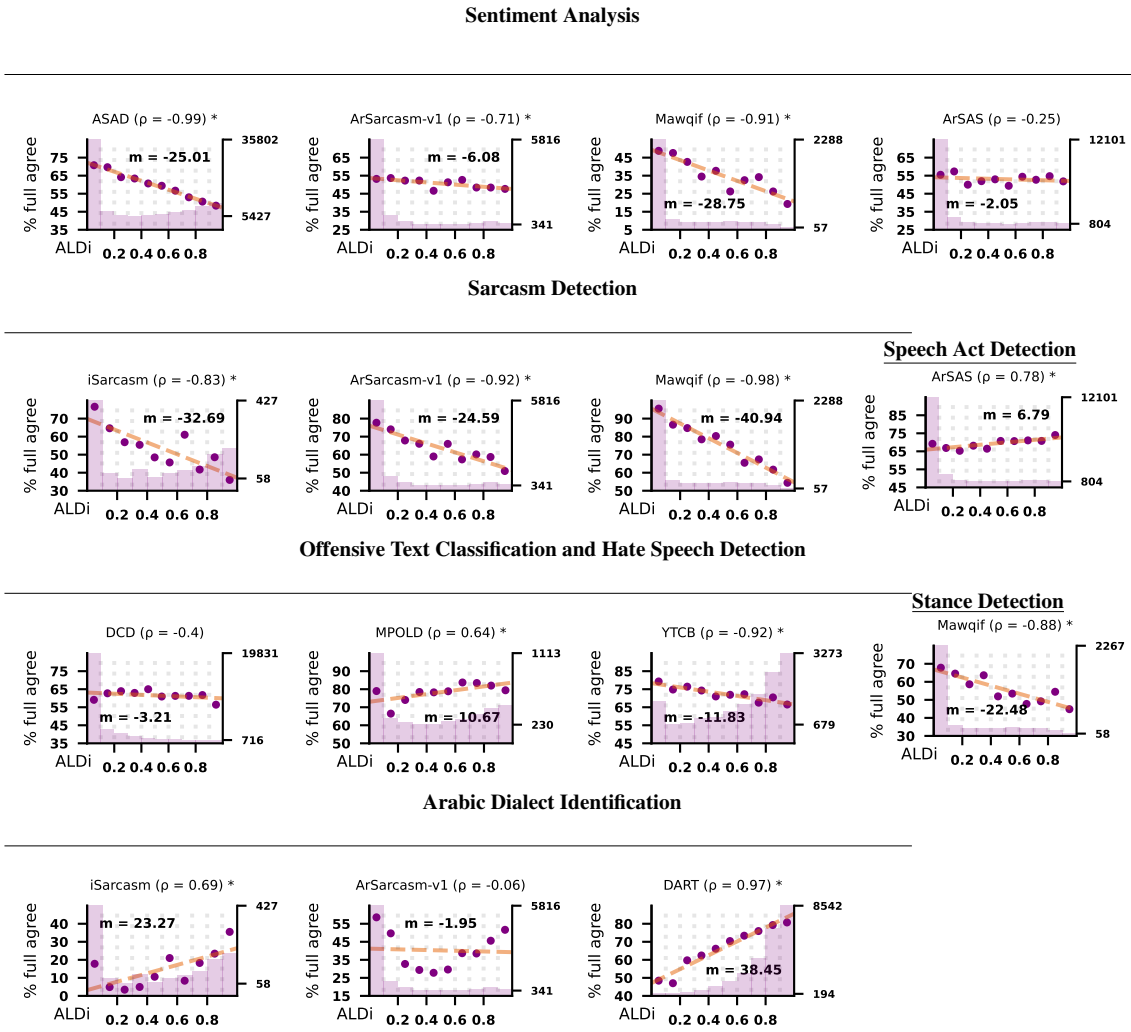


Figure 4.3: Scatter plots showing the relationship between binned ALDi scores (x-axis) and the percentage of samples with full annotator agreement (y-axis). The histogram represents the # of samples per bin (with min and max values for any bin labeled on the right-hand axis). The slope of the best-fitting line (m) is shown, and to enable visual comparison of slopes, all plots have the same y-axis scale (possibly shifted up or down). **Note:** Statistically significant ($p<0.05$) correlation coefficients (ρ) are marked with *.

4.2.3 Analysis of Trends by Class Label

A more nuanced analysis of the non-DI datasets can be done by splitting the samples according to their majority-vote labels. Figures 4.4, 4.5, 4.6, 4.7, and 4.8 visualize the impact of ALDi on the annotator agreement after splitting the samples according

to their majority-vote labels. Please note that the number of samples in the bins for some classes is not enough to draw concrete conclusions (e.g., samples with high ALDi scores for the *Neutral* class of the *ArSAS* and *Mawqif* datasets as per Figure 4.5).

The declining trend for agreement as ALDi scores increase is consistently salient for the negative class of the different tasks (i.e., *Not Sarcastic*, *Neutral* (sentiment), *Non-offensive*). One explanation is that agreeing on one of these labels requires fully understanding the sample, which is expected to be harder for non-native speakers of the dialects of sentences with high ALDi scores. Moreover, the general usage of MSA for formal communications and DA for more personal ones might be biasing the annotators. Further controlled analysis is required to investigate these intuitive explanations.

Conversely, the presence of specific words might be a strong cue for the label of the overall sentence, which might be useful for the positive classes (i.e., *Sarcastic*, *Positive/Negative*, *Obscene/Offensive*). Consequently, annotators might be able to agree on a label for a sentence even if it contains some unintelligible segments. Table 4.3 shows five examples, in which the annotators can use cues to label the whole sentence. Consider the first offensive sentence in the Table. The presence of the MSA insult يا حشرة (you insect) is enough to guess that the sentence is offensive, even if the remaining segment is not fully intelligible. Lastly, note that agreeing on a label does not imply it is accurate, especially when relying on cues for annotation.

Dataset	Sample	Translation of the Underlined Cue	ALDi	Individual Labels
YTCB Hatespeech Detection	كاع واش كاينة و لا حتا حاس يا حشرة المغرب ما عرفكش بيك	you insect	0.98	HateSpeech(3x)
	عمر تروح فدوة للتراب الي اتدوس علية شيعية العراق يا عرب يا جرب	you scabies	0.81	HateSpeech(3x)
	ماكان ناقص كاظم الا بلبل عفآ جحش الخليج يحكي عنو هههههههههههههههه	Gulf's colt	0.98	HateSpeech(3x)
ArSarcasm-v1 Sentiment Analysis	ريكي مارتن كان قبل يعجبني بس بعد كلامه حسيته مقرف يلوع بالجبد	I found him disgusting	0.81	Negative(2x) Positive(1x)
	اطفال دول ولا مش اطفال ؟؟؟ظظ يابتوع #المدارس هههههههههههههههههههه	hahaha	0.94	Positive(2x) Negative(1x)

Table 4.3: Five qualitative samples of high ALDi scores. The underlined segments represent the cues that the annotators might have used to choose a label even if they do not fully understand the sentence. Despite the presence of these cues, the annotators still disagreed on labeling the last two samples. I only provide translations for the underlined cue segments.

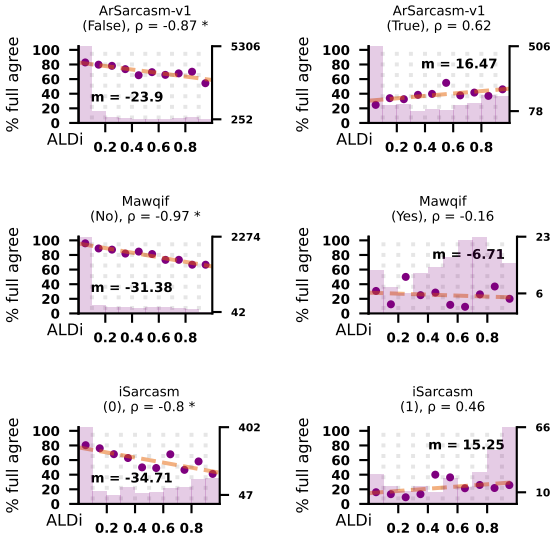


Figure 4.4: The trends for the classes of the Saracasm Detection datasets. Statistically significant correlation coefficients (ρ) are marked with *.

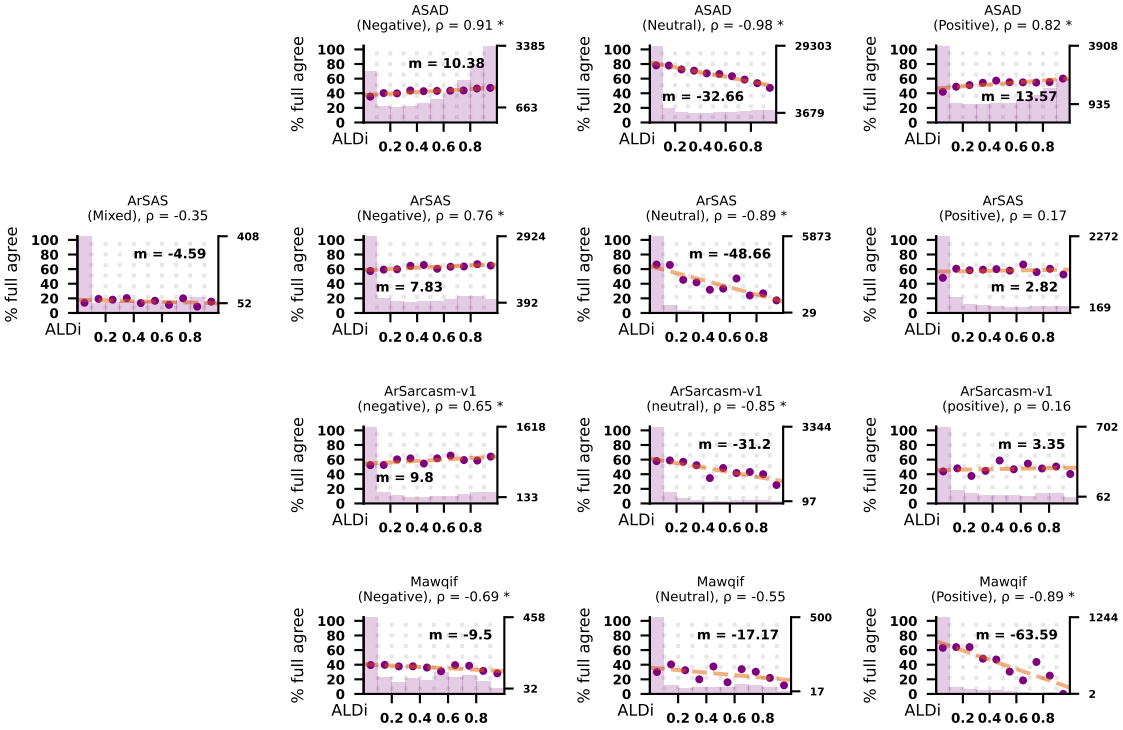


Figure 4.5: The trends for the classes of the Sentiment Analysis datasets. Statistically significant correlation coefficients (ρ) are marked with *.

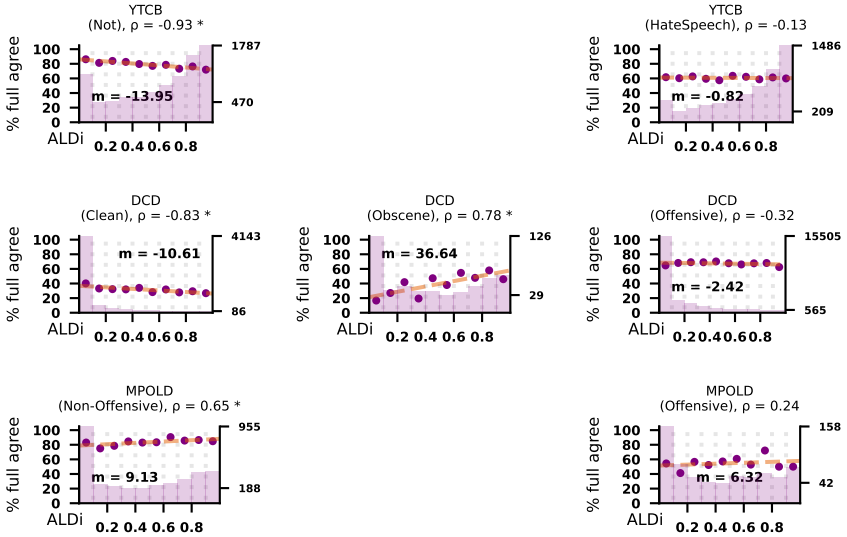


Figure 4.6: The trends for the classes of the Offensive Text Classification and Hate Speech datasets. Statistically significant correlation coefficients (ρ) are marked with *.

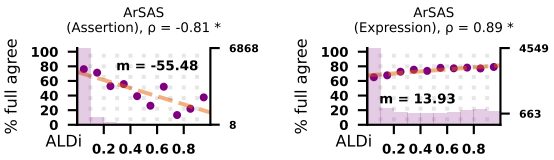


Figure 4.7: The trends for the *Assertion* and *Expression* labels of the ArSAS dataset, which represent 95% of the dataset samples. Statistically significant correlation coefficients (ρ) are marked with *.

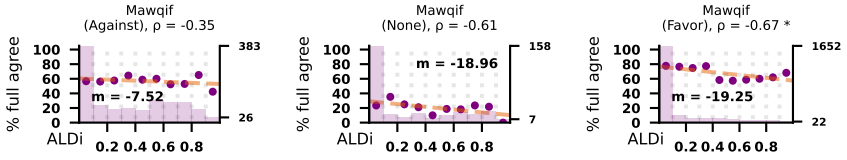


Figure 4.8: The trends for the classes of Mawqif's Stance dataset. Statistically significant correlation coefficients (ρ) are marked with *.

4.2.4 The Anomaly Trends of ArSAS and MPOLD

The ***ArSAS*** dataset stands out as a dataset with a rising trend for the *Speech Act Detection* task and a falling trend for the *Sentiment Analysis* task. Samples of *ArSAS* were jointly annotated for their sentiment and speech act. Despite having six different speech acts, which would arguably make speech act detection harder than sentiment analysis, the *Assertion* and *Expression* classes represent 95% of the samples. Looking at their respective trends shown in Figure 4.7, the two acts show two different behaviors. Most of the assertive samples have ALDi scores <0.2 (arguably, all are MSA ones). Moreover, the number of *Assertion* samples with high ALDi scores is not enough to estimate the *% full agree* for their respective bins. Conversely, the *Expression* act shows higher agreement as the ALDi score increases.

The creators of *ArSAS* noticed that most of the *Assertion* samples were annotated as *Neutral*, while most of the *Expression* samples had polarized sentiment (mostly *Negative*). The annotators might have treated the *Assertion* class as the act for *Objective* sentences, while treating *Expression* as the act for *Subjective* sentences. This is arguably easier than sentiment analysis, which might explain why annotators agree more on the Speech Act label than the Sentiment label for the *ArSAS* dataset. Further analysis is required to explain the trends of this dataset.

MPOLD is the only dataset in Figure 4.6 where the percentage of samples with full agreement increases the higher the ALDi scores are, for the *non-offensive/clean* class. Unlike the two other datasets (*YTBC* and *DCD*), the majority of *MPOLD*'s samples belong to the *non-offensive* class. The annotators could have noticed this label imbalance, treating the *non-offensive* class as the default label, superficially minimizing the reliance on intelligibility to choose this label.

4.2.5 Further Remarks on the Methodology

As described in §4.2.1, each dataset's samples were split into 10 bins of equal width according to their respective ALDi scores. Afterward, the correlation between each bin's midpoint ALDi score and the percentage of samples having full agreement *% full agree* was computed. For each bin, *% full agree* represents the Maximum Likelihood Estimation (MLE) for the probability that all the annotators agree on the same label for the samples of this bin.

Inability to use Interannotator Agreement metrics for some datasets Automated metrics such as Fleiss' Kappa (Fleiss, 1971) attempt to measure the Interannotator Agreement (IAA) while accounting for the random agreement/disagreement between annotators. In principle, it might be possible to perform a version of my analysis using Fleiss' Kappa rather than *% full agree* as the dependent variable. However, computing Fleiss' Kappa would require knowledge of the individual annotations for each sample. Such annotations are not available for the ArSAS (Sentiment/Speech Act), DART, and DCD datasets as described in Table B2. Since I wanted to include as many datasets as possible, I used % full agree instead.

Logistic regression as an alternative analysis tool Binning the data leads to a loss of analytical information, which might impact the results of the analysis, especially if implausible bins' boundaries are used (Wainer et al., 2006).

Logistic regression with binary outcomes is an alternative analysis that alleviates the limitations of binning. Each sample has a continuous ALDi score as the independent variable, and a binary outcome *Full Annotator Agreement (Yes/No)*. After fitting a logistic regression model to predict the binary outcome, the coefficient of the ALDi variable measures the impact of ALDi on the odds of full agreement. If this coefficient is negative, then the odds of full annotator agreement decrease as the ALDi score increases.

Figure 4.9 demonstrates the probability of full agreement of each dataset, in addition to the coefficient of the ALDi score with its 95% confidence interval. For the 8 non-DI datasets with $Coef_{ALDi} < -0.2$, the coefficients can be considered to be statistically significant since the confidence interval does not include zero.

Both analysis tools (correlation analysis and logistic regression) achieve similar results. The same 8 non-DI datasets—ASAD, ArSarcasm-v1 (Sentiment/Sarcasm), Mawqif (Sentiment/Sarcasm/Stance), iSarcasm, and YTCB—have significantly strong negative correlation coefficients as in Figure 4.3, and statistically significant coefficients for the ALDi variable, which are less than -0.2. However, binning the data allows for visualizing the *% full agreement* as a scatter plot, which can reveal whether the relation between ALDi and the agreement is linear or not, in addition to having a visual way for determining how well the best-fitting line models the data.

Impact of data skewness MSA samples are over-represented in some of the considered datasets. However, this is generally unproblematic for the analysis, so I opted not

to discard the MSA samples. For the method described in § 4.2.1, the samples of each bin are independently used to estimate the MLE of full agreement between annotators. Therefore, the over-representation of MSA samples in some datasets does not impact my analysis.

4.2.6 Reflections and Implications

My analysis only considers datasets that are labeled by randomly routing the samples in various dialects to speakers of different dialects. I show that the interannotator agreement decreases as ALDi increases. My intuition is that high-ALDi samples in a dialect are less probable to be fully intelligible to speakers of other dialects compared to the low-ALDi ones—the ones not too divergent from MSA. One might wonder whether a similar trend exists when the annotators speak the same dialect as the samples they label. I think that high-ALDi samples in a specific dialect have no inherent feature that would make them harder to annotate by speakers of this dialect than the low-ALDI ones. In the improbable case that the same trend would exist, then the recommendation that high-ALDi samples need to be handled with more care than low-ALDi ones would still hold. It would be interesting to run a controlled experiment in which the samples of a specific dialect, having different ALDi scores, are labeled by two groups of annotators: a group who speak this dialect, and another group who speak other dialects.

Furthermore, more thorough analyses need to be done to understand how ALDi affects each task, given its unique nature. Knowing the demographic information about the annotators might have allowed for revealing deeper insights into how speakers of specific Arabic dialects understand samples from other dialects. However, this would have required running a controlled experiment re-annotating the 15 datasets, which I hope future work will attempt.

Lastly, I acknowledge that there are multiple reasons for the annotators to disagree, which include the task's subjectivity, the annotators' background, and their worldviews (Uma et al., 2021). However, these factors would have less impact on the annotators' disagreement if a sample is not fully intelligible.

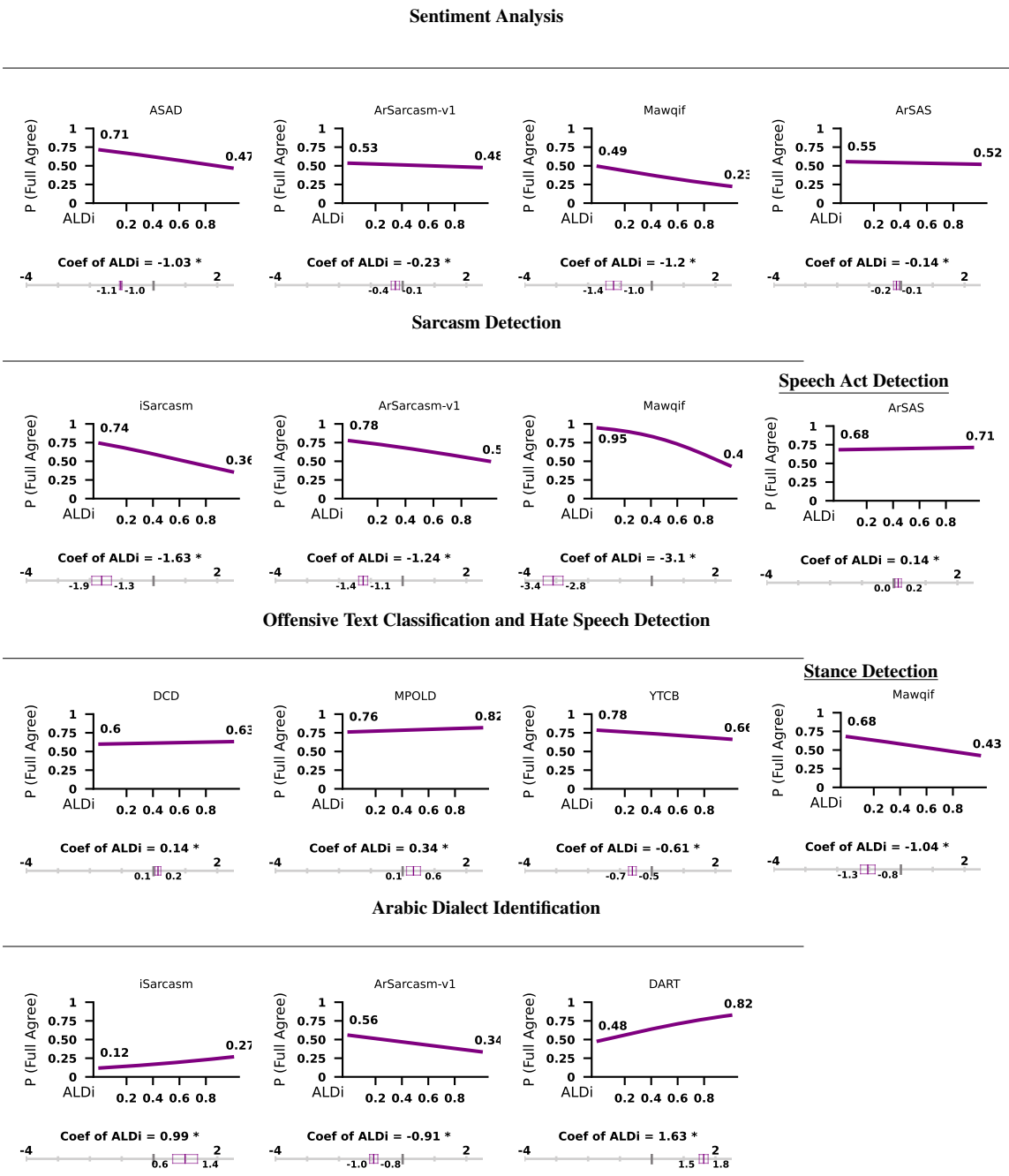


Figure 4.9: For each dataset, plots show the estimated probability of *full agreement* according to each dataset's fitted logistic regression model. Under each plot, the coefficient of ALDi with its 95% confidence interval is visualized. Nearly all datasets (marked with *) have confidence intervals that do not include zero, meaning the effect of ALDi is statistically significant at $p < 0.05$. Negative coefficients indicate that higher ALDi scores predict lower agreement.

4.3 Summary

In this chapter, I first introduce multiple case studies that demonstrate the effectiveness of ALDi in revealing new insights in Arabic text by analyzing the speeches of different Arab presidents. This presents ALDi as a tool for analyzing the various styles that Arabic speakers could employ, which could potentially be of great use for studies in sociolinguistics and computational social sciences.

By analyzing 15 datasets, I then find strong evidence of a negative correlation between ALDi and the full annotator agreement scores for 8 of the 12 non-Dialect Identification datasets. Moreover, for the 3 Dialect Identification datasets, I find that annotators have higher agreement scores for samples of higher ALDi scores, which by definition would have more dialectal features. The combination of more dialectal features in a sentence is more likely to be distinctive of a specific dialect.

Previous research recommended routing samples to native speakers of the samples' dialects for better annotation quality. My analysis indicates that a large proportion of the 6 datasets are samples with ALDi scores < 0.1, which are expected to be MSA samples that can be routed to speakers of any Arabic dialect. Moreover, the lower agreement scores for samples with high ALDi scores show that extra care should be given to these samples. Dataset creators should prioritize routing high-ALDi samples to native speakers of the dialects of these samples, for which the dialects can be automatically identified with higher accuracy, as these samples show more dialectal cues.

The following chapters will investigate the other dimension of variation, *Inter-speaker Variation*. More specifically, I will investigate the drawbacks of the long-lasting idea of modeling Arabic Dialect Identification (ADI) as a single-label classification task. This addresses the *Disjointedness* limitation, where a dialectal sentence is assumed to only be valid in one dialect, be it on the regional, country, or the province/city level.

Chapter 5

Limitations of Single-label Arabic Dialect Identification (ADI)

In chapter 3, I introduced ALDi, a continuous variable that quantifies a sentence's level of dialectness. However, ALDi is a dialect-agnostic metric. Hence, it can not discriminate between the different varieties of DA. Automatic Arabic Dialect Identification (ADI) of text has been used to discriminate between the varieties of DA. ADI has gained great popularity among the Arabic NLP community since its introduction in the early 2000s. Multiple datasets were developed, and yearly shared tasks have been running between 2018 and 2024. Despite attracting lots of attention and effort for over a decade, ADI is still considered challenging, especially for the fine-grained distinction of micro-Arabic dialects at the country and city levels. A lot of the papers conclude by calling for better algorithms and computational models to mitigate the inability of the current ADI systems to achieve high macro-F1 scores. In contrast, I argue that the currently adopted framing of ADI as a single-label classification task is the main limitation, especially for short sentences that might not have enough distinctive cues of a specific dialect as per Table 5.1. This chapter investigates this assumption, taking the first step to answer (***RQ4***) ***How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?***

With ***RQ4*** in mind, I first focus on studying the implications of adopting the single-label framing on the model evaluation process in this chapter. The subsequent chapter then introduces the first multi-label country-level ADI dataset, with manually assigned validity labels for nine different dialects. The main contributions of this chapter can be summarized as follows:

1. Highlighting the limitation of framing ADI as a single-label classification task by empirically estimating the *Maximal Accuracy* for multiple existing ADI datasets.
2. Performing an error analysis for an ADI model by recruiting native speakers of seven different country-level Arabic dialects.
3. Presenting a detailed proposal for how ADI could be framed as a multi-label classification task, covering various aspects such as data curation, annotation, and model evaluation.

Dialects	**Sentence**
Iraq, Jordan, Lebanon, Libya, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Yemen	وين المحطة؟ **Where is the station?**
Iraq, Morocco, Qatar	شنو رقم الرحلة؟ **What is the flight/trip number?**

Table 5.1: The MADAR corpus (Bouamor et al., 2018) has English/French sentences manually translated into different Arabic dialects. The table shows two sentences having the same translation across multiple country-level dialects. Hence, these sentences should not be assigned only a single dialect label if they are used in an ADI dataset.

The work presented herein was reported in the following paper:

- Amr Keleg and Walid Magdy. 2023. *Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification.* In Proceedings of ArabicNLP 2023, pages 385–398, Singapore (Hybrid).

5.1 Background

Language Identification (LI) is the task of identifying the language of a piece of text. For one language, Dialect Identification (DI) makes a finer distinction, aiming at identifying the dialect of a piece of text. Hence, LI and DI could be considered to be the same computational task, yet at different levels of granularity. So far, DI systems have employed the same algorithms and features used for LI, as indicated in a recent survey:

> "As evidenced in this article, from a computational perspective, the algorithms and features used to discriminate between languages, language varieties, and dialects are identical." (Jauhiainen et al., 2019, p. 677)

LI is computationally modeled as a single-label classification task, based on the assumption that "labels are non-overlapping and mutually exclusive, meaning that a text can only be written in one language" (Jauhiainen et al., 2019, p. 678). While this assumption would mostly hold for LI—except for closely related languages—it is apparent that it does not extend to DI, where a piece of text could be valid in multiple dialects. Assigning a **single dialect label** to each sentence, either automatically (e.g., using geotagging) or manually, makes the labels incomplete. This, in turn, could affect the fairness of the evaluation process. Broadly speaking, the need for improving how DI is framed, and consequently the accompanying resources, was previously noted by Althobaiti (2020), who concluded the *Future Directions* section of her survey of Arabic Dialect Identification (ADI) with the following:

> "There is also a need to criticize the available resources and analyze them in order to find the gaps in the available ADI resources."

For country-level ADI, a few qualitative analyses showed that the single-label framing is not suitable, especially since sentences could be valid in the dialects spoken in geographically proximate countries (Bayrak and Issifu, 2022; Khered et al., 2022), with a few proposing framing it as a multi-label task (Kchaou et al., 2019; Touileb, 2020). However, these recommendations did not get much notice from the community. To quantitatively analyze the limitations of the single-label framing of DI, Bernier-Colborne et al. (2023) used the Levenshtein edit ratio to identify near-duplicate sentences of a French DI dataset (**FreCDO**, Găman et al., 2023) that are assigned different dialect labels. They found that about 1% of the dataset are ambiguous sentences (i.e., sentences valid in multiple dialects). Olsen et al. (2023) applied the same technique to the test set of (**MADAR26**, Bouamor et al., 2019), finding that 6.3% of its sentences are ambiguous. However, Bernier-Colborne et al. (2023) noted that their automated method only provides a lower bound estimate for the proportion of ambiguous sentences, with manual human assessment required to accurately estimate this proportion.

5.2 How were the Existing Sentence-level Single-label ADI Datasets Built?

This section overviews the main data annotation methods used to create single-label ADI datasets and their key limitations. Several efforts (listed in Table 5.2) have built ADI datasets using various techniques, which can be grouped into four main approaches:

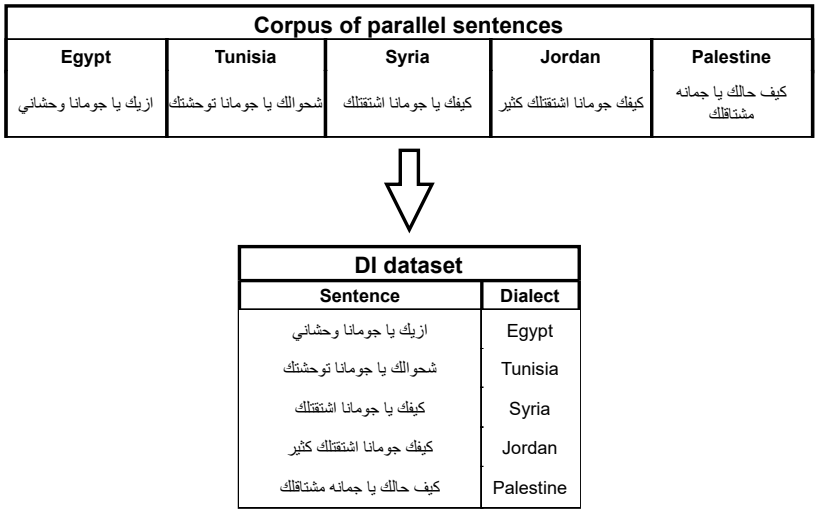


Figure 5.1: A demonstration of how parallel dialectal sentences are transformed into DI samples. The parallel sentences are sampled from the MPCA corpus (Bouamor et al., 2014).

(1) Manual Human Annotation where annotators categorize Arabic sentences into one dialect from a predefined list of dialects (Zaidan and Callison-Burch, 2011; Huang, 2015; Malmasi et al., 2016; Zampieri et al., 2017, 2018).
Limitations: It was found that annotators over-identify their own native dialects (Zaidan and Callison-Burch, 2014; Abu Farha and Magdy, 2022). Therefore, the annotations for sentences that are valid in multiple dialects might be skewed toward the countries from which most of the annotators originate, causing a representation bias. Moreover, accurately determining the Arabic dialect of a sentence requires exposure to the different dialects of Arabic, which might not be a common case for Arabic speakers.

(2) Translation in which participants are asked to translate sentences into their native Arabic dialects (Ho, 2006; Bouamor et al., 2014; Meftouh et al., 2015; Bouamor et al., 2018; Mubarak, 2018). If all the participants are asked to translate the same source sentences, then the dataset is composed of parallel sentences in various dialects. The main application of these datasets is to help develop machine translation systems; however, they are sometimes used for ADI. Figure 5.1 demonstrates how a corpus of parallel sentences is transformed into a corresponding DI dataset.
Limitations: While the labels of the corresponding DI dataset are correct, a source sentence might have the same translation in multiple Arabic dialects, as previously shown in Table 5.1. In such cases, a single-label classifier is asked to predict different

Dialect labels for the same duplicated input sentence.

Moreover, the syntax and lexical items in the translated sentences might be affected by the corresponding syntactic and lexical features of the source sentences, especially if the source sentence is MSA or a variety of DA (Bouamor et al., 2014; Harrat et al., 2017). Such effects might make the translated sentences sound unnatural to native speakers of these dialects.

(3) Distinctive Dialectal Terms where text is curated based on the appearance of a term from a seed list of distinctive dialectal terms. These terms are used to automatically determine the dialect of the text (Alsarsour et al., 2018; Althobaiti, 2022).
Limitations: Building lists of terms that are distinctive of certain dialects is challenging, even when these lists are manually validated, as I later show in §7.3.3 of chapter 7. Moreover, the curated samples are constrained by the diversity of the dialectal terms of these lists. In other words, dialectal sentences that do not have any distinctive terms will be discarded.

(4) Geo-tagging where the text is automatically labeled using information about the location or the nationality of its writer (Mubarak and Darwish, 2014; Salama et al., 2014; Al-Obaidi and Samawi, 2016; Al-Moslmi et al., 2018; Zaghouani and Charfi, 2018; Charfi et al., 2019; El-Haj, 2020; Abdelali et al., 2021; Abdul-Mageed et al., 2020a, 2021b, 2022).
Limitations: While this technique allows for curating data from different Arab countries, it does not consider that speakers of a variety of DA might be living in an Arab country that speaks another variety (e.g., an Egyptian living in Kuwait) (Charfi et al., 2019; Abdul-Mageed et al., 2020a). Moreover, some of the curated sentences might be written in MSA, so the curated sentences need to be split into DA sentences and MSA ones (Abdelali et al., 2021; Abdul-Mageed et al., 2021b, 2022).

Table 5.2: The list of single-labeled ADI datasets categorized by the labeling techniques. I follow the regional categorization of Baimukan et al. (2022). **Ct**/**Cn**/**Re**: the number of cities (provinces), countries, and regions, respectively. *****: The regional dialects are defined as Egypt, Iraq, Levant, Gulf, and Maghreb (Cotterell and Callison-Burch, 2014). †: Sudanese Arabic is considered another regional dialect. **?**: Missing information.

Dataset	Ct/Cn/Re	Description
(1) Manual Labeling		
AOC (Zaidan and Callison-Burch, 2011)	- / - / 5 *	Online comments to news articles, manually labeled three times by crowd-sourced human annotators.
Facebook test set (Huang, 2015) Note: Data is attached to the paper on ACL Anthology.	- / - / 3	2,382 public Facebook posts manually annotated into Egyptian, Levantine, Gulf Arabic, and MSA.
VarDial 2016 (Malmasi et al., 2016) Note: The link provided is not working.	- / - / 4	Sentences sampled from transcripts of broadcast, debate and discussion programs from Al Jazeera. The dialects of these recorded programs were manually labeled. MSA is included as a 5^{th} dialect class for the models. Audio features were used in the 2017 and 2018 editions to allow for building multimodal models.
VarDial 2017 (Zampieri et al., 2017)	- / - / 4	
VarDial 2018 (Zampieri et al., 2018) Note: VarDial 2017 and 2018 used the same data.	- / - / 4	

ArSarcasm-v2 (Abu Farha et al., 2021)	- / - / 4 *	15,548 tweets sampled from previous sentiment analysis datasets, annotated for their dialect (including MSA).
(2) Translation		
Tatoeba (Ho, 2006)	- / 8 / 4	An ever-growing crowdsourced corpus of multilingual translations, that include MSA and eight different Arabic dialects.
MPCA (Bouamor et al., 2014)	- / 5 / 3	2,000 Egyptian Arabic sentences from a pre-existing corpus, manually translated into four other country-level dialects in addition to MSA.
PADIC (Meftouh et al., 2015)	5 / 4 / 2	6,400 sentences sampled from the transcripts of recorded conversations and movie/TV shows in Algerian Arabic and manually translated into four other dialects and MSA.
DIAL2MSA (Mubarak, 2018)	- / - / 4	Dialectal tweets manually translated into MSA.
MADAR6	5 / 5 / 4	10,000 sentences manually translated into five city-level Arabic dialects in addition to MSA.
MADAR26 (Bouamor et al., 2019)	25 / 15 / 5	2,000 sentences manually translated into 25 city-level Arabic dialects in addition to MSA.

(3) Distinctive Lexical Cues		
DART (Alsarsour et al., 2018)	- / - / 5 *	Tweets streamed using a seed list of distinctive dialectal terms, which are used to initially assign a dialect to each tweet, before having them manually verified by crowdsourced annotators.
Twt15DA (Althobaiti, 2022) Note: Data shared as (tweet IDs, labels) only.	- / 15 / 5	Tweets curated by iteratively augmenting lists of distinctive dialectal cues, starting with a seed list for each dialect.
(4) Geo-tagging		
(Mubarak and Darwish, 2014) Note: Not publicly available.	- / ? / ?	Arabic tweets streamed from Twitter, then automatically annotated using the reported user locations of the tweets' authors.
YouDACC (Salama et al., 2014) Note: Not publicly available.	- / 8 / 5 *	Comments to YouTube videos labeled using the videos' countries of origin and the authors' locations.
OMCCA (Al-Obaidi and Samawi, 2016)	5 / 2 / 2	27,912 reviews scrapped from `Jeeran.com`, and automatically labeled using the location of the reviewer.
MASC (Al-Moslmi et al., 2018)	- / 6 / 4	9,141 reviews curated from online reviewing sites, Google Play, Twitter, and Facebook. The country of the reviewer is used as a proxy for the dialect of the review.

Shami (Abu Kwaik et al., 2018)	- / 4 / 1	Sentences in one of the 4 Levantine dialects: (1) manually collected from discussions about public figures on online fora; (2) automatically collected from the Twitter timelines of public figures.
ARAP-Tweet (Zaghouani and Charfi, 2018) Note: No download link on their site.	- / 16 / 5 *	A corpus of tweets from 1100 users, annotated at the user level for the dialect, age, and gender.
ARAP-Tweet 2.0 (Charfi et al., 2019) Note: No download link on their site.	- / 17 / 5 *	A corpus of tweets from about 3000 users, annotated at the user level for the dialect, age, and gender.
Habibi (El-Haj, 2020)	- / 18 / 6 *†	Songs' lyrics labeled by the country of origin of their singers.
QADI (Abdelali et al., 2021) Note: Training data shared as (tweet IDs, labels) only.	- / 18 / 5	Tweets automatically labeled based on the locations of the authors in the user description field. The labels of the testing set of each country were validated by a native speaker of each country's dialect.

Micro-dialects Dataset Abdul-Mageed et al. (2020b)	319 / 21 / 5	More than 277,000 tweets authored by 3,085 users. The geolocated city and country labels of these users were automatically identified and then manually verified.
NADI2020 (Abdul-Mageed et al., 2020a)	100 / 21 / 5	
NADI2021 (Abdul-Mageed et al., 2021b)	100 / 21 / 5	Tweets of users staying in the same province for 10 months, automatically labeled by geotagging the tweets of the selected users.
NADI2022 (Abdul-Mageed et al., 2022)	- / 18 / 5	
NADI2023 (Abdul-Mageed et al., 2023)	- / 18 / 5	
(5) Miscellaneous		
Arabic Dialects Dataset (El-Haj et al., 2018)	- / - / 4 *	12,801 sentences sampled from the AOC dataset, in addition to 3,693 sentences sampled from the *Internet Forums* category of the Tunisian Arabic Corpus (McNeil and Faiza, 010).

5.3 Maximal Accuracy of Single-label ADI Datasets

For a single-label ADI dataset consisting of sentences where each is assigned one dialect label, assume that a percentage $\mathbf{Perc_2}$ of those sentences is valid in two different dialects. For those sentences, only one of the valid dialects is listed as their label. An effective model trained to predict a single label will randomly assign each of these sentences to one of its two respective valid labels. Thus, the expected maximal accuracy for the dataset $\mathbf{E}[\mathbf{Accuracy_{max}}]$ that the model can achieve would then be:

$$\mathbf{E}[\mathbf{Accuracy_{max}}] = (\mathbf{100} - \mathbf{Perc_2}) + \frac{\mathbf{Perc_2}}{\mathbf{2}} \tag{5.1}$$

For example, if 40% of the sentences are valid in two dialects (i.e., $\mathbf{Perc_2} = 40\%$), then the $\mathbf{E}[\mathbf{Accuracy_{max}}]$ of the dataset would be 80%. This becomes worse when a sentence is valid in more than two dialects, to the extent that it can have ten valid dialects in some cases, as shown in Table 5.1. Thus, for a total number of dialects $N_{dialects}$, the equation above can be generalized to:

$$\mathbf{E}[\mathbf{Accuracy_{max}}] = \mathbf{Perc_1} + \sum_{\mathbf{n=2}}^{\mathbf{n=N_{dialects}}} \frac{\mathbf{Perc_n}}{\mathbf{n}} \tag{5.2}$$

where $\mathbf{Perc_1}$ is the percentage of samples that are only valid in one dialect, $\mathbf{Perc_n}$ is the percentage of samples valid in n dialects, $N_{dialects}$ represent the total number of dialects considered, and $\sum_{\mathbf{n=1}}^{\mathbf{n=N_{dialects}}} \mathbf{Perc_n} = \mathbf{100}\%$.

The higher the percentages $\mathbf{Perc_n}$ where $n \in [2, N_{dialects}]$, the lower the maximal accuracy would be. The same pattern would apply to F1 scores. Therefore, a model might be achieving low F1 scores as a consequence of framing DI as a single-label classification task, which might result in high $\mathbf{Perc_n}$ values.

5.4 Estimating the Maximal Accuracy of Datasets

In this section, I will estimate $\mathbf{E}[\mathbf{Accuracy_{max}}]$ for the four country-level datasets, according to Equation 5.2. In order to quantify the percentages $Perc_n$, each sample of a dataset needs to be assessed by native speakers from all the Arab countries. Given the infeasibility of annotating the four datasets, I will estimate the percentages using two methods that provide lower bounds $\widetilde{\mathbf{Perc}}_\mathbf{n}$ of the actual values $\mathbf{Perc_n}$ (i.e., $\widetilde{\mathbf{Perc}}_\mathbf{n} \leq \mathbf{Perc_n}$). Consequently, the estimated maximal accuracy is an upper bound of its true value.

5.4.1 Datasets Derived from Parallel Corpora

These datasets are composed of parallel sentences in different dialects, translated from source sentences in MSA, English, or French. For the four parallel corpora **Multidialectal Parallel Corpus of Arabic (MPCA)** (Bouamor et al., 2014), **PADIC** (Meftouh et al., 2015), **MADAR6**, and **MADAR26** (Bouamor et al., 2018), I transformed the parallel sentences into *(sentence, dialect)* pairs as in subtask (1) of the MADAR shared task (Bouamor et al., 2019). I then mapped the dialect labels for **PADIC**, **MADAR6**, and **MADAR26** from city-level dialects to country-level ones. In case the same sentence is used in different cities within the same country, a single copy is kept. The sentences are then preprocessed by discarding Latin and numeric characters in addition to diacritics and punctuation. Lastly, I estimated the percentages $\widetilde{\mathbf{Perc}}_{\mathbf{n}}$ by computing the percentages of sentences that have the exact same translation in n dialects.

The upper bound for the maximal accuracies of the four corpora lies in the range $[93.9\%, 98.7\%]$ as per Table 5.3. The fact that the maximal accuracy for **MADAR26** is lower than that for **MADAR6** demonstrates that the probability that a sentence is valid in multiple dialects increases as more translations in other country-level dialects are considered.

5.4.2 Datasets of Geolocated Dialectal Sentences

While it is possible to find exact duplicates in datasets of parallel sentences, the probability of having them in datasets of sampled sentences from social media is extremely low. Hence, a different methodology is required to estimate an upper bound of the maximal accuracy that could be achieved for these datasets.

Methodology For the sake of simplicity, let's assume that the samples that were correctly predicted are single-label ones. Moreover, the model's errors can be split into two categories. The first category is the samples for which the model's predictions are not valid. Hence, these samples could also be assumed to have a single label (i.e., the gold standard label). The second category is the samples for which the model's predictions are also valid. Therefore, these samples are valid in at least two dialects (i.e., the gold standard label and the model's prediction).

Dataset	Country-level Dialects	$N_{samples}$	$\sum_{n=2}^{n=N_{dialects}} \widetilde{Perc}_n$	$\widetilde{E}[Accuracy_{max}]$
PADIC	(N=4) Algeria, Palestine, Syria, Tunisia	29,138	5.2%	97.1%
MPCA	(N=5) Egypt, Jordan, Palestine, Syria, Tunisia	4,960	7.8%	95.4%
MADAR6	(N=5) Egypt, Lebanon, Morocco, Qatar, Tunisia	49,476	2.3%	98.7%
MADAR26	(N=15) Algeria, Egypt, Iraq, Jordan, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, Yemen	48,624	9.6%	93.9%

Table 5.3: The estimated percentages and the corresponding expected maximal accuracy for the country-level DI datasets formed using the four parallel corpora. The estimated maximal accuracies are upper bounds of the true maximal accuracies, and we expect the true values to be significantly lower than these estimates. **Note:** The four datasets have city/province-level dialect labels, which were mapped into country-level labels. Sentences in city-level dialects that belong to the same country will be mapped to the same country-level dialect label. Pairs of (sentence, country dialect) are deduplicated.

Model Description The baseline MarBERT-based model of NADI 2022 (Abdul-Mageed et al., 2021a) fine-tuned on the training dataset achieved competitive results on the task's two test sets (macro-averaged F1 scores: 31.39% and 16.94%, accuracies: 47.77% and 34.06%) as compared to the following scores of the best performing model (macro-averaged F1 scores of 36.48% and 18.95%, and accuracies of 53.05% and 36.84%). Consequently, I replicated NADI 2022's baseline, but this time, MarBERT was fine-tuned on the balanced training dataset of NADI 2023.

Dataset Decription I used the QADI dataset (Abdelali et al., 2021) as my test set. QADI's test set covers the same 18 countries as NADI 2023. I decided to analyze the errors of my model on QADI for two reasons: 1) the test sets of the NADI shared tasks are not not publicly released; 2) the dialect labels of the samples of QADI's test set were automatically assigned using geolocations similar to NADI, but the label of each sample was validated by a native speaker of the sample's label, which gives additional quality assurance for QADI over NADI. The model achieves an accuracy of **50.74%** on QADI's test set with the full classification report in Table 5.4. Additionally, Figure 5.2 visualizes how the predictions and labels are confused together.

Manual Error Analysis I recruited annotators to validate the False Positives (FPs) the model makes for seven different dialects. A dialect's FPs are samples for which the model predicts the considered dialect, while the gold standard is another dialect. The annotators are native-speakers of Algerian, Egyptian, Palestinian, Lebanese, Saudi Arabian, Sudanese, and Syrian Arabic, respectively. Each annotator is shown the FPs for their native dialect, one at a time, and is asked to validate them.[1] To this end, an online Qualtrics survey was created. Screenshots of the used survey are shown in §D of the Appendix.

If the annotator found the FP sample to be valid in their native dialect, it means that this sample is valid in at least two different Arabic dialects (i.e., the sample's original label, and the model's prediction).[2] However, it can still be valid in additional dialects, which I did not check for due to the limited number of participants.

[1] I release the judgments through: `https://github.com/AMR-KELEG/ADI-under-scrutiny/tree/master/data`

[2] Annotators are given a third choice *Maybe / Not Sure*, which I count as *No* (i.e., invalid in their dialect).

Dialect	Support	Precision (P)	Recall (R)	F1-score (F1)
Algeria	170	0.63	0.42	0.51
Libya	169	0.45	0.73	0.56
Morocco	178	0.77	0.63	0.70
Tunisia	154	0.63	0.54	0.58
Bahrain	184	0.33	0.29	0.31
Iraq	178	0.69	0.62	0.65
Kuwait	190	0.38	0.43	0.40
Oman	169	0.46	0.51	0.49
Qatar	198	0.37	0.34	0.35
Saudi Arabia	199	0.40	0.44	0.42
UAE	192	0.37	0.53	0.43
Egypt	200	0.65	0.85	0.73
Sudan	188	0.91	0.68	0.78
Jordan	180	0.31	0.47	0.38
Lebanon	194	0.63	0.69	0.66
Palestine	173	0.47	0.43	0.45
Syria	194	0.56	0.31	0.40
Yemen	193	0.55	0.25	0.34
Macro avg.		0.5309	0.5085	0.5072
Weighted avg.		0.5295	0.5074	0.5058
Accuracy		0.5074		

Table 5.4: The evaluation metrics for the predictions of the fine-tuned MarBERT model on QADI's testing set. The model is fine-tuned on NADI 2023's training data.

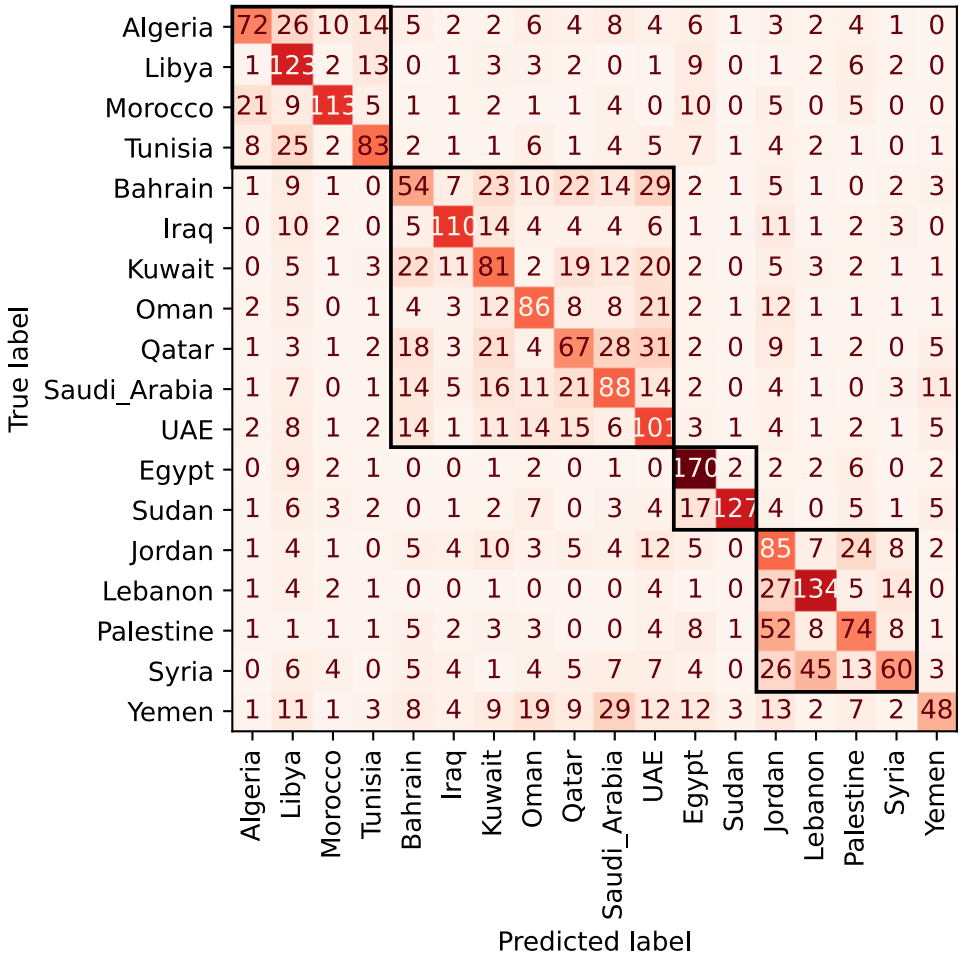


Figure 5.2: The confusion matrix for the predictions of a MarBERT model on QADI's test set. The model was fine-tuned using NADI 2023's training dataset. The black bounding boxes designate parts of the confusion matrix for countries within the same macro-region. This indicates that the model often confuses the dialects of countries within the same region. Yet, a non-negligible proportion of the errors exists outside these boxes, indicating confusion between country-level dialects from different regions.

Validity of the Model's FPs Out of 490 validated FPs, 325 were found to be also valid in the other dialect to which they were classified, which represents ≈ 66% of the validated errors. Having such a great proportion of FPs that are not true errors hinders the ability to properly analyze and improve the ADI models. For Egyptian, Palestinian, Saudi Arabian, and Syrian Arabic, the majority of the FPs are incorrect, as demonstrated in Figure 5.3 (i.e., the model's prediction should be considered to be correct). As expected, dialects grouped in the same region are similar, and thus the FPs of a dialect would generally have labels of other dialects from the same region, as in Figure 5.4.

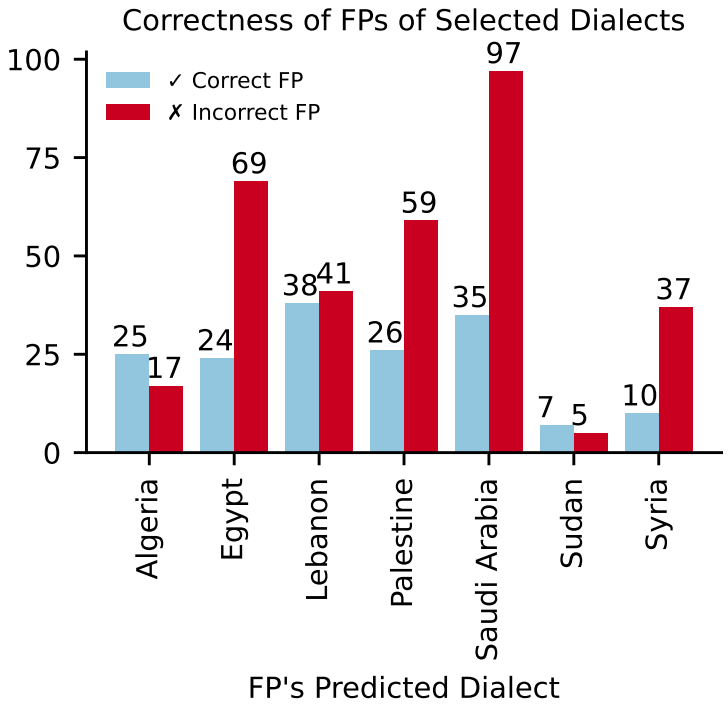


Figure 5.3: The distribution of the annotations for the validity of the False Positives (FPs) in 7 Arabic dialects. **Correct FP** represents the FP samples for which the model's prediction is invalid. **Incorrect FP** the FP samples for which the model's prediction is valid.

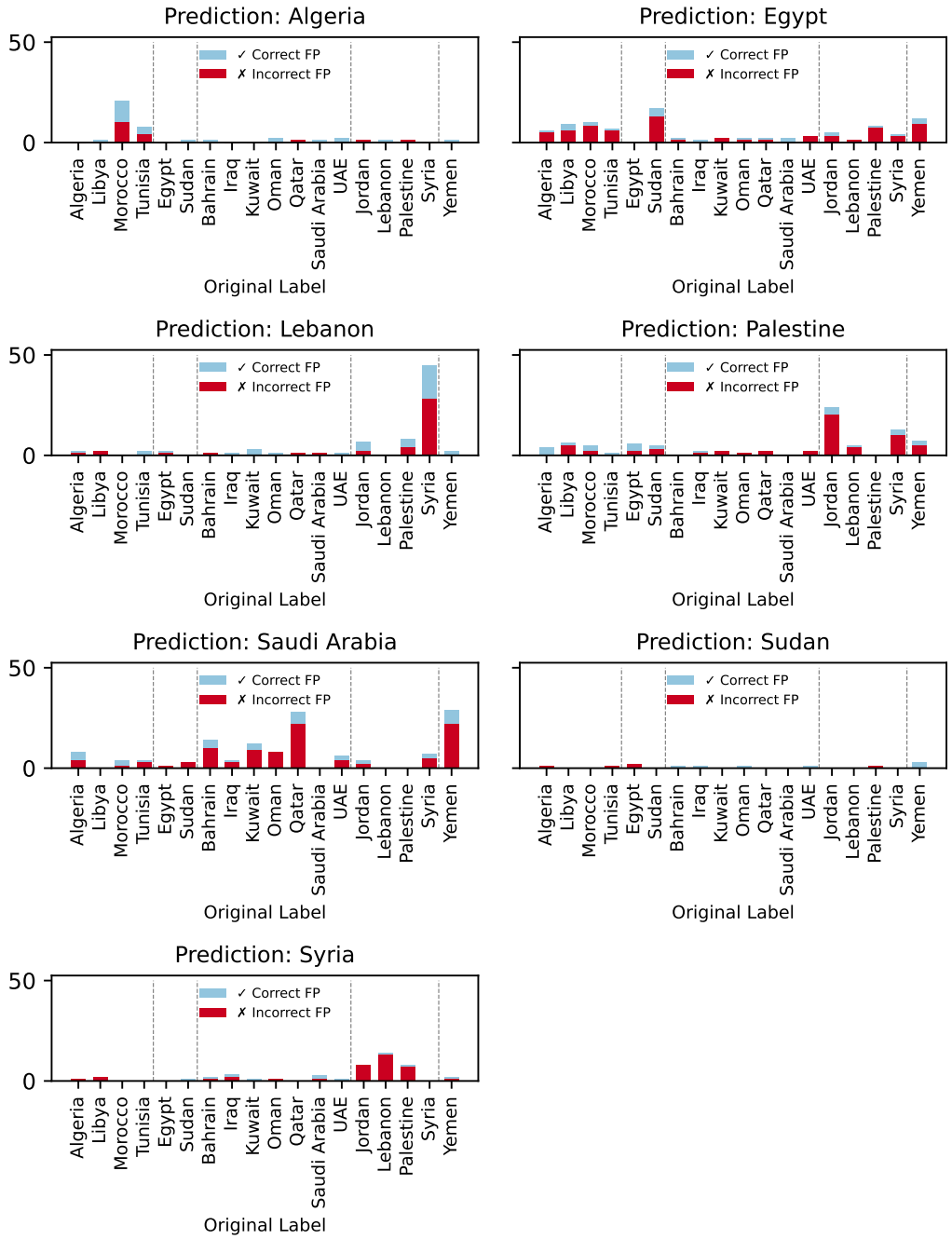


Figure 5.4: The distribution of the original labels for the False Positives (FPs) of the seven validated dialects. **Correct FP** represents the FP samples for which the model's prediction is invalid. **Incorrect FP** represents the FP samples for which the model's prediction is valid.

Table 5.5 lists some examples of samples of the QADI dataset for which the model's predictions do not match the original labels, yet the annotators found these predictions to also be valid.

Model's Prediction	Sentence	Original Label
Algeria	عيشك يبارك فيك و يخليك	Tunisia
	الله يرحمه ربي معك خويا و انا لله و انا اليه راجعون	Morocco
Egypt	يلعن الكورة واليوم اللي شجعت في كورة .	Palestine
	مرتضي صوتوا ضعيف مع كامل إحترامي مايتقارنش بنسيم مجرد مقارنة	Tunisia
Lebanon	حالتنا أهون من حالات كتير في الحاضر و في التاريخ .. و غيرنا كتير نجحوا .	Egypt
	هههههه مين قلك أعصابي تعبانة	Syria
Palestine	بما أنو آخر شهر يا ربي يكونو عاملين خصم عالفلافل	Lebanon
	المشكلة انه فيه ناس ماعندهم عقل عشان تعطيهم على قد عقلهم	Kuwait
Saudi Arabia	والله ماعرف عنه بس جتني الصوره على الخاص وقلت اكيد تذكرونه	Iraq
	اقرا تغريدتي بالكامل وتقرا تغريدة كساب العتيبي وتعال اسال عنها وراح اجيبك	Qatar
Sudan	هههههههه انت رجعتي في كلامك سمحتي سمحتي	Tunisia
	والله يا استاذ عوض دي عربيه	Egypt
Syria	هلق الاستعمار فرض علينا بس الاستحمار نحنا فينا نعمله او ما نعمله	Lebanon
	لاابدا ناس عندهم مبدا	Iraq

Table 5.5: Samples of QADI for which the ADI model's predictions are also valid.

Estimated Upper bound of QADI's Maximal Accuracy If we only consider the 725 samples that were correctly predicted by the model (TPs) in addition to the validated 490 FPs, then we know that 325 samples out of 1215 are at least valid in two different dialects. The $\tilde{\mathbf{Perc}}_{\mathbf{2}}$ for this subset is 26.7%, making the maximal accuracy $\mathbf{E}[\mathbf{Accuracy}_{\mathbf{max}}]$ equal to 86.6%.

Impact on Evaluation Metrics To further investigate the impact of the incorrect FPs on the evaluation metrics, I computed the corrected True Positive value for each dialect $\mathbf{TP}^*$ as $\mathbf{TP}^* = \mathbf{TP} + \mathbf{Incorrect\ FP}$. Using these corrected $\mathbf{TP}^*$ values, I computed corrected precision, recall, and F1-scores. As per Table 5.6, the macro-averaged F1-score increased from 0.56 to 0.72. This clearly confirms my hypothesis that modeling the ADI task as a single-label classification task leads to inaccurate evaluation of the systems.

Dialect	**TP**	**FP**	**TP***	**FP***	**FN**	**P**	**R**	**F1**	**P***	**R***	**F1***
Algeria	72	42	72 + 17 = 89	25	98	.63	.42	.51	.78	.48	.59
Egypt	170	93	170 + 69 = 239	24	30	.65	.85	.73	.91	.89	.90
Lebanon	134	79	134 + 41 = 175	38	60	.63	.69	.66	.82	.74	.78
Palestine	74	85	74 + 59 = 133	26	99	.47	.43	.45	.84	.57	.68
Saudi Arabia	88	132	88 + 97 = 185	35	111	.40	.44	.42	.84	.62	.72
Sudan	127	12	127 + 5 = 132	7	61	.91	.68	.78	.95	.68	.80
Syria	60	47	60 + 37 = 97	10	134	.56	.31	.40	.91	.42	.57
Macro-average						.61	.55	.56	.86	.63	.72

Table 5.6: The impact of the incorrect FPs on the precision $\mathbf{P}$, recall $\mathbf{R}$, and F1-score $\mathbf{F1}$. Error samples for a specific predicted dialect (i.e., FPs of this dialect) that are labeled as valid in this predicted dialect are counted as true positives in the corrected $\mathbf{TP}^*$ score. The corrected $\mathbf{P}^*$, $\mathbf{R}^*$ and $\mathbf{F1}^*$ are based on the corrected value of $\mathbf{TP}^*$.
$\mathbf{P}^* = \frac{\mathbf{TP}^*}{\mathbf{TP}^* + \mathbf{FP}^*}$, $\mathbf{R}^* = \frac{\mathbf{TP}^*}{\mathbf{TP}^* + \mathbf{FN}}$, $\mathbf{F1}^* = \frac{2 * \mathbf{P}^* * \mathbf{R}^*}{\mathbf{P}^* + \mathbf{R}^*}$
Note: P stands for Precision, R stands for Recall, and $F1$ stands for F1-score.

5.5 Proposal for Framing the ADI Task

The previous sections demonstrated the limitations of single-label framing of ADI. More specifically, it can not model the many sentences that are valid in multiple dialects. Two labeling schemes have been previously used for the task of DI to better model these sentences.

(1) Using a *General* label Zaidan and Callison-Burch (2014) asked crowdsourced annotators to label dialectal sentences as being *Egyptian*, *Gulf*, *Iraqi*, *Levantine*, *Maghrebi*, *Other dialect*, or *General dialect*. They used the *General dialect* for sentences that can be valid in multiple dialects. However, the *General dialect* is underspecified, and it is not clear whether it should be used for sentences valid in more than one dialect or for sentences valid in all of the considered dialects.

Unsurprisingly, the authors noted that some annotators barely used the label, while others used it when they were not sure about the dialect of the underlying sentences. Moreover, they noticed that the annotators tended to over-identify their native dialects. Since annotators might not realize that a sentence valid in their native dialect is also valid in other dialects, they could end up choosing their native dialect as the label for this sentence, instead of the *General dialect* label.

2) Using a *Both or Neither* label Zampieri et al. (2024) focused on the binary distinction between two varieties of English, Portuguese, and Spanish. In addition to the two varieties of each language, the annotators are allowed to assign sentences to a third label *Both or Neither*. The evaluation results indicate that the *Both or Neither* label is harder to model computationally than the other variety labels. Moreover, it is not clear how this label could generalize to more than two dialects. The dataset creators acknowledged some limitations of using this third label, calling for modeling improvements.

The discussion above indicates that the *General* and *Both or Neither* labels do not adequately model the sentences valid in multiple dialects. Framing ADI as a multi-label classification task would potentially alleviate all the aforementioned limitations.

5.5.1 ADI as Multi-label Classification

Multi-label classification allows assigning one or more dialects to the same sample. This requires multiple design changes as explained below.

Labeling: Collecting multi-labels for a dataset requires the manual annotation of its samples. Dataset creators need to consider how they collect the annotations, and consequently, who to recruit. An Arabic speaker of a specific dialect would be able to determine if a sentence is valid in their dialect or not (Salama et al., 2014; Abdelali et al., 2021). Althobaiti (2022) found that the average inter-annotator agreement score (Cohen's Kappa) is 0.64, where two native speakers of 15 different country-level Arabic dialects are asked to check the validity of tweets in their native dialects.

While human participants can sometimes infer the macro-dialect of a sentence that is not in their native dialect, it seems quite hard for them to predict the country-level dialects in which the sentence is valid (Abdul-Mageed et al., 2020b).
Recommendation: Ask Arabic speakers to identify if a sentence is valid in their native dialects or not as per (Salama et al., 2014; Abdelali et al., 2021; Althobaiti, 2022). In order to include new dialects, speakers of these dialects need to be recruited.

Modeling: One way of building multi-label classification models is to use multiple binary classifiers. More specifically, a binary classifier is built to decide whether a sentence is valid in one dialect or not. For N dialects, N binary classifiers would be responsible for predicting the labels of a single sample.

Evaluation: For each supported dialect, evaluation metrics like accuracy, precision, recall, and F1-score can be used. Macro-averaging the metrics is a way to measure the average performance of the model across the different dialects.

Extensibility: The multi-label framing is extensible since more labels can be added to a previously annotated dataset. Adding a new dialect class does not invalidate the labels of the other dialect classes. This does not apply to the single-label framing since an annotator would need to select a dialect out of a predefined set of dialects. Changing the set of dialects would require the reannotation of the whole dataset.

5.6 Summary

Single-label classification has been the de facto framing for Arabic Dialect Identification (ADI). In this chapter, I showed that an upper bound for the expected maximal accuracy that an oracle ADI model can achieve is as low as 93.9%. After recruiting native speakers of 7 different Arabic dialects, I found that 66% of a state-of-the-art ADI model's errors are also valid predictions. This hints that a non-negligible portion of the Arabic sentences is valid in multiple dialects, which is the first step to answer (***RQ4***) ***How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?*** Hence, framing ADI as a single-label classification task is a major limitation, which could be the main reason behind the low performance of the current ADI models.

Based on the findings above, I argue that ADI should be framed as a multi-label classification task. To this end, I provided a proposal with detailed design recommendations for various aspects, including data curation, annotation, and model evaluation. The following chapter describes how this proposal was put into practice, in collaboration with the organization team of the NADI 2024 shared task.

Chapter 6

Redesigning Arabic Dialect Identification (NADI 2024)

I concluded the previous chapter by proposing to model ADI as a multi-label classification task. This proposal was based on analyzing the errors of a single-label state-of-the-art ADI model, finding that $\approx 66\%$ of the model's predictions for these samples are also valid. In this chapter, I put this proposal into practice to build the first multi-label ADI dataset. This was done in collaboration with the organization team of the *Fifth Nuanced Arabic Dialect Identification Shared Task (NADI 2024)*.

We sampled 80 tweets—75 in DA and 5 in MSA—from the 14 most populated Arab countries (excluding Somalia, for which available data was not sufficient), for a total of 1,120 tweets. Instead of only annotating the errors of an ADI model, the whole dataset is labeled by speakers of 9 different country-level dialects. We also asked the annotators to label the level of dialectness of the tweets that they deemed valid in their country-level dialect, according to new annotation guidelines.

I was the main member responsible for designing the different aspects and monitoring the annotation of this dataset, which served as the evaluation set for the shared task's first two subtasks: (1) multi-label ADI, and (2) ALDi estimation. I also used the newly created dataset to further investigate the following questions: (***RQ1***) ***How can the concept of Dialect Levels be operationalized in a way that can be effectively estimated?*** and (***RQ4***) ***How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?***

Previous runs of the NADI shared tasks played a vital role in advancing our understanding of the country-level ADI task. Throughout the yearly runs of the shared task between 2020 and 2023, the NADI organizers have been improving the quality of the

training data used for the task as summarized in Figure 2.8. For the fifth run of the NADI shared task, and given the limitations of the single-label framing of the task listed in chapter 5, we (I, in collaboration with the NADI 2024 organizers from CMU Qatar, Edinburgh, MBZUAI, NYU-AD, and UBC) decided to frame the ADI subtask as a multi-label classification one. Additionally, ALDi estimation was introduced as another subtask for the first time.

The work presented herein was reported in the following paper:

- Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. *NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task.* In Proceedings of the Second Arabic Natural Language Processing Conference, pages 709–728, Bangkok, Thailand.

6.1 Dataset Creation

6.1.1 Samples Curation

In order to curate dialectal data with more accurate geolocations, Abdul-Mageed et al. (2020a) collected tweets from 21 Arab countries for 10 months (between January 2010 and October 2019). They then identified the users who had exclusively tweeted from a single province for the whole 10-month period. Lastly, they selected samples of these users' tweets to form the training and evaluation datasets of the NADI 2020 dataset, with the country-level dialect labels of the samples automatically set by mapping the geolocated province/city labels to their respective country labels. Different samples from the same collection of tweets were used in the subsequent runs of the NADI shared task (Abdul-Mageed et al., 2021b, 2022, 2023).

For the NADI 2024 evaluation dataset, UBC members randomly sampled 80 data points from the 14 most populous Arab countries (excluding Somalia, for which sufficient data was not available), using the same source collection of tweets used in the previous NADI datasets. I managed to recruit annotators from the following 10 countries: *Algeria, Egypt, Iraq, Jordan, Morocco, Palestine, Sudan, Syria, Tunisia,* and *Yemen*. However, one of the Jordanian annotators was unable to complete the annotation process in time, so the test set samples have labels for only nine country-level dialects. The remaining four countries—represented in the dataset—from which I could not recruit annotators at the time of organizing the shared task are: *Lebanon, Libya, Saudi*

Arabia, UAE.[1] A significant proportion of the samples geolocated to these four countries are intuitively expected to be labeled as invalid in the dialects of the ten countries from which we recruited the annotators. Yet, including them ensures the dataset's samples cover a wider range of dialects, allowing for the evaluation of the models' ability to correctly identify that most of these samples are not valid in the dialects for which we have validity labels. The UBC members utilized their in-house MSA/DA classification model (acc=89.1%, F_1 score=88.6) introduced in Abdul-Mageed et al. (2021a) to ensure that for each country's 80 geolocated samples, five are in MSA, and 75 are in DA. The overall dataset size for the shared task is 1,120 samples. Each annotator labeled the whole dataset.

The UBC members applied a preprocessing step that normalizes user mentions, URLs, emojis, and numbers, replacing them with the following respective special tokens: *USER, URL, EMOJI, NUM*. I removed these special tokens from the data to prevent confusion for the annotators, but retained the hashtags before labeling the samples. In hindsight, I would rather not normalize the numbers and the emojis, keeping them as they are represented in the original tweet, as I realized that specifically removing the *NUM* tokens made some of the sentences ungrammatical. I recruited the annotators on Upwork, and the annotation process incurred a total cost of $\$1,700$.

6.1.2 Annotation Guidelines

This subsection explains the rationale behind the annotation guidelines we used for the *Validity* and *ALDi* labels of our dataset.

(1) Validity Labels There have been multiple attempts asking annotators to check if sentences are written in their native dialect of Arabic. This was done mainly to validate the dialect labels that are automatically assigned using geolocating methods/distinctive dialectal cues (Alsarsour et al., 2018; Abdelali et al., 2021; Althobaiti, 2022) or to perform error analysis for the predictions of DI systems (chapter 5).

Arabic speakers could have different perceptions of their country-level dialects, depending on their background and exposure to the different speaking communities in their countries. Such differences could impact their understanding of the validity of sentences in their country-level dialects. Previous wordings (shown in Table 6.1) did not instruct the annotators to consider the different dialects spoken in their countries.

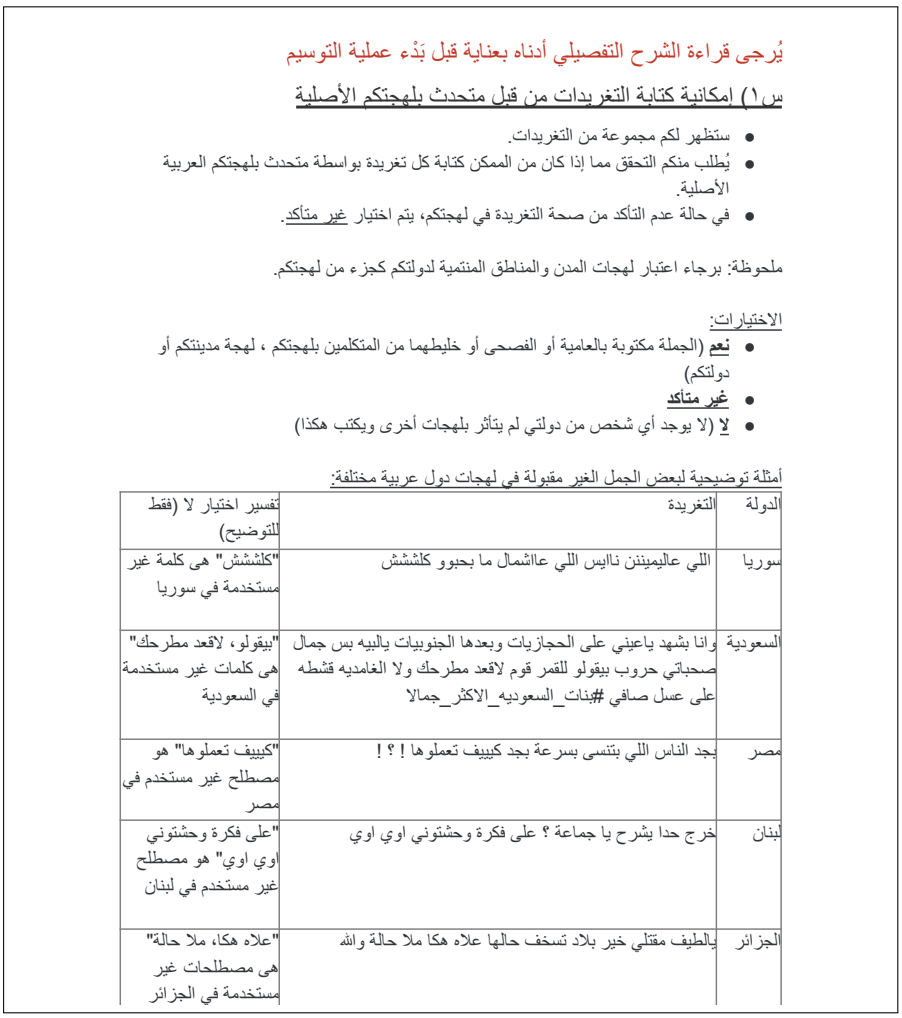
[1] I will introduce *MLADI* in chapter 7, an extended version of the dataset with completed annotations from Jordan, and new annotations from Saudi Arabia.

Wording
(Alsarsour et al., 2018) Asked annotators to label each tweet as either in: (their native dialect, MSA, or other).
(Abdelali et al., 2021) *Is this tweet consistent with the dialect spoken in your country?* (Yes, No).
(Althobaiti, 2022) *Is this sentence written in* `Dialect` *dialect?* (Yes, No); `Dialect` is the demonymic form of the annotator's country (e.g., Egyptian, Moroccan).
Wording used in the error analysis experiment (chapter 5) *Is this sentence valid in your dialect?* (Yes, Not Sure, No).
Current wording used in NADI 2024 ***Is it possible that the tweet is authored by someone who speaks one of your country's dialects?*** (Yes, Not Sure/Maybe, No)

Table 6.1: The different wordings used for checking the validity of sentences in an Arabic dialect.

Current Wording For our dataset, we tried to overcome these limitations by explicitly asking the annotators to consider the different dialects spoken in their respective countries: **Q1)** *Is it possible that the tweet is authored by someone who speaks one of your country's dialects?* **Options:** (a) *Yes*, (b) *Not Sure/Maybe*, or (c) *No*.

Lastly, we provided some examples of tweets not valid in the country-level dialects of *Syria, Saudi Arabia, Egypt, Lebanon, and Algeria*, highlighting the spans of the tweets that made them invalid in each respective dialect, as demonstrated in Figure 6.1.



يُرجى قراءة الشرح التفصيلي أدناه بعناية قبل بَدْء عملية التوسيم

س١) إمكانية كتابة التغريدات من قبل متحدث بلهجتكم الأصلية

- ستظهر لكم مجموعة من التغريدات.
- يُطلب منكم التحقق مما إذا كان من الممكن كتابة كل تغريدة بواسطة متحدث بلهجتكم العربية الأصلية.
- في حالة عدم التأكد من صحة التغريدة في لهجتكم، يتم اختيار غير متأكد.

ملحوظة: برجاء اعتبار لهجات المدن والمناطق المنتمية لدولتكم كجزء من لهجتكم.

الاختيارات:

- **نعم** (الجملة مكتوبة بالعامية أو الفصحى أو خليطهما من المتكلمين بلهجتكم ، لهجة مدينتكم أو دولتكم)
- **غير متأكد**
- **لا** (لا يوجد أي شخص من دولتي لم يتأثر بلهجات أخرى ويكتب هكذا)

أمثلة توضيحية لبعض الجمل الغير مقبولة في لهجات دول عربية مختلفة:

الدولة	التغريدة	تفسير اختيار لا (فقط للتوضيح)
سوريا	اللي عاليمينن نايس اللي عااشمال ما بحبوو كلششش	"كلششش" هى كلمة غير مستخدمة في سوريا
السعودية	وانا بشهد ياعيني على الحجازيات وبعدها الجنوبيات يالبيه بس جمال صحباتي حروب بيقولو للقمر قوم لاقعد مطرحك ولا الغامديه قشطه على عسل صافي #بنات_السعوديه_الاكثر_جمالا	"بيقولو، لاقعد مطرحك" هى كلمات غير مستخدمة في السعودية
مصر	بجد الناس اللي بتنسى بسرعة بجد كيييف تعملوها ! ؟ !	"كيييف تعملوها" هو مصطلح غير مستخدم في مصر
لبنان	خرج حدا يشرح يا جماعة ؟ على فكرة وحشتوني اوي اوي	"على فكرة وحشتوني اوي اوي" هو مصطلح غير مستخدم في لبنان
الجزائر	يالطيف مقتلي خير بلاد تسخف حالها علاه هكا ملا حالة والله	"علاه هكا، ملا حالة" هى مصطلحات غير مستخدمة في الجزائر

Figure 6.1: The guidelines used for annotating the validity of the sentences. Refer to Figure E5 in the Appendix for English translation.

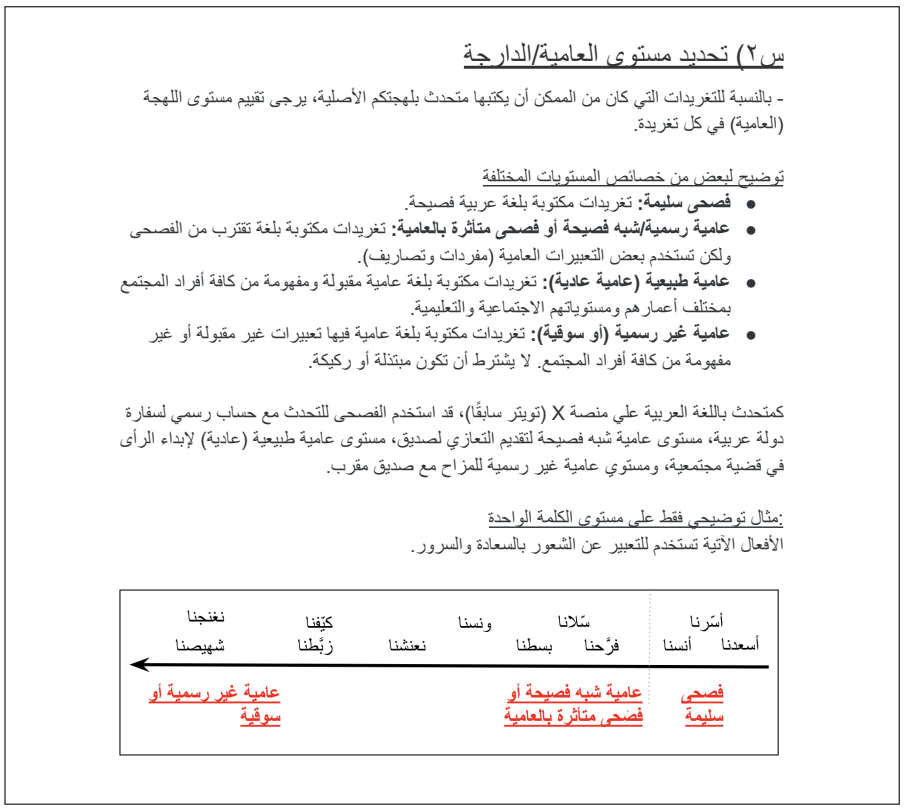
(2) ALDi Labels We follow Zaidan and Callison-Burch (2011)'s setup in which they asked the annotators to assign a discrete ALDi level to each sentence, along a four-level ordinal scale (Refer to §3.2 for further details). For their guidelines, I noticed multiple limitations. First, Level 3 (*Mostly Dialectal* معظمها عامية) could not fully separate between sentences having a word perceived as highly colloquial, and sentences having a majority of dialectal words that are not perceived as highly colloquial on the word level. Moreover, they provided minimal guidelines to the annotators, which might have impacted the interannotator agreement.

Current Guidelines We decided to use the same operationalization of ALDi, while providing more elaborate guidelines to reduce the variability of the assigned ALDi levels for the same sentences. To this end, we provided descriptive labels for the four levels and short descriptions of the sentences expected to belong to each label. Moreover, we included two examples to further explain the concept of ALDi, one on the word level and another on the sentence level, as per Figure 6.2. This would potentially allow for better separation between the different ALDi levels, hence higher IAA scores.

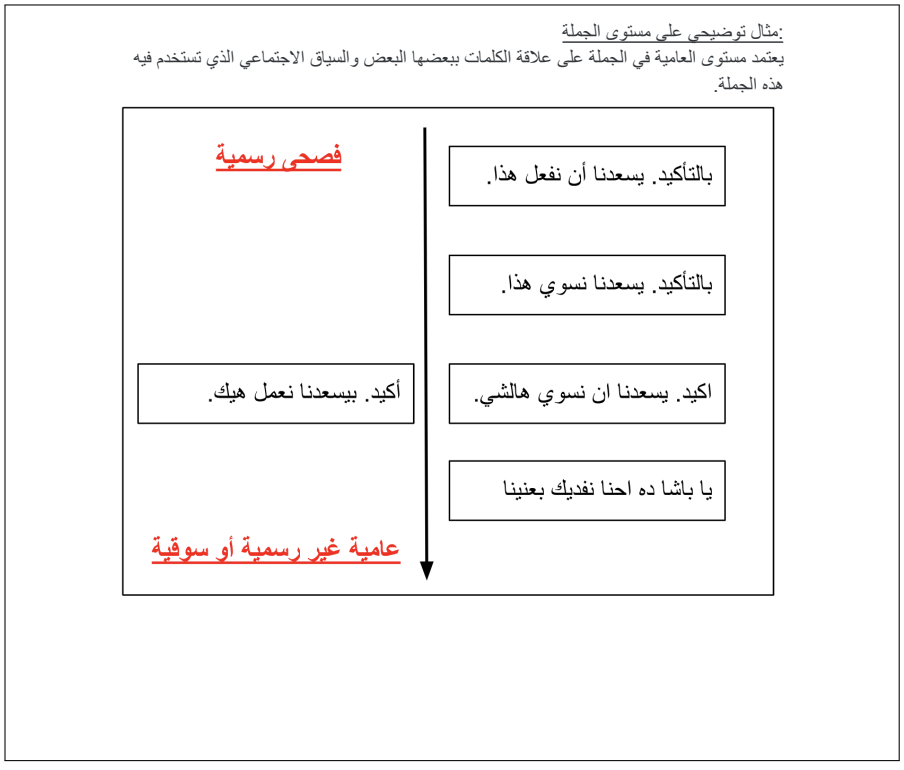
Q2) *What is the Arabic Level of Dialectness (ALDi) of the tweet?* We define the following levels:

0. **Sound MSA**: Tweets written in fluent MSA.
1. **Formal Colloquial or Colloquial-influenced MSA**: Tweets written in a language close to MSA but using some colloquial expressions (lexemes/ morphemes).
2. **Natural/Ordinary Colloquial**: Tweets written in a colloquial language that is accepted and understood by all members of society, of all ages and social/educational levels.
3. **Informal (or Vulgar) Colloquial**: Tweets written in a colloquial language having expressions that are not accepted or understood by all members of society. It does not have to be vulgar or weak.

ALDi Annotation Condition I thought that speakers of a dialect might perceive sentences in other dialects to be highly colloquial, especially if the sentences are not fully intelligible to them. For this reason, an annotator was asked to assign an ALDi rating only if their answer to the first question (validity of a tweet in one of their country-level dialects) is either *Yes* or *Not Sure/Maybe*.



(a) The guidelines for the ALDi Estimation subtask. Refer to Figure E6 in the Appendix for English translation.



(b) An example of different-ALDi variants of a sentence. Refer to Figure E7 in the Appendix for English translation.

Figure 6.2: Screenshots of the guidelines that were provided to the annotators for the ALDi level estimation.

6.1.3 Annotation Process

As previously mentioned, I used Upwork to recruit three native speakers from each of the ten pre-specified countries to annotate the selected 1,120 tweets. For this dataset, I was unable to ensure that each country's annotators represented different provinces or cities, due to the difficulty of doing so on Upwork. Based on the city/province locations of the annotators' native dialects in Figure 6.3, we can see that some countries' annotators are geographically distributed (e.g., Morocco and Algeria), while others share the same city/province dialect (e.g., Egypt).

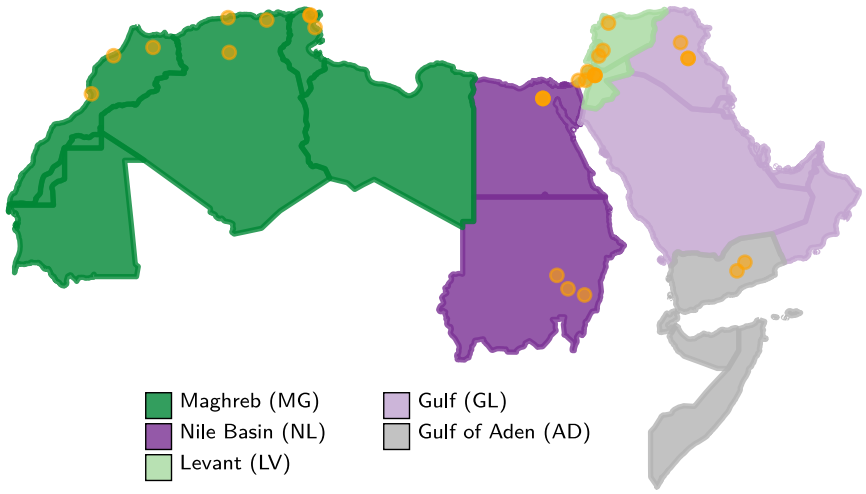


Figure 6.3: The city/province locations of the annotators' native dialects, for the 10 considered country-level dialects. The regional dialects (Maghreb, Nile Basin, Levant, Gulf, and Gulf of Aden) are encoded in different colors according to the groupings presented in Baimukan et al. (2022).

Before inviting the annotators to the main task, I asked them to complete an on-boarding task to get them acquainted with our objectives and clarify any potential misunderstandings. Afterward, the main task annotation process was split into five batches, each containing 224 samples, where feedback was provided to the annotators after each batch to ensure high annotation quality. Annotators were paid $8 after successfully completing each of the six tasks, in addition to a bonus value between $8 and $12 after completing the whole process. After accounting for the platform fees, annotating the dataset cost about $1,700. I was the primary member responsible for planning the dataset creation, setting up the annotation platform, and generating batches. The annotation process quality assurance and monitoring were done in collaboration with Injy Hamed, another member of the shared task organization team.

Onboarding Tasks QADI's test set (Abdelali et al., 2021) has 3,303 tweets geolocated to 18 different Arab countries (including the 14 countries represented in the samples of our dataset). The geolocated label for each tweet was then validated by a native speaker from that country, who checked if the "tweet is consistent with the dialect spoken in their country". Additionally, the test set has 200 tweets automatically classified as written in MSA.

In order to get the annotators acquainted with our annotation guidelines shown in Figure 6.1 and Figure 6.2, I asked them to label 35 tweets from QADI as an initial onboarding task. Each country's onboarding task had 10 tweets labeled as consistent with the dialect(s) spoken in the country according to QADI's annotations and 5 MSA samples.[2] I included MSA samples to ensure that the onboarding tasks contain tweets of potentially different levels of dialectness. Additionally, I had 2 DA samples from 10 other country-level dialects, which would act as negatives for the first subtask (i.e., some of these samples are expected not to be valid in the considered country of the onboarding task). For each country's onboarding task, the composing samples were randomly shuffled. The annotators were not provided with information about the samples' geolocated labels or their distribution across the different country labels.

Quality Assurance and Feedback For each country's onboarding task, and thanks to the labels from QADI, we could perform two automatic checks for assessing the quality of the annotations:

- **Check (1)** The 10 samples geolocated to an annotator's country are expected to (a) be labeled as valid, and (b) have an ALDi level > 0.
- **Check (2)** The 5 MSA samples are expected to be marked as valid by all the annotators, with Level (0) *Sound MSA* as their ALDi.

[2] I noticed that some of QADI's non-MSA samples were classified as being in MSA by the UBC team's closed-source MSA/DA model, so I relied on the model's predictions for considering a sample as MSA.

Annotator #	Sample Geolocated Label														
	Algeria (y/m/n)	**Egypt** (y/m/n)	**Iraq** (y/m/n)	**Jordan** (y/m/n)	**Lebanon** (y/m/n)	**Libya** (y/m/n)	**Morocco** (y/m/n)	**Palestine** (y/m/n)	**Saudi Arabia** (y/m/n)	**Sudan** (y/m/n)	**Syria** (y/m/n)	**Tunisia** (y/m/n)	**UAE** (y/m/n)	**Yemen** (y/m/n)	**MSA** (y/m/n)
Algeria (A)	**6**/0/4 *	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	1/1/**0**	-	-	**5**/0/0
Algeria (B)	**7**/0/3 *	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	2/0/**0**	-	-	**5**/0/0
Algeria (C)	**5**/0/5 *	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	-	**0**/0/5 *
Egypt (A)	0/0/**2**	**8**/0/2	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Egypt (B)	0/1/**1**	**10**/0/0	1/0/**1**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Egypt (C)	0/0/**2**	**10**/0/0	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/1/**1**	0/0/**2**	1/0/**1**	0/0/**2**	0/0/**2**	-	-	**4**/0/1
Iraq (A)	1/1/**0**	1/0/**1**	**6**/1/3	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Iraq (B)	0/2/**0**	0/1/**1**	**7**/3/0	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	-	**3**/2/0
Iraq (C)	2/0/**0**	1/0/**1**	**10**/0/0	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/1/**1**	-	-	**5**/0/0
Morocco (A)	2/0/**0**	2/0/**0**	1/0/**1**	2/0/**0**	2/0/**0**	-	**10**/0/0	1/1/**0**	1/1/**0**	2/0/**0**	2/0/**0**	2/0/**0**	-	-	**5**/0/0
Morocco (B)	2/0/**0**	2/0/**0**	2/0/**0**	2/0/**0**	2/0/**0**	-	**10**/0/0	2/0/**0**	2/0/**0**	1/1/**0**	2/0/**0**	2/0/**0**	-	-	**5**/0/0
Morocco (C)	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	**8**/0/2	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	-	**4**/0/1
Palestine (A)	1/0/**1**	1/0/**1**	0/0/**2**	1/0/**1**	1/0/**1**	-	0/0/**2**	**8**/0/2	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	-	**4**/0/1
Palestine (B)	2/0/**0**	2/0/**0**	1/0/**1**	2/0/**0**	2/0/**0**	-	1/0/**1**	**7**/0/3	0/0/**2**	0/0/**2**	2/0/**0**	1/0/**1**	-	-	**4**/0/1
Palestine (C)	1/0/**1**	1/1/**0**	0/0/**2**	1/0/**1**	2/0/**0**	-	0/0/**2**	**9**/1/0	0/0/**2**	1/0/**1**	1/0/**1**	1/0/**1**	-	-	**5**/0/0
Sudan (A)	0/1/**1**	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	0/0/**2**	0/0/**2**	-	-	**4**/0/1
Sudan (B)	2/0/**0**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	**9**/0/1	0/0/**2**	0/0/**2**	-	-	**5**/0/0
Sudan (C)	2/0/**0**	0/0/**2**	0/1/**1**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	1/0/**1**	1/0/**1**	-	-	**5**/0/0
Syria (A)	0/0/**2**	1/0/**1**	0/0/**2**	1/0/**1**	2/0/**0**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/0/1	0/0/**2**	-	-	**5**/0/0
Syria (B)	2/0/**0**	1/0/**1**	1/0/**1**	1/0/**1**	1/1/**0**	-	0/1/**1**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	1/1/**0**	-	-	**5**/0/0
Syria (C)	0/1/**1**	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/1/0	0/0/**2**	-	-	**1**/4/0 *
Tunisia (A)	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	**9**/0/1	-	-	**2**/0/3 *
Tunisia (B)	1/0/**1**	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	**9**/0/1	-	-	**5**/0/0
Tunisia (C)	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	1/0/**1**	**8**/1/1	-	-	**4**/1/0
Yemen (A)	1/1/**0**	0/0/**2**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/1/**1**	0/0/**2**	0/0/**2**	1/0/**1**	1/0/**1**	-	**8**/1/1	**5**/0/0
Yemen (B)	2/0/**0**	1/0/**1**	0/0/**2**	0/0/**2**	0/0/**2**	-	0/0/**2**	0/0/**2**	1/0/**1**	0/0/**2**	1/0/**1**	1/0/**1**	-	**10**/0/0	**4**/0/1
Yemen (C)	2/0/**0**	1/0/**1**	0/0/**2**	0/0/**2**	1/0/**1**	-	0/0/**2**	0/0/**2**	0/1/**1**	0/0/**2**	1/0/**1**	0/1/**1**	-	**8**/1/1	**5**/0/0

Table 6.2: The distribution of the validity labels for the samples of the onboarding tasks presented as the number of each of the following decisions (Yes/Maybe/No), split into columns according to QADI's geolocated label of the samples. **Note #1:** The bolded value in each column represents the expected decision. **Note #2**: We initially discarded Libya, UAE, and Yemen from our dataset, and thus the onboarding datasets of the other countries do not have samples from these three countries. **Note #3**: I marked the unexpected patterns with *.

Samples from Algeria's Onboarding Task
هه ياي وبتصحح كمان مصدقة نفسها يابنتي مش هيعبرك برضو سايكو بجد هموت
بسم الله الرحمن راح نبدأ نعد واجيب البت دي تعدي معايا
عباس لسه ما فيه وسم لمظاهرة اليوم نزلو وقول وززززع وخليه يوصل ترند عالمي

Table 6.3: Three samples categorically annotated as invalid by the three Algerian annotators, yet are geolocated to Algeria according to QADI's test set.

On inspecting the annotators' performance on the onboarding task, I found that these automated checks are generally satisfied by the annotators, as per Table 6.2. The checks also helped us identify any misinterpretation of the guidelines, and provide feedback to our annotators before labeling the samples of the main task. For instance, one of the Algerian annotators *Algeria (C)* interpreted question (1) as identifying if the tweet matches how an Algerian typically writes, and consequently marked all the MSA samples as not valid in Algerian Arabic. For these sentences, he provided an alternative translation that sounds more natural to him (e.g., for the sample واحلى اخت انتي .. نفتخر انتي اخت مخيرة . و نعتز بيك اكيد ربي يحفظك, he provided the following alternative نزوخ و نستعرف بيك ربي يحفظك). Another Syrian annotator *Syria (C)* chose *Maybe/ Not Sure* for 4 out of the 5 MSA samples, as she understood the first question as if the sentence could have *only* been written by a speaker of Syrian Arabic. Her explanation for her choice is: *For the MSA sentences, it is hard to accurately identify the speaker's dialect, so I chose unsure*.

Moreover, I noticed that for the onboarding task of Algeria, three samples geolocated to Algeria (Table 6.3) are labeled as not valid in Algerian Arabic by the three Algerian annotators. Given that (a) the members of the team (I and Injy) responsible for the annotation process are native speakers of Egyptian Arabic only, and (b) the labels for the main task are not validated (i.e., assigned based on geolocation only), we could not judge whether these are errors in the labels provided by our annotators or if QADI's validated labels for these samples were incorrect. Therefore, we decided not to use these checks as automatic measures for accepting or rejecting the labels provided by the annotators for the main task batches and resorted to manually inspecting the annotations by the end of each annotation batch, as elaborated next.

Main Task Batches Following the onboarding task, I invited the annotators to label the task's data, split into 5 batches of 224 samples each. I ran the annotation batches over 5 weeks (1 batch per week) to ensure a higher annotation quality.

By the end of each batch, and as done for the onboarding task, we used the two aforementioned checks to inspect the quality of the annotations. Moreover, we compared the labels provided by the annotators of each country against each other. We also kept track of the quality using automatic IAA metrics, namely Fleiss' Kappa (κ) for the *Validity* label, and Krippendorff's alpha (α) for the *ALDi* label. For the first batch, we flagged all the instances of disagreement, asked the annotators to relabel them, and wrote comments in case these flagged instances were deemed invalid. This allowed us to have a better assessment of the reasons for disagreement and provide the annotators with tailored feedback accordingly.

For the following three batches, we tried to categorize clear patterns of disagreement between the annotators (e.g., an annotator systematically disagreeing with the other annotators) and discussed them individually with the annotators to rectify them in future batches. We have only asked them to relabel the samples of high disagreement in case we could not determine a pattern for the disagreement. For the last batch, we resorted to asking the annotators to relabel the samples of disagreement, to get an approximate evaluation of the impact of this process on the aggregated labels.

Analysis of the IAA Scores Table 6.4 demonstrates how the IAA scores (Fleiss' Kappa for the *Validity* label, and Krippendorff's Alpha for the *Validity* label) changed as the annotation process progressed. First, the values hint at acceptable levels of agreement between the annotators for both subtasks. However, we notice that the range of the IAA scores differs from one country to another, especially for the *Validity* label. The variation in the ranges of the IAA scores could be attributed to (a) the level of homogeneity between the dialects spoken in each country, and (b) the annotators' representativeness/knowledge of the different dialects spoken in their countries. Recruiting annotators from different regions within the same country (e.g., the case of the Algerian annotators) could increase the possibility of disagreement compared to when they all came from the same region (e.g., the case of the Egyptian annotators, where all are from Cairo).

Regarding the annotators' performance, I noticed that the agreement between the annotators categorically increased by asking them to reannotate the high-disagreement sentences for their validity in their country-level dialects. That said, the impact of

Country	Fleiss' Kappa (κ)					Krippendorff's Alpha (α)				
	Batch 1	**Batch 2**	**Batch 3**	**Batch 4**	**Batch 5**	**Batch 1**	**Batch 2**	**Batch 3**	**Batch 4**	**Batch 5**
Algeria	.49 (59) .58 (59)	.49 (80) -	.54 (65) -	.51 (67) .61 (70)	.48 (59) .59 (59)	.745 (59) .663 (59)	.615 (80) -	.708 (65) -	.513 (67) .536 (70)	.715 (59) .666 (59)
Morocco	.62 (34) -	.42 (28) .81 (53)	.27 (49) .5 (50)	.36 (41) .53 (46)	.48 (43) .62 (47)	.823 (34) -	.691 (28) .76 (53)	.767 (49) .742 (50)	.768 (41) .811 (46)	.687 (43) .742 (47)
Tunisia	.43 (28) .56 (41)	.67 (47) -	.64 (33) .71 (40)	.53 (31) .7 (31)	.46 (32) .71 (30)	.798 (28) .787 (41)	.738 (47) -	.8 (33) .808 (40)	.71 (31) .722 (31)	.664 (32) .698 (30)
Egypt	.58 (62) .7 (61)	.63 (69) -	.56 (68) -	.64 (81) .74 (81)	.69 (74) .79 (74)	.845 (62) .796 (61)	.828 (69) -	.791 (68) -	.791 (81) .788 (81)	.862 (74) .82 (74)
Sudan	.57 (68) .72 (74)	.53 (76) -	.58 (84) -	.56 (81) .74 (81)	.67 (77) .79 (78)	.765 (68) .746 (74)	.537 (76) -	.657 (84) -	.696 (81) .727 (81)	.624 (77) .643 (78)
Palestine	.4 (114) .58 (111)	.41 (59) -	.52 (72) -	.51 (61) .69 (67)	.54 (71) .74 (66)	.731 (114) .645 (111)	.752 (59) -	.673 (72) -	.633 (61) .739 (67)	.559 (71) .573 (66)
Syria	.39 (83) .56 (89)	.49 (92) -	.49 (87) -	.59 (109) -	.53 (90) .57 (98)	.845 (83) .829 (89)	.709 (92) -	.866 (87) -	.751 (109) -	.774 (90) .796 (98)
Iraq	.59 (58) -	.52 (44) -	.59 (50) -	.61 (51) .69 (62)	.59 (53) .64 (57)	.677 (58) -	.684 (44) -	.724 (50) -	.733 (51) .776 (62)	.795 (53) .816 (57)
Yemen	.46 (101) .55 (104)	.57 (99) -	.52 (81) -	.45 (76) -	.45 (94) .5 (94)	.561 (101) .498 (104)	.495 (99) -	.457 (81) -	.568 (76) -	.397 (94) .433 (94)

(a) Validity Labels (b) ALDi Labels

Table 6.4: The detailed IAA scores for each of the 5 main annotation tasks, computed independently for each country's 3 annotators. The second line for each country represents the IAA scores after providing feedback to the annotators and asking them to reannotate the samples of high disagreement. The number of sentences valid in each country-level dialect after applying majority voting is shown (in brackets).

this relabeling process on the number of valid sentences according to the majority voting is minimal for the last annotation batches. This increase in the agreement scores post-relabeling was not as consistent for the ALDi levels, in which we sometimes noticed insignificant decreases. This could be attributed to the subjectivity of the ALDi Estimation task, compared to the Validity task. Lastly, the annotators' performance, measured by the IAA scores, was consistent across the different annotation batches, showcasing the effectiveness of our process.

Country	Validity Labels			ALDi Labels
	Fleiss κ	N valid	N ¬valid	Krip. α
Algeria	0.56	333 (205)	787 (666)	0.66
Morocco	0.62	230 (152)	890 (784)	0.74
Tunisia	0.67	189 (129)	931 (879)	0.75
Egypt	0.69	353 (257)	767 (682)	0.82
Sudan	0.67	393 (283)	727 (619)	0.66
Palestine	0.59	375 (245)	745 (587)	0.68
Syria	0.54	475 (305)	645 (543)	0.79
Iraq	0.61	271 (171)	849 (738)	0.73
Yemen	0.52	454 (291)	666 (477)	0.50

Table 6.5: Interannotator agreement scores – Fleiss' Kappa (κ) for Subtask 1 and Krippendorff's Alpha-interval method (α) for Subtask 2 – for the full dataset. I also report the number of valid, not valid sentences out of the 1,120 according to majority voting, while showing the number of sentences with complete agreement (in brackets). **Note:** The country-level Krippendorff's Alpha scores are computed for their respective valid samples, for which ALDi ratings of this country exist.

Overall Interannotator Agreement (IAA) Scores The country-level Fleiss' Kappa (κ) (Fleiss, 1971), and Krippendorff's Alpha (α) (Krippendorff, 2004) scores for the whole dataset in Table 6.5 indicate moderate to substantial agreement between the annotators for both subtasks. Moreover, there is no noticeable variation among the scores across the countries, except for the α score for the Yemen annotators, which is slightly lower than those of the other countries. More specifically, I noticed that some sentences rated as *Sound MSA (Level 0)* by some Yemeni annotators are annotated as *Natural/Ordinary Colloquial (Level 2)* by others.

6.1.4 Label Aggregation Techniques

Validity A sentence is considered valid in a country-level dialect if among the three annotators from the respective countries: a) one of them answered *Yes*, and b) another answered *Yes* or *Maybe*. On average, the same-country annotators fully agreed on the validity of more than 66% of the valid samples, and the invalidity of more than 85% of the invalid samples, as per Table 6.5.

ALDi For each sentence, the ordinal ALDi levels assigned by the annotators from the different countries are aggregated into a single numeric value $\in [0, 1]$. Discrete ALDi levels (0, 1, 2, 3) are transformed into the following numeric values $(0, \frac{1}{3}, \frac{2}{3}, 1)$. The algebraic mean of these numeric values is used as the overall ALDi score for the sentence.

As mentioned in §6.1.2, annotators only assigned ALDi levels to sentences they rated as valid in their country-level dialect. Consequently, the number of ALDi annotations per sentence can range from 0 to 3*N, where N is the number of countries from which annotators are recruited. If a sentence is deemed invalid according to the majority vote label (Subtask 1) for a country-level dialect, we discard the respective ALDi annotations (if any) assigned by annotators from this country.

6.1.5 Formation of Development/Test Sets

I used 120 samples from the first batch as the development sets shared with the participating teams. The first batch's remaining samples and the samples of the 4 succeeding batches form the test sets. For the *ALDi* label, samples that are not valid in the considered dialects of the corresponding set have no assigned ALDi ratings and thus are not released as part of the ALDi estimation development/test sets.

6.2 Shared Task Description

6.2.1 Subtask 1 – Multi-Label Dialect Identification

In this subtask, we proposed multi-label dialect identification at the country level. The objective is to evaluate the feasibility of using single-label Arabic dialect identification datasets to train a multi-label system that can predict *all* dialects in which a given sentence is valid.

Training Data We provided participants with the *training* splits of following datasets: `NADI-2020-TWT`, `NADI-2021-TWT`, and `NADI-2023-TWT` Abdul-Mageed et al. (2020a, 2021b, 2023).

Dev and Test Data We provided the new multi-label development set, as explained in §6.1.5. This dataset has 120 samples with manually assigned validity labels of eight different Arab countries: *Algeria, Egypt, Jordan, Palestine, Sudan, Syria, Tunisia,* and

Sentence	GEO	Valid in	ALDi
اللهم انت ربي لا اله الا انت خلقتني وانا عبدك وانا علي عهدك ووعدك ما استطعت اعوذ بك من شر ما صنعت #اذكار_الصباح_و_المساء	SA	DZ, EG, JO, PS, SD, SY, TN, YE	0.00
#تريكه_في_كاس_العالم شاهد ماذا قال #الغندور علي المطالبه برجوع ابو تريكه للعب مع مصر في كاس العالم	EG	DZ, EG, JO, PS, SD, SY, TN, YE	0.15
لمن الحياه ترسل ليك رساله	SD	PS, SD, YE	0.58
وين يلعب هذا ما شفته	AE	DZ, PS, YE	0.64
الحمد لله الجو برد الايمات يلي فاتوا الواحد مغموم اتقول مكسد بين فرادي تينه شمام	LY	DZ	0.83
ايش دخل عارك ياحسن زميطه هذي العينات اللي يشتي له قيرعي	YE	YE	1.00

Table 6.6: Sample sentences from the development dataset with their geolocated country (**GEO**), valid dialect labels (Subtask 1), and ALDi scores (Subtask 2). **DZ**: Algeria, **EG**: Egypt, **JO**: Jordan, **LY**: Libya, **PS**: Palestine, **SA**: Saudi Arabia, **SD**: Sudan, **SY**: Syria, **TN**: Tunisia, **AE**: UAE, **YE**: Yemen.

Yemen. Examples of those sentences are provided in Table 6.6. We do not restrict systems to these eight dialects. Hence, we include two undisclosed dialects in our test data and ask participants to develop their models such that they can predict all valid dialects out of the 18 country-level ones from NADI 2023. The undisclosed dialects are *Iraq* and *Morocco*. Accordingly, the test set contains 1,000 samples covering nine dialects.[3]

Restrictions Subtask 1 operated under a ***closed-track*** policy where participants are allowed to *only* use the datasets we provide for system training.

6.2.2 Subtask 2 – ALDi Estimation

This subtask assesses the ability to estimate aggregated ALDi scores of tweets, a genre different from the online comments in the AOC dataset used in chapter 3.

[3]I also note that one of the Jordanian annotators did not complete the labeling process on time, and so I did not include the labels from Jordanian annotators in the test sets.

Training Data We provided the teams with the AOC-ALDi dataset's training split (chapter 3).

Dev and Test Data A second layer of annotation for manual ALDi levels was used for the same samples of Subtask 1.

Restrictions Subtask 2 operated under an ***open-track*** policy, allowing participants to train their systems on any additional datasets of their choice, provided that they explain the sources of the data and how it is used, and that these additional training datasets are public at the time of submission.

6.2.3 Evaluation Metrics

The official evaluation metric for Subtask 1 is the macro-averaged F_1 score. More specifically, we compute the F_1 score independently for each country in the evaluation dataset (eight for the development set and nine for the test set), then compute the average of these individual-country F_1 scores.[4] Additionally, we report system performance in terms of `Precision`, `Recall`, and `Accuracy` for submissions to Subtask 1. The metric for Subtask 2 is the `Root Mean Square Error (RMSE)`.

6.2.4 Submission Rules

Participating teams were allowed to submit up to *five* runs for each subtask. For each team, only the submission with the highest score was retained. While the official results were exclusively based on a blind test set, we requested participants to include their results on the development splits in their papers. To facilitate the evaluation of participant systems, we established a CodaLab competition for scoring each subtask.[5] Since subtasks 1 and 2 are proposing new tasks, the organization team is happy to share the individual labels of the development/test sets for these two subtasks with researchers interested in analyzing them.[6]

[4]Participating teams submitted validity predictions for the 18 countries of the training sets. I plan to rerun the evaluation upon collecting labels for more country-level dialects.

[5]The CodaLabs for the first two subtasks are available at: `Subtask 1`, and `Subtask 2`.

[6]Interest could be expressed through the following form: `https://forms.gle/gdgTToxG2tH5xT27A`

Team	Affiliation	Task
dzNLP (Lichouri et al., 2024)	USTHB, Algeria	1
Elyadata (Karoui et al., 2024)	Elyadata, Tunisia	1
NLP_DI (Kanjirangat et al., 2024)	Dalle Molle Ins. of A.I., Switzerland	1
AlexUNLP-STM (Sakr et al., 2024)	Alexandria University, Egypt	2
ASOS (Nacar et al., 2024)	Prince Sultan University, KSA	2
CUFE	Cairo University, Egypt	2

Table 6.7: List of teams that participated in NADI-2024 shared task. Teams with accepted papers are cited.

6.2.5 Participating Teams

At the testing phase, a total of 17 valid entries were submitted by six unique teams. The breakdown across the subtasks is as follows: *ten* submissions for Subtask 1 from *three* teams, and *seven* submissions for Subtask 2 from *three* teams. Five out of the six teams that participated in either of the first two subtasks submitted system description papers, as indicated in Table 6.7.

6.3 Shared Task Baselines and Results

6.3.1 Baselines

I developed baseline (BL) models for the first two subtasks for comparison against the teams' systems, as described below. These models were not shared with participating teams during the competition.

Subtask 1 Baselines I used a fine-tuned single-label ADI system with a softmax activation function to develop two baselines.[7] The first predicts the most probable labels such that their cumulative probability is $> 90\%$. The second assumes the sentence is only valid in the most probable prediction. Lastly, I implemented a Random baseline that generates random binary predictions for the validity of the sentences in the considered dialects.

[7]The fine-tuned baseline model can be accessed through `https://huggingface.co/AMR-KELEG/NADI2024-baseline`.

Subtask 2 Baselines I first used the *Sentence ALDi* model, introduced in §3.3.1, as the supervised baseline. The second baseline is based on the distribution of the ALDi scores for the development set (Figure 6.6), where I implement a model that generates a constant score of 0.67 for all the sentences. In the third baseline, I use a Random ALDi generator ($\in [0, 1]$).

6.3.2 Shared Task Results

Subtask 1 ***Elyadata*** came first with a macro-averaged F_1 score of 50.57%, being the only team to beat the *Top 90% baseline* model as per Table 6.8.

Subtask 2 ***ASOS***, the top-performing team, achieved the lowest RMSE of 0.1403, while ***AlexUNLP-STM*** achieved a similar RMSE of 0.1406, coming second. As shown in Table 6.9, all the teams managed to improve over the baselines.

Rank	System	Macro-average			
		Accuracy (↑)	**Precision (↑)**	**Recall (↑)**	**F_1 score (↑)**
1	**Elyadata**	$67.50_{\pm 3.7}$	$46.48_{\pm 10.1}$	$\mathbf{57.09_{\pm 5.1}}$	$\mathbf{50.57_{\pm 7.1}}$
BL I	Top 90%	$73.40_{\pm 6.1}$	$60.67_{\pm 14.5}$	$39.22_{\pm 14.6}$	$45.09_{\pm 11.3}$
2	**NLP_DI**	$71.88_{\pm 5.6}$	$53.64_{\pm 10.2}$	$37.42_{\pm 11.0}$	$43.27_{\pm 9.4}$
BL II	Random	$50.14_{\pm 1.6}$	$30.43_{\pm 8.8}$	$50.15_{\pm 2.1}$	$37.15_{\pm 7.2}$
BL III	Top 1	$\mathbf{73.42_{\pm 7.6}}$	$\mathbf{76.82_{\pm 10.6}}$	$17.77_{\pm 10.8}$	$27.30_{\pm 12.6}$
3	**dzNLP**	$71.38_{\pm 7.2}$	$63.22_{\pm 10.7}$	$12.87_{\pm 3.8}$	$20.98_{\pm 5.2}$

Table 6.8: Systems' performance on the test set of Subtask 1. The standard errors of these macro-averaged metrics are also reported.

Rank	System	RMSE (↓)
1	**ASOS**	**0.1403**
2	**AlexUNLP-STM**	0.1406
3	**CUFE**	0.2001
BL I	Sentence ALDi	0.2178
BL II	Constant (0.67)	0.2361
BL III	Random	0.3521

Table 6.9: Systems performance on Subtask 2 test set.

6.3.3 General Description of Submitted Systems

A summary of approaches employed by the various teams is provided in Table 6.10. I briefly describe the teams' systems for each subtask.

Task	Team	Metric	Features		Techniques						
			N-gram	TFIDF	C-ML	NNs	PLM	Ensemble	Post-Poc.	Cont. L.	D-Aug.
1	Elyadata	50.57				✓	✓	✓			
	NLP_DI	43.27				✓	✓	✓	✓		
	dzNLP	20.98	✓	✓	✓			✓			
2	ASOS	0.1403				✓	✓	✓			✓
	AlexUNLP-STM	0.1406				✓	✓	✓		✓	

Table 6.10: Summary of approaches used by participating teams NADI 2024 shared task. Teams are sorted by their performance on the official metric of each subtask. *C-ML* (Classifcal ML) indicates any non-neural machine learning methods such as naive Bayes and support vector machines. The term *NNs* refers to any model based on neural networks (e.g. RNN, CNN, and Transformer) trained from scratch. *PLM* refers to neural networks pretrained with unlabeled data such as MARBERT and has less than 1B parameters. Approaches also included contrastive loss (*Cont. L.*) and data augmentation (*D-Aug.*)

Subtask 1 The winning team, ***Elyadata***, extracted dialectal vocabularies from the training data, and used them to augment the labels of the single-label training dataset. They then used a max pooling layer to merge the predictions of a MARBERT-based ensemble model, forming an array of logit predictions. Lastly, they optimized a threshold using the development set to convert the logits into multi-label predictions. ***NLP$_{DI}$*** used a sentence embedding model to encode the datasets of the training set. For each sentence S_{test} of the test dataset, it is embedded into the same space, and then a classifier is used to identify the nearest 10 sentences to its embedding. These 10 sentences' labels are used as the predictions for sentence S_{test}. ***dzNLP*** trained a support-vector machine (SVM) on top of word-level and character-level ngram features. Their system description paper does not explicitly mention whether their system predicts a single label or multiple labels for an input sentence. The low recall scores they achieve hint that their system generates a single label (see Table 6.8).

Subtask 2 ***ASOS*** added a regression head on top of MARBERT's [CLS] embedding. More specifically, they had three hidden layers with non-linear activation functions on top of the [CLS] embedding, with a final linear output layer that predicts a scalar ALDi estimation. They then fine-tuned the whole model using the provided training dataset. ***AlexUNLP-STM*** used the median of an ensemble of regression heads with sigmoid activation on top of AraBERT, trained to minimize contrastive and RMSE losses. Noticeably, their model's performance dropped when non-Arabic letters were discarded. We observed that code-switching affected the annotators' ALDi judgments differently, which is in line with the team's justification for the performance drop. ***CUFE*** did not submit their system details.

6.3.4 Detailed Analysis of Subtask 1 Results

As described in §6.1.1, 70 out of the 1,120 samples used to form the development and test sets for Subtasks 1 and 2 are automatically identified as being in MSA. For Subtask 1, these MSA samples are expected to be labeled as valid in all the considered dialects. On checking the validity labels for these samples, we indeed found that they are mostly deemed valid in all the considered country-level dialects. The developed systems are expected to predict that these sentences are valid in all the considered dialects.

Consequently, I report their performance on the automatically identified DA and MSA samples, respectively, in Table 6.11. Since the MSA samples represent a small proportion of the development and test sets, I find that the models' performance on the DA samples is not different from their overall performance reported in Table 6.8.

Rank	System	Macro-average			
		Acc. (↑)	**Prec.** (↑)	**Recall** (↑)	**F_1 score** (↑)
1	**Elyadata**	$69.27_{\pm 4.3}$	$43.07_{\pm 11.0}$	$\mathbf{62.17_{\pm 5.6}}$	$\mathbf{49.85_{\pm 8.3}}$
BL I	Top 90%	$75.99_{\pm 6.2}$	$57.08_{\pm 15.4}$	$41.92_{\pm 14.9}$	$45.21_{\pm 10.3}$
2	**NLP_DI**	$74.41_{\pm 6.2}$	$49.56_{\pm 11.0}$	$39.70_{\pm 11.4}$	$43.02_{\pm 9.3}$
BL II	Random	$50.06_{\pm 1.8}$	$26.09_{\pm 9.3}$	$49.92_{\pm 3.3}$	$33.31_{\pm 8.4}$
BL III	Top 1	$\mathbf{77.40_{\pm 8.0}}$	$\mathbf{75.20_{\pm 11.3}}$	$20.52_{\pm 11.7}$	$30.37_{\pm 12.8}$
3	**dzNlp**	$75.42_{\pm 7.7}$	$61.21_{\pm 11.7}$	$15.17_{\pm 5.3}$	$23.61_{\pm 6.7}$

(a) DA samples

Macro-average			
Acc. (↑)	**Prec.** (↑)	**Recall** (↑)	**F_1 score** (↑)
$40.68_{\pm 9.4}$	$95.92_{\pm 3.4}$	$39.02_{\pm 10.6}$	$54.58_{\pm 9.1}$
$34.23_{\pm 15.3}$	$96.43_{\pm 4.2}$	$30.75_{\pm 16.2}$	$44.65_{\pm 17.8}$
$33.51_{\pm 15.2}$	$97.40_{\pm 3.1}$	$30.56_{\pm 15.6}$	$44.48_{\pm 16.1}$
$\mathbf{51.43_{\pm 5.6}}$	$94.36_{\pm 4.5}$	$\mathbf{51.34_{\pm 6.0}}$	$\mathbf{66.30_{\pm 5.5}}$
$13.26_{\pm 7.7}$	$\mathbf{100.00_{\pm 0.0}}$	$7.81_{\pm 7.4}$	$13.71_{\pm 11.3}$
$10.22_{\pm 5.3}$	$85.19_{\pm 31.9}$	$5.22_{\pm 4.4}$	$9.49_{\pm 7.7}$

(b) MSA samples

Table 6.11: The performance of the systems submitted to Subtask 1 on the DA and MSA samples of the test set. The systems are ordered according to their macro-averaged F1 scores on the whole test set as indicated in Table 6.8. **Note:** The standard errors of these macro-averaged metrics are also reported.

For the MSA samples, I notice that the macro-average Recall needs to be improved. A two-stage solution could be proposed in which a classifier first identifies if a sentence is in MSA or DA. MSA sentences can be predicted to be valid in all the considered dialects with high accuracy. Conversely, the validity labels for the DA samples could be identified using another multi-label dialect identification system.

Regional Level Performance The results in Tables 6.8, 6.11 indicate that there is room for improvement for the multi-label ADI systems to be reliably able to accurately operate on the country-level. To further analyze the results, I group the nine country labels of the test set into macro-regional dialects according to Baimukan et al. (2022) as follows: *Maghreb* (Algeria, Tunisia, Morocco), *Nile Basin* (Egypt, Sudan), *Levant* (Palestine, Syria), *Gulf* (Iraq), and *Gulf of Aden* (Yemen). For each region, a sample is considered valid in the region if it is valid in any of the region's countries for which we have validity labels. For example, a sample annotated as valid in Algeria, Tunisia, and Sudan will be considered valid in *Maghreb* and *Nile Basin*. I similarly consider the systems' predictions for the same nine countries and aggregate them into macro-regional dialects.

The models' performance predicting the macro-regional dialects is higher than that for country-level ones as per Table 6.12. However, the improvement is not as great as might have been expected, indicating that even multi-label macro-regional dialect identification is a challenging task. In chapter 7, I will explain how I extended the labels in our test set to cover more countries, especially from the *Gulf* region.

Rank	System	Macro-average				Individual Region F_1 score (↑)				
		Acc. (↑)	Prec. (↑)	Recall (↑)	F_1 score (↑)	$Maghreb_3$	$Nile_2$	$Levant_2$	$Gulf_1$	$Gulf\ of\ Aden_1$
1	**Elyadata**	$68.02_{\pm 4.1}$	$52.25_{\pm 12.0}$	$67.16_{\pm 5.4}$	$\mathbf{58.18_{\pm 8.9}}$	55.42	68.54	**67.81**	45.21	**53.89**
BL I	Top 90%	$\mathbf{73.07_{\pm 4.7}}$	$62.54_{\pm 14.0}$	$54.28_{\pm 14.8}$	$56.16_{\pm 12.3}$	**61.08**	**69.58**	64.20	**51.17**	34.76
2	**NLP_DI**	$71.73_{\pm 4.6}$	$57.50_{\pm 12.2}$	$49.65_{\pm 13.3}$	$53.09_{\pm 12.7}$	54.71	67.97	65.65	36.69	40.43
BL II	Random	$46.91_{\pm 5.8}$	$35.52_{\pm 9.8}$	$\mathbf{69.01_{\pm 15.3}}$	$46.19_{\pm 10.9}$	45.59	56.40	58.06	27.91	43.00
BL III	Top 1	$72.52_{\pm 7.5}$	$\mathbf{77.74_{\pm 13.6}}$	$28.87_{\pm 12.5}$	$40.25_{\pm 14.8}$	50.59	60.45	31.84	40.55	17.80
3	**dzNlp**	$69.94_{\pm 7.6}$	$68.39_{\pm 10.0}$	$21.85_{\pm 9.1}$	$32.42_{\pm 11.5}$	44.14	44.27	34.09	25.55	14.03

Table 6.12: The performance of the systems submitted to Subtask 1, in predicting multi-label macro-regional dialects for the DA samples of the test set. In addition to the Macro-average F1 score, the individual F1 score for each region is reported. **Note:** the countries representing the regions are: *Maghreb* (Algeria, Tunisia, Morocco), *Nile* (Egypt, Sudan), *Levant* (Palestine, Syria), *Gulf* (Iraq), and *Gulf of Aden* (Yemen).

6.3.5 Performance of Baselines Based on Models Released after the Shared Task

Given the rapid improvement in the field, one might wonder how recent models would fare on both subtasks. To this end, I tested two new sets of baselines:

Fine-tuning Multilingual Embedding Models In this setup, I replicate the baseline (BL I) of both tasks, with the only difference of replacing the underlying BERT-based model with the Embedding-Gemma (300M) model (Schechter Vera et al., 2025). Table 6.13 and Table 6.14 show that changing the backbone from an Arabic-specific model to a more powerful multilingual model hurts the performance in both MLADI and ALDi estimation, as indicated by the decrease in the performance of the BL I' models in comparison to the BL I models. Adaptors have been used as a more efficient way to fine-tune models; however, full fine-tuning remains to be the best-performing method in medium and high-resource settings (i.e., ALDi Estimation) (Chen et al., 2022).

Prompting Instruction-tuned Large Language Models (LLMs) Instead of fine-tuning task-specific models, prompting has been proposed as a more generic paradigm. Aya Expanse (32B) is a powerful open-source multilingual model, supporting 8 languages, including Arabic (Dang et al., 2024). The model was instructed to assess the possibility that a sentence is acceptable in one of the dialects used in each of the nine considered dialects of Subtask 1 independently, in a zero-shot and a few-shot setup, as shown in Figure 6.4. For the few-shot setup, two exemplar samples from the development set (one positive and one negative) were added to the prompt.

The prompting method in the zero-shot setup (LLM I) achieved the highest F1-score of 51.85, thanks to a higher macro-averaged recall of 65.06 compared to only 57.09 for the top-ranked team (Elyadata). Moreover, the few-shot setup achieved a much higher recall of 92.89, but at the expense of precision, yielding a lower macro-averaged F1-score of 46.82. While prompting provides a competitive baseline, further efforts are still needed to solve the MLADI task.

Rank	System	Macro-average			
		Accuracy (↑)	Precision (↑)	Recall (↑)	F_1 score (↑)
LLM I	**Aya-expanse (32B) $_{\text{zero}}$**	$64.9_{\pm 3.7}$	$44.61_{\pm 10.9}$	$65.06_{\pm 5.6}$	$\mathbf{51.85_{\pm 7.3}}$
1	Elyadata	$67.50_{\pm 3.7}$	$46.48_{\pm 10.1}$	$57.09_{\pm 5.1}$	$50.57_{\pm 7.1}$
LLM II	**Aya-expanse (32B) $_{\text{few}}$**	$38.07_{\pm 6.5}$	$31.83_{\pm 8.3}$	$\mathbf{92.89_{\pm 2.5}}$	$46.82_{\pm 9.4}$
BL I	Top 90%	$\mathbf{73.40_{\pm 6.1}}$	$\mathbf{60.67_{\pm 14.5}}$	$39.22_{\pm 14.6}$	$45.09_{\pm 11.3}$
BL I'	**Top 90% (Gemma)**	$70.17_{\pm 5.6}$	$49.86_{\pm 12.6}$	$34.62_{\pm 15.3}$	$38.91_{\pm 12.8}$

Table 6.13: Systems' performance on the test set of Subtask 1. The standard errors of these macro-averaged metrics are also reported. The three newly tested baselines (LLM I, LLM II, BL I') are shown in bold.

Rank	System	RMSE (↓)
1	ASOS	**0.1403**
BL I	Sentence ALDi	0.2178
BL I'	**Sentence ALDi (Gemma)**	0.2241

Table 6.14: Systems performance of different models on Substask 2's test set, compared to the performance of the newly tested Sentence ALDi (Gemma) baseline.

{
"*role*": "user",
"*content*":
حدد إذا كانت الجملة الأتية مقبولة في أحد اللهجات المستخدمة في «دولة س». أجب ب نعم أو لا فقط.
الجملة: «جملة»
}

{
"*role*": "user",
"*content*":
Specify if the following sentence is acceptable in one of the dialects used in <Country X>. Answer with yes or no only.
The sentence: <sentence>
}

Figure 6.4: The prompt used to test the Aya-exanse model's ability to determine if a sentence is acceptable in each of the 9 country-level dialects of Subtask 1 in the zero-shot setup, with its English translation.

6.4 Discussion

Analysis of the Labels Distribution Figure 6.5 shows that 153 samples are labeled as invalid in all of the nine considered countries, and 310 samples are valid in only one dialect. I also report that 657 samples are valid in 2 or more country-level dialects (i.e., 58.7% of the samples). This percentage is only expected to increase when validity annotations from other country-level Arabic dialects are collected.

The aggregated ALDi scores have a multimodal distribution as per Figure 6.6. The first mode is related to the automatically identified MSA samples in the dataset (70 in total). All of these samples are assigned ALDi scores <0.2, and are judged as valid in all or almost all of the considered country-level dialects. Conversely, the ALDi scores for the automatically identified DA samples are distributed around a score of 0.66 (the numeric value corresponding to Level 2 (*Natural/Ordinary Colloquial*).

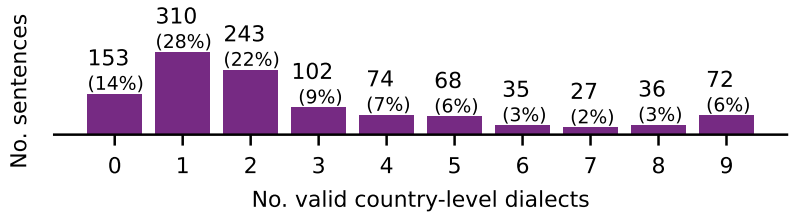


Figure 6.5: The distribution of the number of valid country-level dialects out of 9 countries for the full dataset.

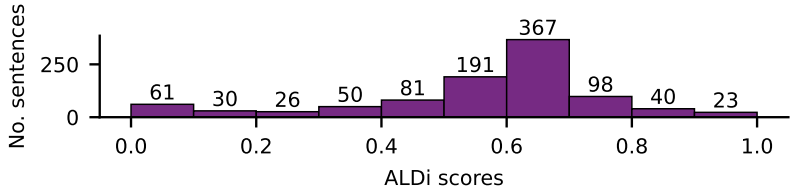


Figure 6.6: The distribution of the ALDi scores for the samples of the full dataset.

Precision of Geolocated Labels Although geolocation can alleviate the need for manually annotating the samples (Abdul-Mageed et al., 2021b), it can be error-prone (Abdul-Mageed et al., 2020a; Abdelali et al., 2021). For the 1,050 DA samples of the development and test sets, we can estimate the precision of the geolocated labels by comparing them to the manual validity labels as demonstrated in Figure 6.7.

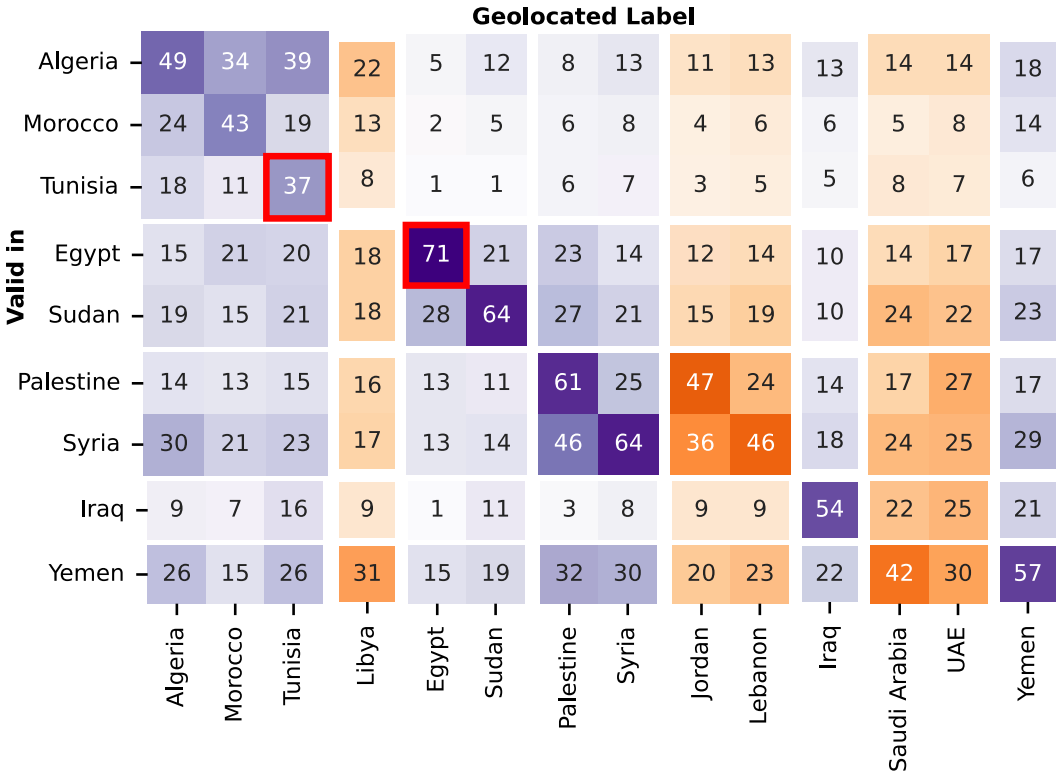


Figure 6.7: The number of DA samples valid in the annotators' country-level dialects (rows) across the 14 countries to which the samples are geolocated (columns). Each row represents the distribution of the geolocated labels for the sentences valid in the row's country-level dialect. Orange columns indicate the countries not represented by the annotators. The max cell value is 75.

Based on this method, I find that the precision of the geolocated labels could be as high as 94.6% (71 of 75 samples) for Egypt, and as low as 49.3% (37 of 75 samples) for Tunisia.

Impact of Named Entities ADI models, trained on single-label data, can make spurious connections between named entities (e.g., specific locations) and country-level labels (Abdul-Mageed et al., 2020b). In `NADI-2021-TWT`, for example, 52 samples out of the 66 mentioning لبنان (Lebanon) are geolocated to and labeled as *Lebanon*. Such spurious connections might be the reason why the following n-grams عراق, لبنان, تونس, يمن are among the most discriminative for the dialects of *Iraq*, *Lebanon*, *Tunisia*, and *Yemen*, respectively (AAlAbdulsalam, 2022). The manual annotation process used for this dataset greatly alleviates this limitation.

6.5 Lessons Learned

I share my reflections on the creation of evaluation datasets for Subtasks 1 and 2.

Subtask 1 Complexity Previous research asking Arabic speakers to check the validity of sentences in their native dialects (See Table 6.1) reported moderate to high agreement between the annotators (only two per country) for *most* of the considered regional-level and country-level dialects. Unlike previous works, we asked the annotators to judge all samples, rather than those geolocated to their own respective countries. Therefore, our annotation task is possibly harder than previous ones, which is reflected in the IAA scores in Table 6.5.

Subtask 1 Labels' Set From a task design perspective, I observed that the frequency of usage of the *Maybe (Not sure)* label varies across annotators. For this reason, including this particular label (rather than using a binary *Valid/Not Valid* setup), needs to be further investigated to understand its implications on the aggregated validity labels.

Annotation Quality Monitoring Two authors—I and Injy Hamed—who are speakers of Egyptian Arabic were responsible for monitoring the quality of the annotations, providing feedback, and marking the samples with high disagreement for reannotation. I believe that having dialect leads who are native speakers of the different Arabic dialects would allow for better monitoring of the annotation process. I hope that our shared task will inspire future collaborative research to extend the labels of our evaluation dataset to include more country-level dialects.

Subtask 2 Guidelines The Krippendorff's alpha (α) scores in Table 6.5 are higher than the corresponding alpha score of 0.63 for the AOC-ALDi dataset (see §3.2.2). Two potential reasons could explain this improvement in agreement. First, the alpha scores for the new dataset are computed independently for each country's three annotators. Annotators from the same country are more likely to provide similar ALDi ratings than annotators from different countries (the latter was the case for the AOC-ALDi dataset). Having a more specified description of each dialect level in the new guidelines might have also contributed to this.

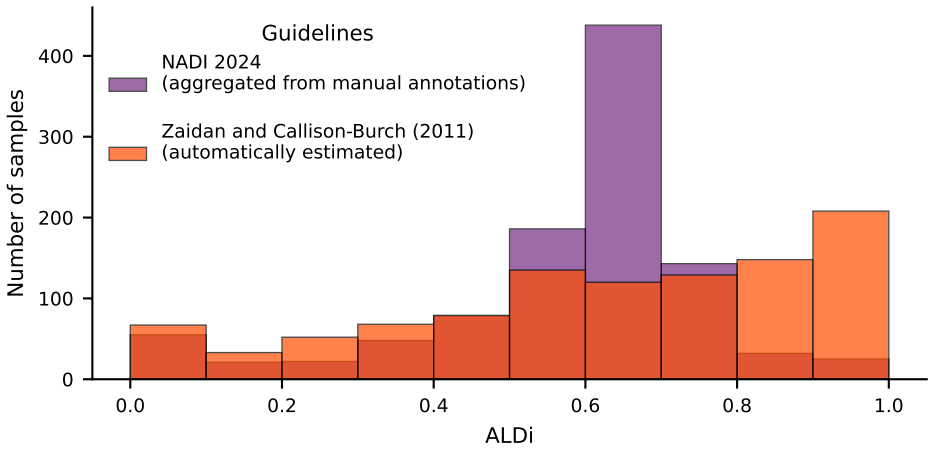


Figure 6.8: The distribution of the ALDi scores assigned to the NADI 2024 samples: (1) aggregated from the manual annotations according to the newly proposed guidelines, and (2) automatically estimated by the *Sentence-ALDi* model, based on Zaidan and Callison-Burch's (2011) guidelines.

To further investigate the differences between the newly proposed guidelines and AOC's guidelines, I employ a similar setup to the one in §3.4. For each sentence of the NADI 2024 dataset, I automatically estimate an ALDi score using the *Sentence ALDi* model, which would be a proxy for the ratings provided according to the AOC-ALDi guidelines. Afterward, I contrast these automatically estimated scores to their corresponding ALDi scores, aggregated from the individual ratings according to the new guidelines. Figure 6.8 shows the histograms of the ALDi scores according to both guidelines. The bins corresponding to low ALDi scores seem to be similar for both guidelines. However, a notably large number of sentences have ALDi scores in [0.6, 0.7] according to the new guidelines compared to a much smaller number of sentences having scores in this range according to the *Sentence ALDi* guidelines. This indicates that many sentences ended up rated as *(L2) Natural/Ordinary Colloquial*, which corresponds to a numeric value of $\frac{2}{3} \approx 0.67$. Conversely, the *Mixed* level according to Zaidan and Callison-Burch's (2011) guidelines would be mapped to this numeric value. Potentially, sentences rated as *Natural/Ordinary Colloquial* could be rated as *Mixed* or *Mostly Dialectal* according to the proportion of the colloquial terms in these sentences that are shared with MSA. This could explain the noticed difference in the number of sentences within the [0.6, 0.7] range. Moreover, this is a limitation of the new guidelines that future work should try to solve, to better distinguish between colloquial sentences that share a lot of terms with MSA and others that do not share as many terms with MSA.

Another limitation of the new guidelines is that they rely on sociolinguistic factors to distinguish between the last two dialect levels *(L2) Natural/Ordinary Colloquial*. This could potentially introduce biases where dialects linked to less prestige are rated as more colloquial than others linked to higher prestige. However, qualitatively inspecting the sentences with high ALDi scores ([0.8, 1]) according to the new guidelines reveals that the vast majority of them are vulgar. In our dataset, there are no clear cases of sentences with high ALDi scores that are in dialects generally associated with lower prestige. One potential reason is that speakers of these dialects might refrain from using them online, as indicated by Mohamed Eida et al. (2024) for Sa'idi Arabic—an Arabic variety used in upper Egypt.

6.6 Summary

In this chapter, I described the process of creating the first multi-label ADI dataset in collaboration with the NADI shared task's organizers, following the proposal provided in the previous chapter. The dataset comprises 1,120 geolocated samples, uniformly distributed across 14 different Arab countries, with manually assigned validity labels and dialectness ratings by 27 annotators from 9 countries (3 annotators each). The dataset was used as the evaluation set of the first two subtasks of the NADI 2024 shared task: multi-label dialect identification and Arabic level of dialectness (ALDi) estimation. For the former, the top-performing team was the only one that beat the baseline, achieving a macro-averaged F1 score of 50.57. This indicates a clear gap in building effective multi-label ADI models. For ALDi estimation, the three participating teams outperformed the *Sentence-ALDi* baseline model, with the first-place team achieving an RMSE of 0.1403 on the newly introduced evaluation dataset. This once again shows that ALDi could be effectively estimated using NLP models addressing (***RQ1***) ***How can the concept of Dialect Levels be operationalized in a way that can be effectively estimated?***

In addition to serving as an evaluation set for the shared task, the new dataset allowed for further investigation of different questions related to ADI. For (***RQ4***) ***How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?*** it shows that a large proportion (58.7% of the dataset) is valid in multiple dialects. The next chapter will further investigate widely held assumptions about Arabic and its dialects, using an extended version of the dataset. It will also analyze the impact of the annotators' native dialects on their corresponding ALDi ratings.

Chapter 7

Revisiting Common Assumptions about the Arabic Dialects in NLP

Successful Arabic NLP systems need to handle both the interspeaker and intraspeaker variations, yet some literature rests on certain assumptions about Arabic dialect variation. In this chapter, I identify three common assumptions that were progressively adopted by the Arabic NLP community, in addition to a fourth one that was recently introduced.[1] The assumptions impact different aspects such as distinguishing between the varieties of DA (Asm. 1, Asm. 2, and Asm. 4), and dialectal samples curation (Asm. 3). However, their validity is neither backed by enough linguistic studies nor quantitatively assessed, making them anecdotal. While they were useful in achieving progress in tasks like Arabic Dialect Identification (ADI)[2], inaccuracies in these assumptions might hinder further progress. My analysis focuses on the text modality, but the findings could apply to the speech modality. It could also benefit linguists studying the Arabic varieties. I systematically examine the assumptions below, with the RQs relevant to each of them:

Asm. 1 A DA sentence is usually valid in only one regional dialect.

→ **(*RQ4*)** *How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?*

Asm. 2 Only short sentences can be valid in multiple dialects.

→ **(*RQ4.1*)** *Do most sentences valid in multiple dialects have a short length?*

[1] Limitations of the 4 assumptions are discussed qualitatively in the literature, but are ignored or perceived as minor.

[2] As of the 15th of December 2024, 618 papers on Semantic Scholar (Jones, 2015) match “Dialect Identification", out of which 173 (≈28%) match “Arabic Dialect Identification". However, ADI is still unsolved, as shown in chapter 6.

Asm. 3 Distinctive dialectal words (e.g., برشة *bršḥ* for Tunisian Arabic) can be curated to infer the dialect of sentences containing any of them.

→ **(*RQ4.2*)** *Are current ad-hoc lists of dialectal lexical terms distinctive enough to ensure that a sentence is valid in a dialect and not valid in other dialects if it contains a term of this dialect's list of lexical terms?*

Asm. 4 For a sentence valid in multiple dialects, speakers of these dialects consistently provide similar ratings of the sentence's level of dialectness.

→ **(*RQ3*)** *Do speakers of different dialects (varieties) share similar perceptions of a sentence's Dialect Level?*

The work presented herein was reported in the following paper:

- Amr Keleg, Sharon Goldwater, and Walid Magdy. 2025. *Revisiting Common Assumptions about Arabic Dialects in NLP.* In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria.

7.1 Background

In this section, I describe how the four assumptions were progressively adopted.

7.1.1 The Groupings of Arabic Dialects

Along the vast geographical area over which Arabic speakers are distributed, different varieties of DA are spoken. Varieties spoken within geographically proximate areas are commonly grouped into regional dialects. An example of such groupings is: the Levant (Lebanon, Jordan, Palestine, Syria), Nile Basin (Egypt, Sudan), Gulf (Saudi Arabia, Oman, Qatar, Bahrain, United Arab Emirates, Iraq), Gulf of Aden (Yemen, Djibouti, Somalia), and Maghreb (Morocco, Tunisia, Algeria, Mauritania, Libya).[3] Regional groupings recognize the within-region similarities while assuming minimal overlap between the regional varieties.

The last two chapters quantitatively showed that country-level ADI should be modeled as a multi-label classification task. However, they did not investigate how the task should be framed at the regional level. Early efforts in ADI used single-label classification to distinguish between a subset of the regional varieties, including MSA as an independent variety/class (Biadsy et al., 2009; Zaidan and Callison-Burch, 2011).

[3] A canonical grouping of the Arabic dialects does not exist (Habash, 2010; Abdul-Mageed et al., 2018).

This adoption of single-label classification implicitly accepts **Asm. 1** at the regional level; i.e., that sentences are usually only valid in one regional dialect. (confer §2.5.1 for more details). Although current ADI models are generally built at the country level, any overlap that exists at the regional level will still persist when these regions are divided into countries. Consider a sentence valid in two regions Reg_A and Reg_B. The same sentence should be valid in at least one country of Reg_A and one country of Reg_B.

Sentence Length and ADI Most ADI datasets use sentence-like units (e.g., tweets). Sentence-level classification models would usually struggle with short sentences. Moreover, it is commonly believed (**Asm. 2**) that most multi-label samples are very short. These two reasons might explain why ADI has continued to be modeled as a single-label classification task.

7.1.2 Dialectal Lexical Cues

Although dialects differ at many linguistic levels (phonological, lexical, syntactic), one of the easiest types of cues to identify in text is lexical cues (Kaye and Rosenhouse, 1997). These cues are *distinctive* of a particular dialect if they are not shared with other dialects. Some papers provide qualitative examples of these cues like هطعش *hTʕš*[4] - eleven) for Yemeni (Al-Shargi et al., 2016) and برشة (*bršħ* - a lot) for Tunisian (McNeil, 2018; Abdelali et al., 2021).

Distinctive cues have been widely used to build DA datasets. To this end, ad-hoc lists of lexical cues were compiled to collect dialectal samples from websites or social media platforms. These lists were either directly used (Al-Sabbagh and Girju, 2012; Alshutayri, 2017; Alshargi et al., 2019), or first validated by speakers of different dialects to ensure their distinctiveness (Almeman and Lee, 2013; Zaghouani and Charfi, 2018; Alsarsour et al., 2018; Mubarak, 2018).

It is acknowledged that the diversity of the curated samples is limited by the lists of cues (Abdul-Mageed et al., 2020b). However, the precision and distinctiveness of these cues are assumed to be high without quantitatively measuring them (**Asm. 3**), which I revisit in this chapter.

[4]Transliteration follows HSB scheme (Habash et al., 2007).

7.1.3 Differences in ALDi Perceptions

In chapter 3 and chapter 6, a sentence's ALDi ratings provided by annotators speaking different dialects are averaged together to compute a gold-standard score for this sentence. *AOC-ALDi* relied on randomly routing each sentence to speakers of different dialects. The guidelines used in §6.1.1 tried to mitigate the expected variation by ensuring the annotators provide ALDi ratings to sentences only if they are valid in their country-level dialects. However, this still assumes that sentences valid in multiple dialects are rated similarly by speakers of these different dialects, overlooking the impact of the annotator's native dialect on the provided ALDi ratings (**Asm. 4**)

7.2 Data

For my analysis, I release an extended version of the NADI 2024 dataset (cf. chapter 6), that I call the ***MLADI*** (Multi-label ADI) dataset.[5] The original dataset has 1,120 tweets, of which only 70 were automatically identified as MSA and 1,050 as DA. The DA samples' geolocations are uniformly distributed across the 14 most populated Arab countries, excluding Somalia, for which data is not sufficiently abundant. 27 annotators were recruited from 9 Arab countries (3 each): Algeria, Morocco, Tunisia, Egypt, Sudan, Palestine, Syria, Iraq, and Yemen. For each sample in the dataset, the annotators **(a)** identified if a speaker of one of their country-level dialects could have authored the tweet. If an annotator answered (a) as yes, then the sentence is also **(b)** rated for its ALDi as MSA (L0), Colloquial-influenced MSA (L1), Normal Colloquial (L2), or Informal (or Vulgar) Colloquial (L3).

To extend the dataset's labels, I recruited three annotators from Saudi Arabia, using the same annotation guidelines introduced in section 6.1. This enhances the dataset's coverage of the various Arab dialects, particularly Gulf Arabic. I also completed the annotation collection from the three Jordanian annotators. Figure 7.1 shows the cities/provinces from which the annotators originate.

The Interannotator Agreement scores for the two new dialects (Jordanian and Saudi) are reported in Table 7.1. For the validity labels of each country, I compute the chance-corrected Fleiss' Kappa (κ) score, finding adequate agreement between the annotators of both countries. For the ALDi ratings, I use Krippendorff's Alpha interval method (α) to compare the numeric values of the ratings for each country's valid samples,

[5] An accompanying ADI leaderboard is released at: `https://huggingface.co/spaces/AMR-KELEG/MLADI`

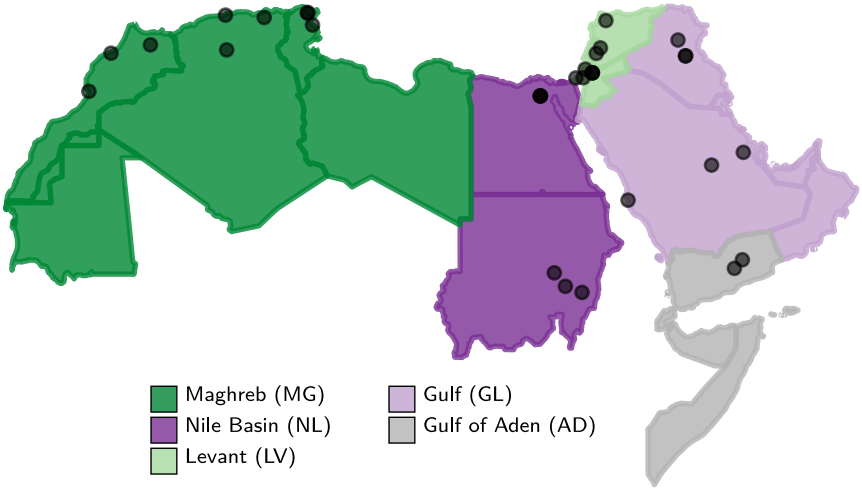


Figure 7.1: A map of the Arab world. The black dots indicate the provinces/cities from which the annotators originate. Regional dialects (Maghreb, Nile Basin, Levant, Gulf, Gulf of Aden) are encoded as different colors according to the groupings of Baimukan et al. (2022).

penalizing disagreements differently according to their assigned values. The range of the *alpha* scores is -1 to 1, with zero indicating chance agreement. Hence, 0.62 and 0.65 signify that the annotators' agreement is substantially better than random, despite the subjectivity of the task. These IAA scores are similar to the ones reported in Table 6.5 for the other dialects of the NADI 2024 dataset.

Country	**Validity labels**			**ALDi ratings**
	Fleiss κ	**N valid**	**N ¬valid**	**Krip. α**
Jordan	0.56	617 (455)	503 (367)	0.62
Saudi Arabia	0.62	476 (328)	644 (490)	0.65

Table 7.1: The Interannotator agreement scores for the validity labels and ALDi ratings, Fleiss' Kappa (κ) for Validity labels and Krippendorff's Alpha - interval method (α) for ALDi ratings. *N valid* and *N ¬valid* represent the number of samples whose majority vote labels are *valid* and *not valid*, respectively, with the number of sentences with complete agreement reported (between brackets).

Following the same aggregation methods in §6.1.4, I use majority voting to identify the validity of each tweet in each of the 11 country-level dialects, and for ALDi, I transform the ratings from discrete levels (L0, L1, L2, L3) into numeric values $(0, \frac{1}{3}, \frac{2}{3}, 1)$. A sentence's ratings, for the dialects in which the sentence is valid (according to the majority voting), are averaged to estimate a dialect-agnostic ALDi score.

7.3 Analysis

In this section, I investigate each of the four aforementioned assumptions, using 978 out of the 1,050 DA samples, after discarding 72 samples that are not labeled as valid in any of the 11 considered country-level dialects.[6]

7.3.1 Challenging Asm. 1 - Arabic Dialects Rarely Overlap

At least 28 different ADI datasets assign a single regional/country-level dialect to each sentence (cf. Table 5.2 in chapter 5). Single-label classification was shown not to be suitable for country-level ADI both qualitatively (Kchaou et al., 2019; Touileb, 2020; Bayrak and Issifu, 2022; Khered et al., 2022) and quantitatively (chapter 5; chapter 6;Olsen et al., 2023). However, single-label classification might still be thought of as suitable for ADI on the level of regional dialects, under the assumption that they rarely overlap.

Method Using the regional grouping proposed in (Baimukan et al., 2022), I form 5 regional-level validity labels from the 11 country-level labels as follows: **1) Nile Basin (NL):** Egypt, Sudan, **2) Gulf (GL):** Iraq, Saudi Arabia, **3) Gulf of Aden (AD):** Yemen, **4) Maghreb (MG):** Tunisia, Algeria, Morocco, and **5) Levant (LV):** Jordan, Palestine, Syria. A sentence is valid in a regional dialect if it is valid in at least one of the considered region's countries. Afterward, I count the number of regional dialects in which each sentence is valid.

Results A majority 56% of sentences (544 in total) are valid in multiple regional dialects, as shown in Figure 7.2. This large cross-regional overlap exists despite the fact that the MSA samples were discarded.[7] Notably, 116 of these DA samples (a non-negligible $\sim$ 12%) are valid in all regional dialects.

[6]The code is released at: `https://github.com/AMR-KELEG/MLADI-assumptions-revisiting`

[7]Refer to §F of the Appendix for a similar analysis on the country-level.

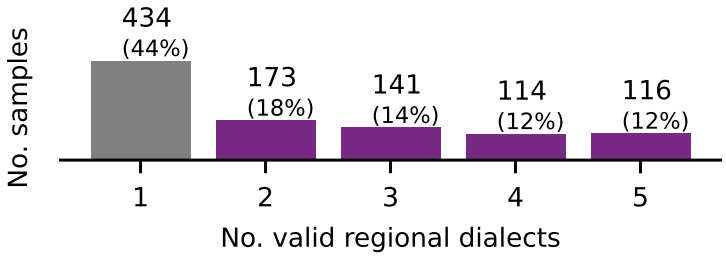


Figure 7.2: The histogram of the number of valid dialects on the regional level. Only 44% of the DA samples are confined to single-region dialects.

Further Analysis Unlike the other dialects, the Gulf of Aden (represented by Yemen) has only 11 single-region samples as per Figure 7.3. Hence, it might not be prominently different from some of the subdialects spoken in other regions, challenging the recognition of *Gulf of Aden* as a regional dialect (Habash, 2010; Abdul-Mageed et al., 2018).

The Levant, Gulf, and Gulf of Aden have a substantial number of samples shared with other regional dialects, with the Levantine dialect sharing more than the other two dialects. Looking at the distribution of the multi-region samples in Figure 7.4, a large number of the 2-region samples are between pairs of these three regions (e.g., 46 valid in GL and LV, 20 valid in AD and GL) and a majority of 61 samples of the 3-region ones are valid in these regions. Additionally, LV has a substantial number of 38 samples shared with NL, 15 shared with MG, and 18 shared with both NL and MG. This explains how LV shares more samples with other dialects than GF and AD.

For the remaining two dialects (NL and MG), both share fewer samples with other dialects, with NL sharing more samples than MG. 62 samples (a majority of the 4-region samples) are valid in all regions but MG. This is a sign of the dichotomy between the Eastern dialects of Arabic spoken in the Maghreb and the other dialects spoken in the West of the Arab world (Kaye and Rosenhouse, 1997). Still, MG shares more with other dialects than previously assumed.

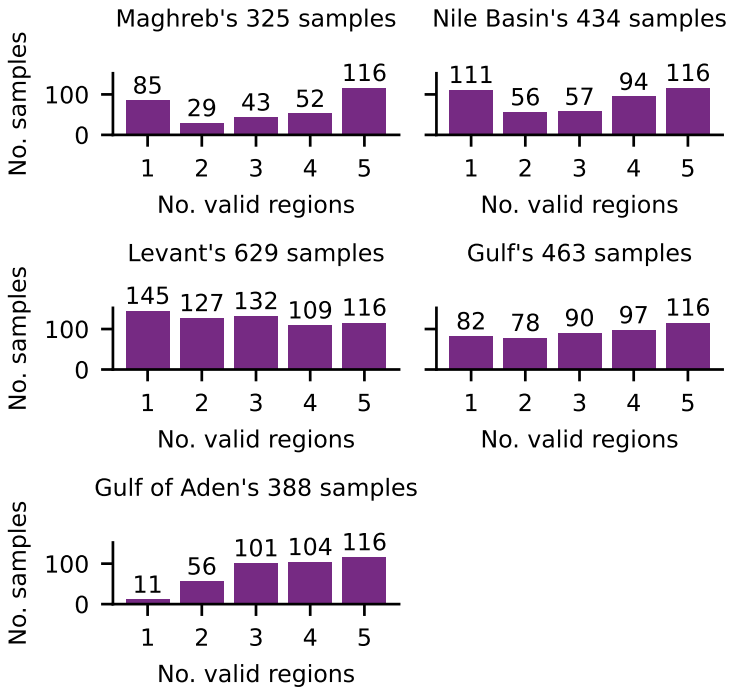


Figure 7.3: The total number of valid regional dialects for each region's valid samples. **Note:** The regions' samples are not mutually exclusive (e.g., the same 116 samples valid in the five regions are in all distributions).

Estimating the Maximal Accuracy for MLADI's regional labels When framing a multi-label task as a single-label one, there is an expected maximal accuracy that an oracle model can achieve (check §5.3). For a sample with multiple valid labels, the gold-standard label and the prediction of the oracle model will both be randomly selected from the sample's set of valid labels. For the model's prediction to be considered correct, both the randomly sampled gold standard label and the model's prediction must match. I introduced Equation 5.2 in chapter 5 for estimating the expected maximal accuracy given the distribution of the number of labels in which a sentence is valid. Applying the formula to the regional-level labels of the 978 DA samples I used for our analysis, we get an expected maximal accuracy of 63.06% as per Equation 7.1. Such a low accuracy upper bound provides more evidence for modeling the task as a multi-label classification one.

$$\mathbf{E}[\mathbf{Accuracy_{max}}(\mathbf{MLADI_{regional}})] = Perc_1 + \sum_{n=2}^{n=N_{dialects}} \frac{Perc_n}{n} = 44 + \frac{18}{2} + \frac{14}{3} + \frac{12}{4} + \frac{12}{5} \approx 63.06\% \tag{7.1}$$

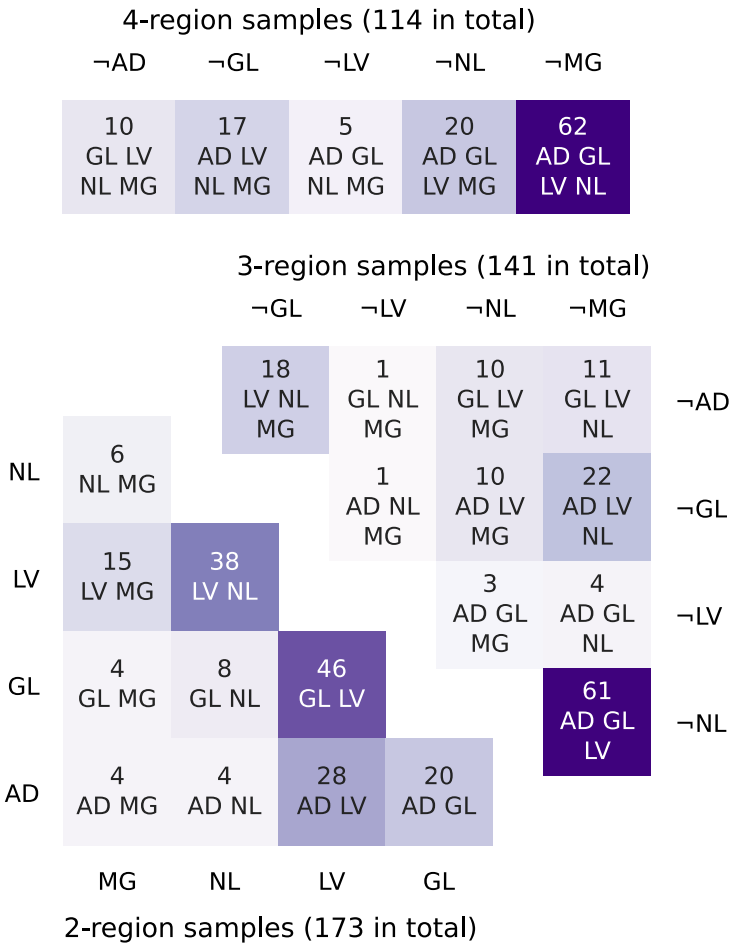


Figure 7.4: The distribution of the 2-region, 3-region, and 4-region samples across the different combinations. Each combination has its regions indicated in its respective cell. **Note:** GL/¬GL means valid/not valid in Gulf.

Implications Substantial overlap exists between the regional dialects, which contradicts the general perception that they are distinguishable from each other. As previously mentioned, this overlap will still exist when the regions are split into countries. Hence, ADI is a multi-label task on the regional and country levels.

Classifying *Gulf of Aden* as a distinct regional variety requires reevaluation, given the limited number of samples only valid in this region. To this end, dialectal categorizations that are not based on the spurious country borders could be considered.[8]

[8]Glottolog (`glottolog.org / resource / languoid / id / arab1395`) and Ethnologue (`ethnologue.com/language/ara/`) recognize 37 and 28 Arabic dialects, respectively. Some of these dialects are not confined by the borders between the countries.

7.3.2 Challenging Asm. 2 - Only Short Sentences' Dialects are Ambiguous

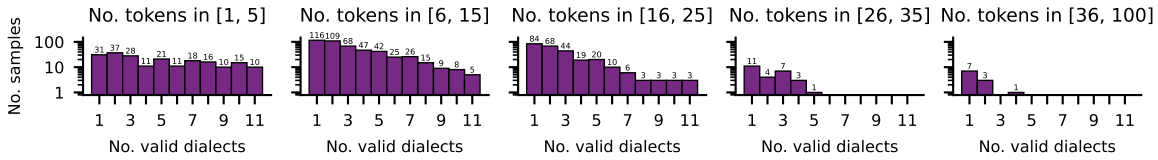
In the context of ADI, sentence length is discussed from two points of view (POVs). *POV #1* explicitly mentions that the dialect of extremely short speech segments/text sentences can be ambiguous. Hence, it is infeasible for humans, and consequently machines, to assign a single dialect to these segments (Alorifi, 2008) and sentences (El-Haj et al., 2018; Alsarsour et al., 2018; Abu Kwaik and Saad, 2019; Althobaiti, 2022). *POV #2* empirically finds that the longer the segment/sentence gets, the higher the performance of a single-label ADI system is, for speech (Biadsy et al., 2009; Shon et al., 2020) and text (Zaidan and Callison-Burch, 2014; Salameh et al., 2018; AlKhamissi et al., 2021; Abdelali et al., 2021; Bayrak and Issifu, 2022). This can be attributed to a decline in dialect ambiguity as sentences get longer.

Method I examine the assumption by computing Spearman's correlation between the sentence length (as the number of tokens) and the number of valid dialects on the country level. Additionally, I study the histograms of the number of valid dialects for five different ranges of sentence lengths.

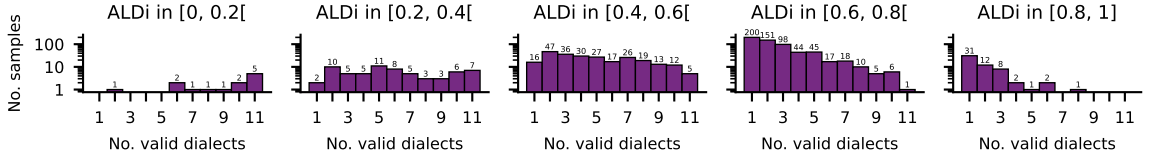
Results According to Figure 7.5a, the majority of trivially short sentences are valid in multiple dialects as per *POV #1*. However, *POV #1* overlooks the large number of moderately long sentences (16-25 tokens) that are also valid in multiple dialects. Additionally, despite long sentences being valid in a smaller number of dialects, confirming *POV #2*, there is only a weak negative Spearman's correlation coefficient (-0.28) between the sentence length and its number of valid dialects.

Further Analysis On replicating the analysis by replacing the sentence length with the ALDi score, a stronger negative correlation (-0.52) is realized.[9] Figure 7.5b also indicates that sentences of ALDi scores < 0.2 are generally valid in most of the dialects. Samples with ALDi scores $\in [0.2, 0.4[$ seem to be evenly probable across the different number of validity labels. The distribution then shifts to be more and more right-skewed for the subsequent ranges of ALDi scores.

[9] A coefficient of -0.45 is realized when replacing the aggregated manually-assigned ALDi scores with ones automatically estimated using the Sentence-ALDi model (chapter 3).



(a) Sentence length (measured as the number of tokens). **Note:** $\rho(\textit{Sentence Length}, \#\textit{ valid dialects}) = -0.28$



(b) ALDi scores (averaged across all ratings). **Note:** $\rho(\textit{ALDi}, \#\textit{ valid dialects}) = -0.52$

Figure 7.5: The distribution of the sentences (log scale) and the number of valid country-level dialects according to different ranges of sentence length **(a)** and ALDi scores **(b)**. **Note:** Since the MSA samples were automatically discarded from my analysis dataset, there are very few samples with low ALDi scores ($\in [0, 0.2]$). However, the histogram of this bin is expected to be left-skewed (i.e., MSA samples are expected to be valid in all dialects).

Implications Previous assumptions about sentence length are either incomplete (*POV #1*) or not sufficiently accurate (*POV #2*). Moreover, a sentence's ALDi score correlates moderately with the number of dialects in which it is valid, making it a better predictor than sentence length. As a proxy of a sentence's number of valid dialects, ALDi could guide the predictions of a multi-label ADI system.

7.3.3 Challenging Asm. 3 - Dialects' Distinctive Lexical Cues

Method For each of DART's (Alsarsour et al., 2018) and DIAL2MSA's (Mubarak, 2018) lists of regional-level distinctive cues, I identify sentences of my dataset that match at least one of the lexical cues.[10] I normalize the sentences and lists of cues to handle common typos of the same characters (e.g., ة is normalized to ه and أ, آ, إ are normalized to ا) (Kholy and Habash, 2012; Darwish and Magdy, 2014). Exact matching is then used between the lexical cues and the whitespace tokenized sentences' tokens.

[10] I could not get access to the lists of Almeman and Lee (2013); Zaghouani and Charfi (2018); Alshargi et al. (2019).

For each dialect, I report the number of samples matching at least one of its distinctive cues (M). Then, I count the number of matching samples manually annotated as valid in this dialect (M_{Val}), and the number of matching samples that are only (i.e., exclusively) valid in this dialect (M_{Exc}). Precision (P), Distinctiveness (D), and Recall (R) of each list are computed as $P = \frac{M_{\text{Val}}}{M}$, $D = \frac{M_{\text{Exc}}}{M}$, and $R = \frac{M_{\text{Val}}}{N_{\text{Val}}}$; where ($N_{Val}$) is the total number of samples valid in the considered dialect.

Adhering to the regional groupings used in both lists, I aggregate the 11 country-level validity labels into the following regions: **1) Egypt**, **2) Iraq**, **3) Gulf:** Saudi Arabia, **4) Maghreb:** Algeria, Morocco, Tunisia, **5) Levant:** Jordan, Palestine, Syria. Sudan and Yemen were ignored in both lists, so I considered them as **6) Others**.

Results Table 7.2 shows the results. The extremely low range of recall values for both manually validated lists confirms that relying on these lists of cues limits the number of matching samples. Conversely, the range of the precision scores is generally high (yet not perfect), except for the cues of Gulf Arabic. The Egyptian Arabic cues have a low precision score (0.6) for DART and extremely low distinctiveness values (0.35 and 0.38) for both lists. The samples' validity in the Maghreb, Levant, and Gulf regions is only defined by the subset of the region's countries from which I could recruit annotators. Hence, the precision scores for these regions might improve after collecting annotations for more country-level dialects. However, the non-perfect Distinctiveness scores indicate that some cues of these regions are used in other regional dialects, even when the cues were manually validated for their distinctiveness by the lists' creators.

Qualitative Analysis On manually inspecting the matching samples, I found that DART's three matching cues of Gulf Arabic (شنو *šnw*, علامك *ʿlAmk*, مواعين *mwAʿyn*) are indeed dialectal terms that are valid in other regional dialects, hence are not indicative of Gulf Arabic. Additionally, other terms are false friends, having different meanings in MSA and DA varieties, and are not distinctive of a specific dialect in the absence of context. For instance, the terms (ماشي *mAšy* and حد *Hd*) have the meanings *okay* and *someone* in Egyptian Arabic. However, they have different meanings in MSA (*walking* and *limit*). The MSA sense of these terms could be used in the context of other dialects, as demonstrated in the examples below, which both use the term حد *Hd* (underlined in the examples).

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	60	36	21	287	.60	.35	.13	271	28
IRQ	7	6	6	204	.86	.86	.03	120	7
MGH	21	16	14	325	.76	.67	.05	273	13
LEV	32	29	25	629	.91	.78	.05	240	11
GLF	9	0	0	407	.00	.00	.00	200	3

(a) DART's five regional lists.

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	53	43	20	287	.81	.38	.15	28	19
MGH	45	36	31	325	.80	.69	.11	60	26
LEV	38	34	34	629	.89	.89	.05	31	11
GLF	0	-	-	407	-	-	.00	9	0

(b) DIAL2MSA's four regional lists.

Table 7.2: The Precision (*P*), Distinctiveness (*D*), and Recall (*R*) of each region's cues. **Note:** For each region's list, I report the number of samples of my dataset matching any of the cues (*M*) of which valid (M_{Val}) and of which exclusively valid (M_{Exc}), in addition to the total number of valid samples (N_{Val}). The last two columns represent the total number of regional cues (**C**) and the number of cues that match any of the samples ($\mathbf{C_{Mat}}$).

Example (1) uses the term حد with its Egyptian meaning (*someone*) and is labeled as valid in Egyptian, whereas (2) uses the term with its MSA meaning (*limit*) and is labeled as valid in Algerian and Tunisian. Therefore, the term حد *Hd* can not be considered a valid cue to Egyptian Arabic, as assumed in DART.

(1) دا الفرق بين حد اهله عرفوا يربوه وحد تاني منفعش فيه التربيه.

'This is the difference between a well-mannered and a bad-mannered person.'

(2) الي حد الان مازال حظوظ تونس كبيره هاك تراو الفرق الكبار.

'So far, Tunisia still has great chances, this is how big teams are.'

Impact of Using a Geolocation-based Filtration Step The TWT15DA is an ADI dataset built by iteratively augmenting lists of lexical cues of 15 country-level dialects using geolocated tweets having any of these cues, then streaming more geolocated tweets using the augmented lists (Althobaiti, 2022). For each country, the new cues to be added are non-MSA unigrams **(a)** in the tweets geolocated to this country, that **(b)** have high PMI values based on the following equation: $PMI(Unigram,\ Country) = log(\frac{P(Unigram,\ Country)}{P(Unigram)*P(Country)})$; where the probabilities are computed using maximum likelihood estimation. Therefore, the same unigram could have PMI scores for multiple countries (e.g., كيفاش *kyfAš* in Algerian, Moroccan, and Tunisian Arabic lists with PMI scores of 2.07, 1.55, 1.19). Hence, these cues are not necessarily distinctive of a single country-level dialect. However, the author defines the *cues* as "words used in one or more Arabic dialects but never used in MSA, thereby distinguishing Arabic dialects from MSA".

Country	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
Morocco	52	22	5	163	.42	.10	.13	410	45
Algeria	41	23	2	265	.56	.05	.09	421	38
Tunisia	62	19	1	123	.31	.02	.15	407	48
Egypt	33	23	18	287	.70	.55	.08	172	35
Jordan	51	30	1	547	.59	.02	.05	180	37
Syria	50	28	9	406	.56	.18	.07	94	28
Iraq	21	13	12	204	.62	.57	.06	179	18
Yemen	8	5	2	388	.62	.25	.01	137	8
Saudi	43	20	6	407	.47	.14	.05	145	26

Table 7.3: Lexical cues of the TWTDA15 datasets. **Note (1) :** For each region's list, I report the number of samples of my dataset matching any of the cues (*M*) of which valid (M_{Val}) and of which exclusively valid (M_{Exc}), in addition to the total number of valid samples (N_{Val}). The last two columns represent the total number of regional cues (**C**) and the number of cues that match any of the samples ($\mathbf{C_{Mat}}$). **Note (2):** The table lists the nine countries that are common between the labels of my dataset, and the lists of TWT15DA, which did not include *Palestine* and *Yemen*.

I replicate the analysis for the TWT15DA dataset, and report the precision, recall, and distinctiveness scores in Table 7.3. Notably, the lists have a low range of precision scores $[0.31, 0.70]$, and an even lower range of distinctiveness scores $[0.02, 0.57]$.

For the TWT15DA dataset, each sample should have at least a cue for one of the dialects. However, the assigned label is based on the sample's geolocation, rather than the dialects associated with the cues. Hence, to assign a sample to a country-level dialect, the sample should **(a)** have a lexical cue of this dialect and **(b)** be geolocated to this country. To simulate this two-step method for each country's/region's list, I replicate the same method above, but then only consider the matching samples that are geolocated to the considered country/region. The results of applying this post-processing step for the three lists of cues (DART, DIAL2MSA, and TWT15DA) are reported in Table 7.4. The effectiveness of this step is better understood by contrasting the results in Table 7.2 and Table 7.3 to those in Table 7.4.

The range of the precision significantly improves to values > 0.9 for the three lists, except for the lists of Tunisia and Jordan in TWT15DA. The distinctiveness scores also improve, yet to much lower ranges compared to the precision. This hints that filtering out the samples that match any of a region's cues, yet are not geolocated to this region, minimizes the impact of matching false friends of these cues, which are intuitively expected to be in samples geolocated to other regions.

Unsurprisingly, limiting the samples to ones geolocated to each list's region causes a decrease in the recall values, as all the samples valid in this region's dialect that are not geolocated to the region are pre-filtered. Another drawback of this geolocation-based step is that the samples' geolocations are not always available.

Implications More rigor is needed in building lists of distinctive dialectal words, especially when the curated sentences need to be surely valid in a specific dialect and/or exclusively valid in this dialect. Using a second validation step (e.g., information about the geolocation of the sentence's author) could increase the precision of the dialects assigned based on the cues' associated dialects. However, this does not ensure distinctiveness and further decreases the recall.

7.3.4 Challenging Asm. 4 - ALDi Perceptions across Dialects

Inspired by earlier work (Zaidan and Callison-Burch, 2011), I introduced the idea of ALDi estimation as an essential task. To this end, I employed two datasets providing pairs of sentences with their corresponding aggregated ALDi scores: *AOC-ALDi* (chapter 3) and the newly introduced *MLADI* dataset. For the former, three annotations per sentence were sought by randomly assigning the sentences to speakers of different

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	20	20	13	287	1.0	.65	.07	271	10
IRQ	6	6	6	204	1.0	1.0	.03	120	7
MGH	15	15	14	325	1.0	.93	.05	273	11
LEV	24	22	20	629	.92	.83	.03	240	8
GLF	0	0	0	407	-	-	.00	200	0

(a) DART's 5 regional lists.

Region	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
EGY	25	25	15	287	1.0	.6	.09	28	10
MGH	34	32	29	325	.94	.85	.10	60	22
LEV	36	33	33	629	.92	.92	.05	31	11
GLF	0	0	0	407	-	-	.00	9	0

(b) DIAL2MSA's 4 regional lists.

Country	M	M_{Val}	M_{Exc}	N_{Val}	P	D	R	C	C_{Mat}
Morocco	15	14	5	163	.93	.33	.09	410	22
Algeria	13	13	2	265	1.0	.15	.05	421	18
Tunisia	12	9	1	123	.75	.08	.07	407	15
Egypt	13	13	11	287	1.0	.85	.05	172	14
Jordan	15	12	1	547	.80	.07	.02	180	14
Syria	12	11	5	406	.92	.42	.03	94	12
Iraq	11	11	11	204	1.0	1.0	.05	179	13
Yemen	4	4	2	388	1.0	.50	.01	137	6
Saudi	9	9	4	407	1.0	.44	.02	145	7

(c) TWT15DA's 9 country-level lists.

Table 7.4: The Precision (*P*), Distinctiveness (*D*), and Recall (*R*) of each region's/country's cues, when only the matching samples that are geolocated to the region/country are considered. **Note:** For each region's list, I report the number of samples geolocated to this region, matching any of its cues (*M*) of which valid (M_{Val}) and of which exclusively valid (M_{Exc}). The total number of samples valid in this region (N_{Val}) is reported irrespective of their geolocations. The last two columns represent the total number of regional cues (**C**) and the number of cues that match any of the samples ($\mathbf{C_{Mat}}$).

dialects (Zaidan and Callison-Burch, 2011). For the latter, 33 annotators rated the ALDi of each sentence only when it was valid in their country-level dialect. Both datasets used the mean of a sentence's ALDi ratings as its gold-standard ALDi score. The implicit assumption is that ALDi scores do not depend on the annotator's native dialect; however, this has not been empirically validated. I have shown (§7.3.2) that even sentences with moderate ALDi scores can be valid in multiple dialects. For these sentences, systematic differences in the ALDi ratings could exist around each sentence's mean ALDi score for annotators speaking different dialects.

Method I compute the Mean Difference (MD) of country-level ALDi scores for each pair of countries. MD is computed for a pair of countries r and c, with N_{rc} sentences valid in both, as

$$\mathrm{MD}(r,c) = \frac{1}{N_{rc}} \sum_{i=1}^{N_{rc}} (\mathrm{ALDi}_r[i] - \mathrm{ALDi}_c[i]),$$

where $\mathrm{ALDi}_r[i]$ and $\mathrm{ALDi}_c[i]$ are the averages of sentence i's ALDi ratings provided by the annotators of r and c respectively.

Results Figure 7.6 summarizes the results. The top three (orangish) rows indicate that when sentences are valid in one of the Maghreb's countries and another non-Maghrebi country, the annotators from the Maghrebi country rate these sentences to be less colloquial than the non-Maghrebi ones. The difference (e.g., MD(Morocco, Saudi) = -0.29) can be close to $\frac{1}{3}$, which is the difference between two consecutive levels of ordinal ALDi ratings, which are mapped to the following respective numeric values $(0, \frac{1}{3}, \frac{2}{3}, 1)$. A similar pattern holds true for Iraq to a lesser extent. Conversely, Saudi annotators assign higher ALDi scores to sentences common with other dialects. Many of the country-level differences are statistically significant, with Standard Errors < 0.035. However, these differences could arise simply because the annotators differ randomly in their mean scores, independent of dialect. So we might see an apparent difference between country groups if I happened to get annotators with higher means in some countries than in other countries. Due to having only three annotators per country, it is not possible to conclusively test for an effect of dialect (separate from annotator) at the country level, although the consistent trends in the visualization are suggestive. Instead, I test for regional-level differences between annotators, as described below. If additional annotations from each country are obtained in the future, a similar test could be used at the country level.

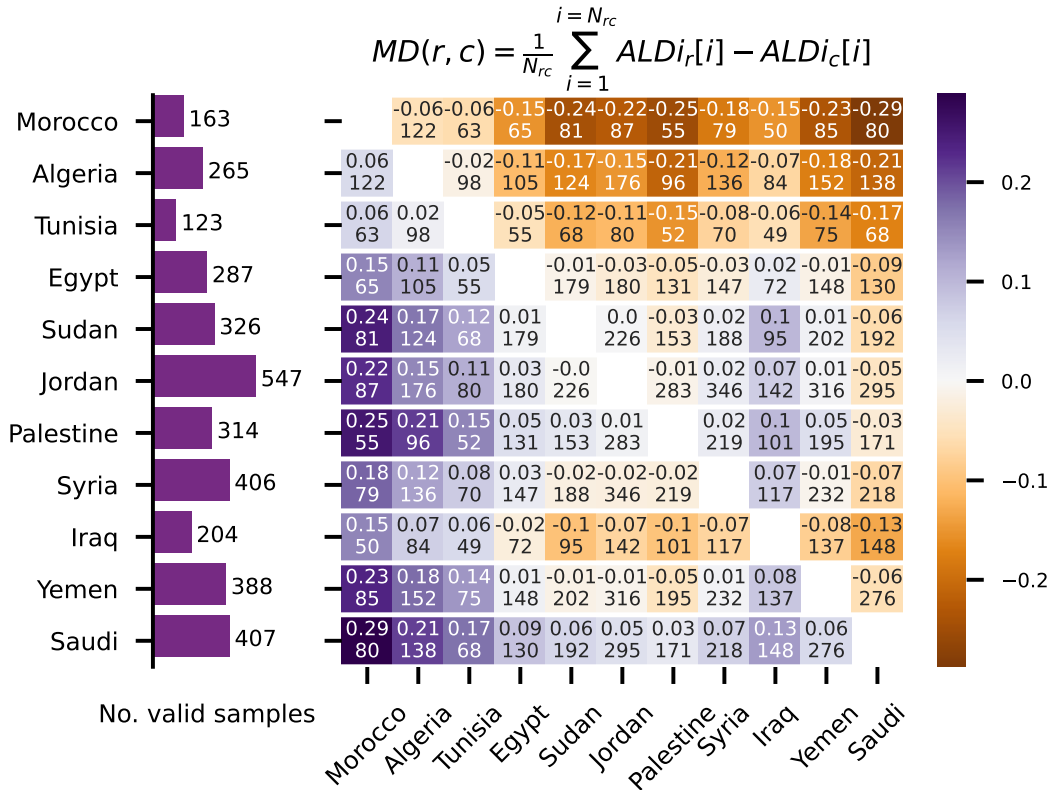


Figure 7.6: **(Left)** The number of valid samples per country (with countries ordered such that same-region ones are consecutive). **(Right)** Mean difference (MD) of row country's (r) and column country's (c) ALDi scores, for the N_{rc} sentences valid in both (N_{rc} is shown as the bottom number in each cell).

Statistical Analysis I use a one-sided permutation test to assess whether the differences between two groups of annotators (G_A, G_B)—of sizes $|G_A|$ and $|G_B|$ respectively—can be attributed to the groups' dialects. First, I compute the MD score between the observed groups' mean ALDi scores (MD_{obs}), for the N_{AB} sentences valid in both groups. The null hypothesis stipulates that the observed mean difference is attributed to annotator variation, independent of their native dialects. A large number of pairs of groups $\{(A', B')\}$ with sizes $|G_A|$, $|G_B|$ are sampled (50k in our case). The pairs of groups (A', B') are formed by random shuffling and distributing all the annotators into two groups. MD scores for each pair are computed for the same N_{AB} sentences.[11] The p-value is the percentage of the shufflings with MDs $\leq$ the observed grouping's mean difference (MD_{obs}).

[11] In some permutations, I discard the small proportion of sentences that have no ALDi ratings for one of the groups.

I consider the annotators of each region as a group, merging Gulf and Gulf of Aden into one region based on the findings of §7.3.1. Accordingly, I find significant MDs of -0.09, -0.13, -0.14 between the ALDi scores averaged across the annotators of Maghreb against those of Nile Basin, Levant, and Gulf/Gulf of Aden, with p-values of 0.007, 0.00002, and 0.0002, respectively. Similarly, Nile Basin's annotators provide significantly lower ALDi scores than Levantine annotators, with MD of -0.05 (p-value = 0.04). Differences between other pairs are not statistically significant.

Discussion There is a general impression that the Arabic dialects are not equally distant from MSA, with some researchers claiming certain dialects—e.g., Gulf Arabic (Zaidan and Callison-Burch, 2014) and Palestinian Arabic (Kwaik et al., 2018)—are closer to MSA than others, which could explain the MDs I found for samples shared between different countries/regions.

Implications Further analysis is required before taking these MDs as an objective measure of a variety's divergence from MSA. Figure 7.3 indicates that all regions—except *Gulf of Aden*—have many samples not shared with other regions. Single-region samples could still be highly divergent from MSA. Moreover, people's perception of dialectness is influenced by how they use MSA terms colloquially. For example, both خمر *xmr* and خمرة *xmrħ* are valid MSA terms for *wine*. The Holy Qur'an mentions the former, while the latter is more colloquially used in Egypt. Hence, Egyptians might link the first to CA/MSA, and the latter to DA. Consider some MSA lexical items that are shared with dialect D_A but not with dialect D_B. Sentences with these items could be rated as more colloquial by speakers of D_A than D_B. Lastly, sentences valid in multiple varieties could share the same surface form but have different semantics in each variety.

7.4 Further Implications in NLP

The recent improvements to how the varieties of Arabic are computationally modeled—introduced in chapter 3, chapter 5, and chapter 6—are being used in multiple applications, such as better routing of samples to annotators (as shown in chapter 4), evaluating the LLMs' dialectal capabilities (Robinson et al., 2025), and building better recommendation systems (Alshabanah and Annavaram, 2025). Hence, validating widely held assumptions about Arabic could lead to further progress in automatic ADI and many other tasks/applications.

For example, Arabic NLP researchers used manually curated lists of words/phrases in various applications like compiling dialect-specific pretraining data (Gaanoun et al., 2024), creating datasets for sentiment analysis (Refaee and Rieser, 2014), and offensive text classification (Chowdhury et al., 2020). Therefore, my finding—that some terms share the same orthographic form but have different semantic meanings/senses in various varieties of Arabic—has implications for building datasets for tasks beyond ADI.

Moreover, parallels of the first three assumptions exist beyond Arabic. For example, the overlap between different dialects of the same language has already been noted for other languages such as English, French, and Spanish (Bernier-Colborne et al., 2023; Zampieri et al., 2024; Lopetegui et al., 2025). Our findings argue for modeling dialect identification as a multi-label classification task, even on macro-regional levels. In addition, sentence length has been discussed as an important predictor of language identification models' performance (Baldwin and Lui, 2010), especially for closely-related languages and dialects (Tiedemann and Ljubešić, 2012; Blodgett and O'Connor, 2017; Kanjirangat et al., 2022). I show that the conscious *Dialect Level* choice that Arabic speakers make—operationalized as ALDi—is a better predictor of the number of dialects in which a sentence is valid than its length. Speakers of other languages make similar conscious decisions about how much they adhere or diverge from the standard variety of their language (e.g., Shoemark et al., 2017). For these languages, modeling the sentences' divergence from the language's standard variety, as ordinal/quantitative variables, could also provide better predictors of a sentence's validity in multiple dialects than the sentence's length.

7.5 Summary and Moving Forward

I identified four common assumptions regarding Arabic dialects, and systematically studied them by extending the annotations of a previous dataset to cover more country-level dialects. My analysis shows that these assumptions oversimplify some details that, in turn, impact how tasks are framed, datasets are created, and models are trained.

In particular, my main findings and recommendations address different research questions of this thesis, and can be summarized as follows:

RQ4 ***How prevalent are the sentences valid in multiple dialects, on the country and macro-regional levels?*** Arabic dialects overlap considerably at both the country and regional levels, so ADI should be modeled as a multi-label task at both levels.

RQ4.1 ***Do most sentences valid in multiple dialects have a short length?*** ALDi scores (but not sentence length) provide a good proxy of a sentence's validity in multiple dialects, which could be used to inform annotation and modeling decisions. Nevertheless, researchers should be aware that speakers of different dialects may systematically differ in their ALDi annotations of the same sentences.

RQ4.2 ***Are current ad-hoc lists of dialectal lexical terms distinctive enough to ensure that a sentence is valid in a dialect and not valid in other dialects if it contains a term of this dialect's list of lexical terms?*** Existing lists of supposedly distinctive lexical cues are less distinctive than previously thought. More rigorous validation is needed for such lists in the future.

RQ3 ***Do speakers of different dialects (varieties) share similar perceptions of a sentence's Dialect Level?*** Future work should study whether sentences with diverging ratings by speakers of different dialects have different semantic meanings in these dialects.

Chapter 8

Conclusion and Future Work

This thesis investigated two main hypotheses—forming two complementary themes—that would enable the improvement of the representation of Arabic varieties in NLP models. I will summarize the thesis's contributions according to these two themes, related to intraspeaker variation and interspeaker variation, respectively.

8.1 Summary of Contributions

Hypothesis 1 states that '*Written Arabic sentences of a specific dialect exist over a spectrum of Dialect Levels, with pure Standard Arabic and highly Colloquial Arabic as its extremes.*'

I proposed a new quantitative variable, the *Arabic Level of Dialectness (ALDi)*, to represent this spectrum on the sentence level. *ALDi* is defined as the **extent by which the sentence diverges from standard language**, and is operationalized as a continuous variable in [0, 1]. To assign an ALDi score for a sentence, manually provided ratings on an ordinal scale are mapped to numeric values. To aggregate a sentence's ALDi ratings into a single ALDi score, the algebraic mean of the ratings' corresponding numeric values is used, which better represents the interannotator variation than using majority voting to aggregate the ratings. After analyzing the ratings of dialectness of a large existing dataset of 127,835 online comments, I found adequate interannotator agreement (Krippendorff's α of 0.63).

To automatically estimate ALDi, I proposed fine-tuning a regression head on top of an Arabic BERT model, which proved to provide an effective method that can even generalize to different dialects. For intrinsic evaluation, the model achieved an RMSE of 0.18, demonstrating its ability to distinguish between the varying levels of the two

extreme values: 0 for Standard Arabic and 1 for highly colloquial Arabic. For extrinsic evaluation, the model generates an intuitive distribution of ALDi scores for two parallel corpora, comprising sentences in Standard Arabic and their counterparts in dialects, such as Moroccan and Tunisian, which are not well-represented in the model's training dataset.

Automatic ALDi estimation models are a powerful tool that could be of great use in many applications. Sociolinguistic studies could rely on using ALDi as a linguistic variable. I demonstrate its effectiveness in identifying the different styles that Arab presidents use in their speeches. This analysis could have only been done qualitatively before. Moreover, I refine previously proposed guidelines for annotating Arabic datasets, which recommended routing the samples to speakers of the samples' dialects. While appealing, this recommendation does not address the current lack of robust ADI systems that can automatically identify the dialects of each sample, and ignores the inability to extensively recruit speakers of some Arabic dialects. After analyzing the impact of ALDi on IAA for 15 different datasets, I conclude that high-ALDi samples should have a higher priority of being routed to speakers of these samples than low-ALDi samples. **Hypothesis 2** states that '*A substantial number of written sentences are valid in more than one variety of Arabic, especially when fine-grained geographical taxonomies of these varieties are used (e.g., country level or city/province level).*'

Despite expecting that some sentences would be valid in multiple dialects, the prevalence of these sentences in reality is substantially higher than expected. A preliminary investigation revealed that for about 66% of the errors of a SOTA single-label Arabic Dialect Identification model, the predictions are also valid, indicating the invalidity of the *Disjointedness* assumption, where a sentence is considered only valid in a single dialect. To investigate this further, I built the first multi-label ADI dataset called *MLADI*, with validity and ALDi labels assigned by 33 annotators from 11 Arab countries (chapter 7). Among the 978 dialectal sentences of the MLADI dataset that are valid in one of the 11 countries, only 25% are valid in a single country (out of 11 countries), and only 44% are valid in the dialects of a single region (out of 5 macro-regions). Hence, Arabic Dialect Identification should be modeled as a multi-label task, even on the macro-regional level. In collaboration with the NADI 2024 organizers, we presented the first attempt to model Arabic Dialect Identification as a multi-label task. The extended version of the shared task's dataset (the *MLADI* dataset) is hosted as a leaderboard to support further research in this direction.

Unlike the common conception that only short sentences could be valid in multiple dialects, I show that the conscious *Dialect Level* choice that Arabic speakers make—operationalized as ALDi—is a better predictor of the number of dialects in which a sentence is valid than its length. I also demonstrate that existing lists of supposedly distinctive lexical cues are less distinctive than previously thought. More rigorous validation is needed for such lists in the future, so that one can assume that a sentence is valid in a single dialect if it contains a lexical cue of this dialect.

8.2 Limitations and Going Forward

The contributions above have only opened the door for further investigations related to building a better representation of the different dialect levels and improving the performance of ADI systems. In this section, I outline some limitations of the current framing and propose potential future directions to mitigate them.

8.2.1 Going Beyond Written Arabic

All the contributions of this thesis are based on studying textual data (either written text or transcribed speech). It is conceivable that similar findings could be achieved for speech. However, incorporating the speech signal might require changes to how the different varieties of Arabic are represented and distinguished from each other. A major aspect would be related to the differentiation between dialects and accents. For speech inputs, Bafna and Wiesner (2025) recently showed that *language identification* systems tend to wrongly correlate accent with language (e.g., wrongly identifying Dutch-accented English as Dutch). Abdullah et al. (2025) reached a similar finding when investigating the manual annotations of a regional-level speech ADI dataset. While the two annotators they recruited achieved perfect agreement in categorizing dialectal speech, they disagreed on the classification of 2.3% of radio broadcast segments as being in MSA or being in a dialect. In this case, one annotator might have classified these segments as MSA irrespective of any cues of an accent. Conversely, the other annotator might have relied on the speaker's accent to identify their dialect, even if the speaker intended to produce MSA. Hence, speech ADI systems need to reconsider how to differentiate between accents and dialects. Building an ALDi estimation model for speech inputs would also necessitate defining how the different accents should impact the ALDi scores.

Moreover, parallels of some findings could exist for other languages. For instance, there is growing proof that Dialect Identification should be framed as a multi-label task for different languages such as English, French, and Spanish (Bernier-Colborne et al., 2023; Zampieri et al., 2024; Lopetegui et al., 2025). It would also be interesting to apply the concept of *Dialect Levels* to other languages, such as Swiss German, where a Standard variety coexists with non-standardized ones.

8.2.2 Reevaluating the Groupings of the Arabic Varieties

The varieties of Arabic are generally taxonomized into groups according to geographical factors, which are mostly constrained by the borders between the different Arab countries. Throughout this thesis, limitations of the adopted geography-based taxonomies arose. For instance, I found that a small proportion of MLADI's samples (only 11 samples to be specific) are exclusively valid in *Gulf of Aden* (represented by Yemen), unlike the other regions, which had at least 82 samples exclusively valid in their dialects. This provides preliminary evidence that the Gulf of Aden might not stand out as a separate regional variety of Arabic. Moreover, defining Arabic varieties on the country level, according to these countries' borders, does not acknowledge that speaking communities near the borders of two countries might be speaking the same variety of Arabic. This was clearly the case for the Algerian annotators recruited to annotate MLADI, where one of them lived in the east of Algeria, and hence spoke a variety (dialect) of Algerian Arabic that is closer to those spoken in Tunisia than the other two Algerian annotators. Hence, she labeled some sentences as valid in Algerian Arabic, which the other two Algerian annotators deemed as not valid, given their lack of knowledge of dialects spoken in the east of Algeria.

Therefore, there is a necessity to adopt new taxonomies that are linguistically grounded and not merely based on superficial borders between countries. Instead of only considering 22 varieties of Arabic on the country level, Glottolog and Ethnologue recognize 37 and 28 Arabic dialects, respectively. Moreover, the varieties of both taxonomies are not restricted by the geographical borders. Further investigations are needed to assess the effectiveness of incorporating these taxonomies into NLP models and whether there is a need for adopting more fine-grained taxonomies than the macro-regional/country-level ones.

8.2.3 ALDi is not a Completely Objective Distance Metric

ALDi is a variable that I newly introduced and defined as the *the* ***extent by which the sentence diverges from the standard language***. By definition, ALDi is a distance metric. However, we saw multiple signs that it is not an entirely objective metric in this thesis. For instance, significant differences exist in the ALDi ratings provided by annotators from the Maghreb compared to annotators from the Gulf and the Gulf of Aden. Significant differences also exist between Maghreb's annotators and annotators from the Levant and the Nile Basin, yet with a smaller effect size. Hence, ALDi annotations are influenced by the native dialect of each annotator, and further refinements of the annotation guidelines could attempt to mitigate this.

Moreover, *MLADI*'s annotators were only asked to provide ALDi ratings to sentences that they deemed as valid in their country-level dialects. My intuition was that a speaker of one dialect might perceive sentences in other dialects as more colloquial than sentences in their dialect. However, if we assume that some dialects are indeed closer to MSA than the others, then sentences in these MSA-proximate dialects could conversely be rated as more dialectal by speakers of these dialects than speakers of other dialects, as the former group of speakers could be using these sentences colloquially. Future work should investigate whether averaging a sentence's ALDi ratings, provided by all annotators, would be more accurate as a distance metric than averaging those provided only by annotators who perceived the sentence as valid in their country-level dialect.

8.2.4 Handling Code-switching to Other Languages

This thesis focused on studying online comments and tweets written in Arabic. While these samples would have diglossic code-switching, I found that the *AOC-ALDi* dataset is nearly free of samples code-switched between Arabic and other languages. Preliminary investigations indicate that the *Sentence-ALDI* model, fine-tuned on the *AOC-ALDi* dataset, might not be providing a consistently interpretable estimation of code-switched samples (check Table 8.1). That said, it is not easy to define how the model should handle code-switching to other languages, including the extreme cases of having whole sentences in languages other than Arabic.

	Sentence	ALDi
Original Sentence	related to research جزء actually كان في	0.332
Single Latin Span	جزء ليه علاقة بالبحث actually كان في	0.577
	related to research كان في فالحقيقة جزء	0.236
MSA Translation	كان هنالك حقاً جزء متعلق بالبحث	0
EGY Translation	كان في فالحقيقة جزء ليه علاقة بالبحث	0.413
English Translation	There was actually a part related to research	0.026

Table 8.1: A codeswitched sentence from the ZAEBUC corpus (Hamed et al., 2022), with multiple corresponding translations. The sentences' ALDi scores are automatically estimated using the *Sentence-ALDi* model.

The MLADI dataset has a few code-switched samples. During the annotation process, I noticed that the annotators rated these differently. Code-switching to English for Named Entities did not generally affect the ALDi ratings. Other cases of code-switching had some impact on the ratings. More specifically, some annotators rated the sentences as Level 2 (Normal Colloquial). In contrast, another group of annotators rated them as Level 3 (Informal Colloquial). When asked, the annotators referred to the levels' descriptions in the guidelines, where Level 3 sentences are described as *Tweets written in a colloquial language having expressions that* ***are not*** *accepted or* ***understood*** *by all members of society*. Some annotators considered English/French spans as normal colloquial ones that are accepted and understood by the whole community, hence chose Level 2. Conversely, the other group of annotators found code-switching to other languages unacceptable and rated these code-switched sentences as Level 3.

8.2.5 ALDi and Mutual Intelligibility

In §4.2, I showed that on randomly routing samples to random Arabic speakers, the interannotator agreement (IAA) between them decreases as the ALDi score of the samples increases. It is conceivable that high-ALDi samples in one dialect are less intelligible to speakers of other dialects than low-ALDi samples. However, relying on IAA is not enough to draw a causal relation between a sentence's ALDi score and its intelligibility to speakers of dialects other than the sentence's dialect.

In reality, testing mutual intelligibility between different languages—such as Dutch, English, and German—is a complex task for many reasons, one of which is the difficulty

of finding participants of one language with minimal or no exposure to the other languages (Nieder and List, 2024). This is even more difficult for the dialects of Arabic, as speakers of these dialects are more likely to be exposed to other dialects than speakers of different languages. For example, Egyptian Arabic is widely understood across the Arab world, as a vast majority of series, films, and songs consumed by the whole Arab world were produced in Egyptian Arabic. More broadly, S'hiri (2013) mentioned that the exposure to the Maghrebi varieties is insignificant compared with the Maghrebi exposure to the other dialects.

Consequently, the relationship between ALDi scores of sentences in a dialect and their intelligibility to speakers of other dialects is expected to be influenced by exposure to different dialects.

8.2.6 Investigating the Differences Between MSA and CA

Another potential direction to extend this thesis is to study the differences between MSA and CA. ALDi aims at studying how a sentence diverges from Standard Arabic. Given this definition, it is worth checking how it treats CA text. The Holy Qur'an is considered the most prominent book written in CA. To get some understanding of how CA is currently represented, I used the *Sentence-ALDi* model to automatically estimate the ALDi scores of the Qur'an's verses, after removing any diacritics. The distribution of ALDi scores in Figure 8.1 shows that the vast majority of the sentences are assigned low scores. Out of 6,256 considered verses, only 263 had ALDi scores > 0.11 ($\approx 4.2\%$), and only 5 had ALDi scores > 0.66. This suggests that the model cannot distinguish between the stylistic differences between MSA and CA, treating CA as equivalent to MSA.

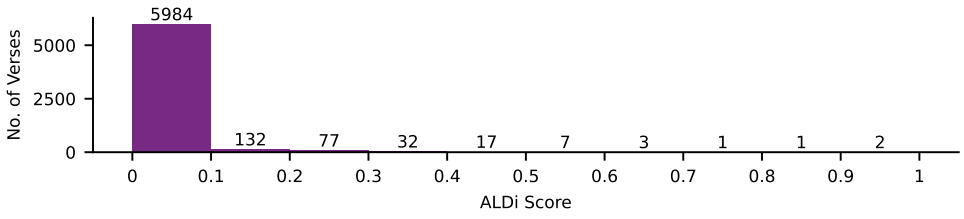


Figure 8.1: A histogram of the ALDi scores for the verses of the holy Qur'an, a book that is generally perceived as an example of Classical Arabic (CA).

Appendices

A Discarded Samples from AOC-ALDi

As mentioned in §3.2.2, I discarded 2,038 comments that have the majority of their ALDi annotations either set to *Not Arabic* or are missing. Five different categories of such comments were identified as per Table A1. These categories include sentences that have only punctuation marks, are written in English or Arabizi (Romanized Arabic), are just links to sites or emails, or have HTML encoded characters or formatting tags.

Reason for Discarding	Sentence	Source	ALDi Ratings
Symbols	؟؟؟؟؟	Cmnt (Y7)	¬ Arabic (x13), Missing (x2)
	********	Cmnt (Ri)	¬ Arabic (x3)
English	ممكن تلبسgloves to protect the baby from infection !	Cmnt (Ri)	¬ Arabic (x2), MSA (x1)
	I agree with you that racism exists in the United States; I also know it exists in Arab countries as well. Just remember that America elected a black president with 360 electoral college votes. In terms of numbers, that means a sweeping majority. Lets learn to be better than the Americans by developing our own democratic systems for a change...ccc	Cmnt (Gh)	¬ Arabic (x3)
	very nice...	Cmnt (Ri)	¬ Arabic (x3)
Arabizi	ya zamalek ya 7arameyaaaa	Cmnt (Y7)	¬ Arabic (x2), Most (x1)
	ma howeh el blogs m3abbiyeh el denya ? ya3ni law doctor el jam3a bedo yet3ab shway w yekteb articles, ma kan 3emel blog men zaman.	Cmnt (Gh)	¬ Arabic (x2), Most (x1)
URLs and Emails	http://elbeet-elmuslim.ace.st/forum.htm	Cmnt (Y7)	¬ Arabic (x3)
	Ahmad.altamimi@alghad.jo	Cntrl (Gh)	¬ Arabic (x3)
Presence of HTML	█████ 5000 √DONE	Cmnt (Y7)	¬ Arabic (x3)
	د. أشرف بلبع	Cntrl (Y7)	¬ Arabic (x3)
	بيتهيألى قربنا قوى من سبتمبر ١٩٨١	Cmnt (Y7)	¬ Arabic (x2) Most (x1)

Table A1: Examples of the discarded AOC comments with majority labels set to Not Arabic or missing. **Note**: **Cmnt** stands for comment, **Cntrl** stands for control sentence, **Y7**: Youm7, **Ri**: AlRiyadh, **Gh**: AlGhad.

B Detailed Description of the Datasets Used in §4.2

I noticed some discrepancies between the number of samples reported in the papers and the number of samples in the corresponding raw datasets. Despite following the filtration steps described in the papers, some of the datasets had more samples than the ones in the publicly released version, as indicated in Table B2. Additionally, the *ArSarcasm-v1*, *Mawqif (Stance Task)*, *Mawqif (Sentiment/Sarcasm Tasks)*, and *ASAD* had 516, 170, 151, and 191 samples with less than three annotations, respectively, that I decided to discard from my analysis.

Additionally, I did not include the MLMA dataset (Ousidhoum et al., 2019) in my analysis. The number of samples in the raw annotation files that MLMA's creators shared was too small compared to the number of samples in the public dataset with majority-vote labels. I also discarded another dataset, for which there was a significant discrepancy between the released dataset and its description in the respective paper.

Table B2: A detailed description of the distribution of the majority-vote labels and the data/paper discrepancies in the datasets with individual annotator labels included in my study. **Note 1:** *No Majority* means that multiple labels have the same majority number of votes for Individual/Proportion labels, and Confidence < 0.5 otherwise. **Note 2:** Some of the samples of the *ASAD*, *ArSarcasm-v1*, *Mawqif* datasets have more than three annotations, despite the fact that the former two are supposed to have only three annotations per sample.

Dataset	**Task (# labels)**	**Labels**	**Distribution of Majority-vote Labels**	**Dataset/Paper Discrepancy**
Deleted Comments Dataset (DCD) (Mubarak et al., 2017)	Offensive (3)	Confidence	Offensive (80.31%) Clean (17.76%) Obscene (1.58%) No Majority (0.35%)	-
MPOLD (Chowdhury et al., 2020)	Offensive (2)	Individual	Non-Offensive (83.12%) Offensive (16.88%)	-
YouTube Cyberbullying (YTCB) (Alakrot et al., 2018)	Offensive (2)	Individual	Not (61.38%) Hate-Speech (38.62%)	-
ASAD (Alharbi et al., 2021)	Sentiment (3)	Individual	Neutral (67.83%) Negative (15.33%) Positive (15.19%) No Majority (1.65%)	The authors shared with us the raw annotation file of which I analyze 100,484 samples with three annotations or more, as opposed to the 95,000 in the released dataset.

ArSAS (Elmadany et al., 2018)	Sentiment (4)	Confidence	Negative (35.38%) Neutral (33.45%) Positive (20.51%) No Majority (6.07%) Mixed (4.59%)	-
	Speech Act (6)	Confidence	Expression (55.07%) Assertion (38.63%) Question (3.32%) No Majority (1.81%) Request (0.67%) Recommendation (0.31%) Miscellaneous (0.18%)	
ArSarcasm-v1 (Abu Farha and Magdy, 2020)	Dialect (5)	Individual	MSA (67.56%) Egypt (19.37%) No Majority (5.83%) Gulf (3.61%) Levant (3.46%) Maghreb (0.18%)	The samples in the raw annotation artifact shared by the authors has 10,641 samples, as opposed to the 10,547 samples in the released dataset.
	Sarcasm (2)	Individual	False (84.24%) True (15.7%) No Majority (0.06%)	
	Sentiment (3)	Individual	neutral (49.45%) negative (32.57%) positive (14.58%) No Majority (3.4%)	

Mawqif (Alturayeif et al., 2022)	Sarcasm (2)	Individual	No (95.97%) Yes (3.78%) No Majority (0.25%)	The authors annotated the same samples for sentiment/sarcasm and stance separately. This was done across eight different annotation jobs (4 each), for which the authors shared the raw annotation files with us. The number of samples in these files is 4,093 for sentiment/sarcasm and 4,079 for stance, of which 3,942 and 3,909 have three or more annotations. The released dataset is reported to have 4,100 samples.
	Sentiment (3)	Individual	Positive (41.15%) Negative (31.46%) Neutral (22.68%) No Majority (4.72%)	
	Stance (3)	Individual	Favor (60.5%) Against (27.65%) None (7.7%) No Majority (4.14%)	
iSarcasm's test set (Abu Farha et al., 2022)	Dialect (5)	Individual	MSA (32.29%) Nile (31.36%) Gulf (16.5%) No Majority (15.79%) Levant (2.21%) Maghreb (1.86%)	The dataset having the individual annotator labels is released as an artifact accompanying the following paper (Abu Farha and Magdy, 2022).
	Sarcasm (2)	Individual	0 (82.07%) 1 (17.93%)	

DART (Alsarsour et al., 2018)	Dialect (5)	Proportion	GLF (24.27%) EGY (21.69%) IRQ (21.64%) LEV (16.22%) MGH (16.18%)	-

C Ethical Considerations of the Annotation Processes

Error Analysis Experiment (chapter 5)

The study was approved by the research ethics committee of the University of Edinburgh School of Informatics, with reference number 207712.

MLADI Dataset (chapter 6, chapter 7)

The dataset has a few samples with offensive language. The annotators were asked to provide consent confirming their agreement to annotate these samples at the start of the annotation process. The annotation process I followed for creating the evaluation dataset of the first two subtasks was approved by the research ethics committee of the University of Edinburgh, School of Informatics, with reference number 839548.

Despite having three annotators per country, my crowdsourced annotators are skewed toward younger age groups and have/are pursuing higher education degrees. Therefore, I acknowledge that my results could be representative of the perceptions of specific demographics within each country.

The analyzed tweets' geolocations are uniformly balanced across 14 different Arab countries, covering a wide range of Arabic dialects. However, I acknowledge that some sub-dialects are not well represented online, as shown by (Mohamed Eida et al., 2024) for the Sa'idi Arabic variety of Egypt. Moreover, the data does not have Arabic sentences written in Latin script (known as Arabizi). Arabizi is prominently used in the Maghreb region (Younes et al., 2015), and to a lesser extent in other countries such as Lebanon and Egypt (Tobaili, 2016).

D The Annotation Guidelines Used in chapter 5

Figure D1, Figure D2, Figure D3, and Figure D4 show screenshots of the *Qualtrics* survey I created to analyze the errors of the country-level ADI model presented in §5.4.2 of chapter 5.

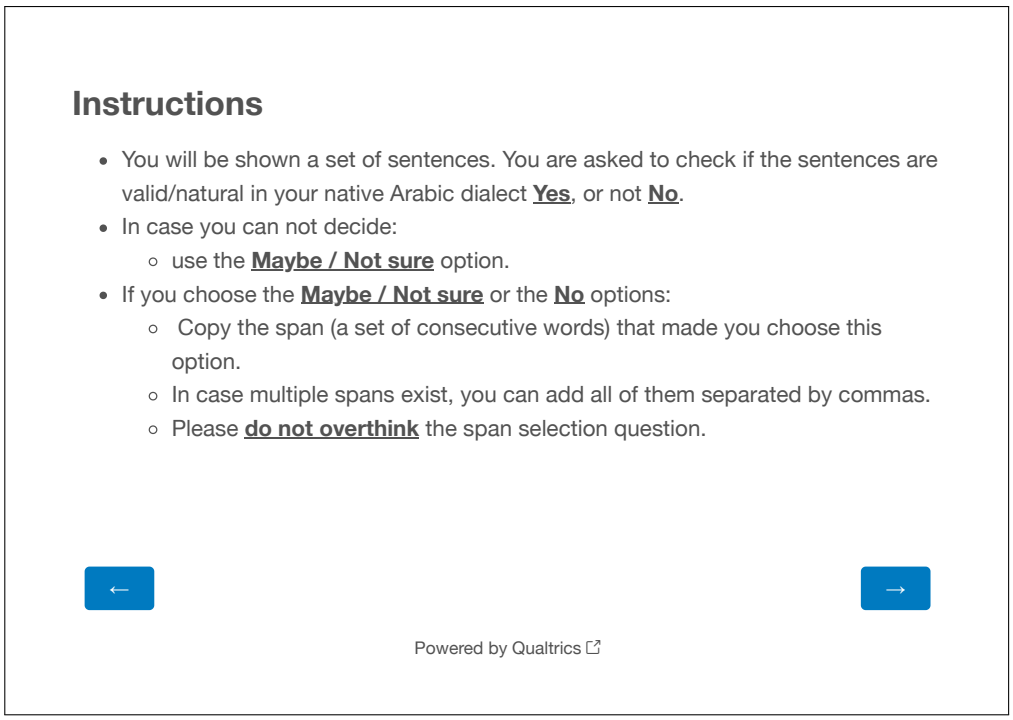


Figure D1: Screenshot of the instruction provided to the participants of the error analysis survey.

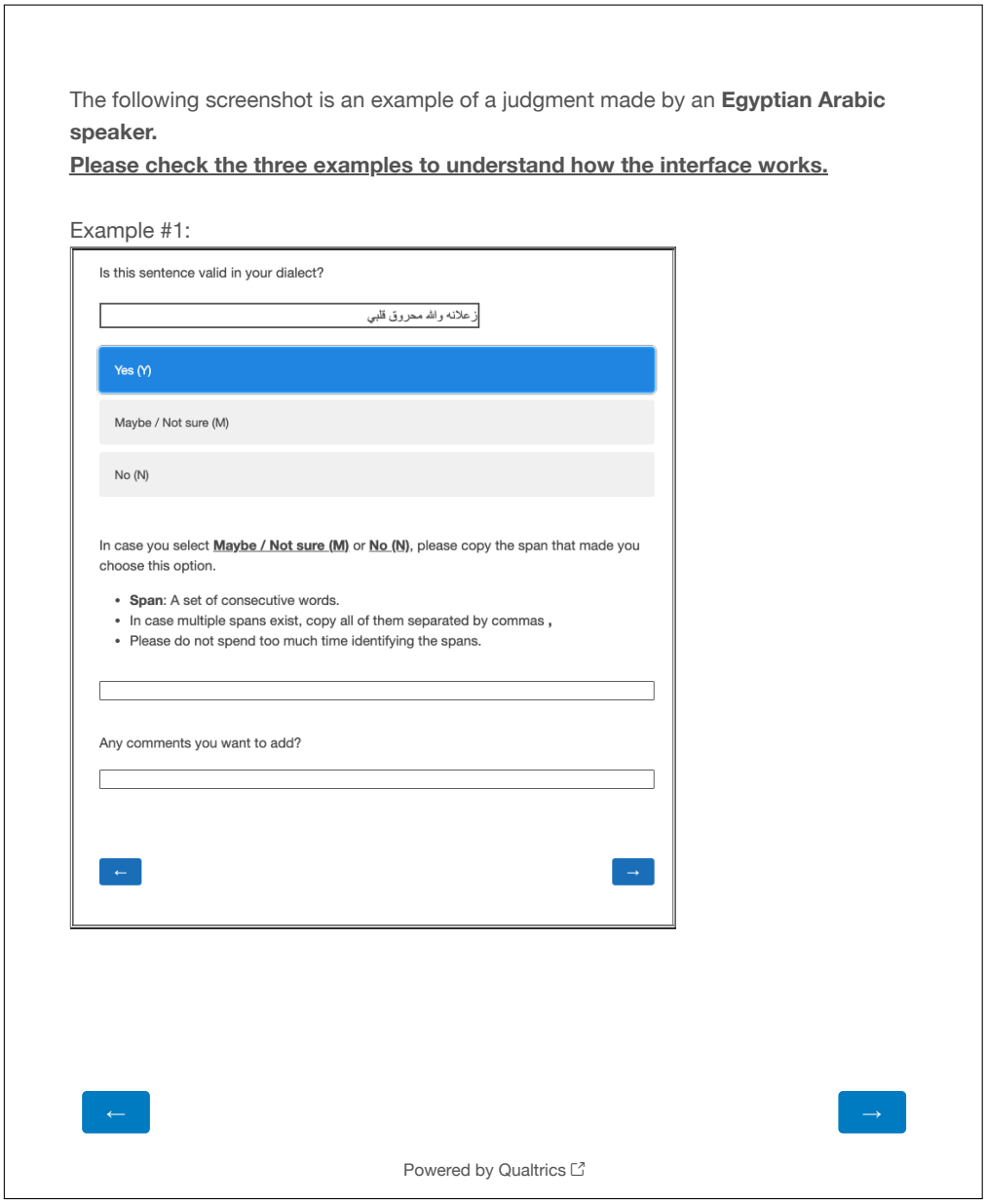


Figure D2: Screenshot of the first example shown to the participants of the error analysis survey.

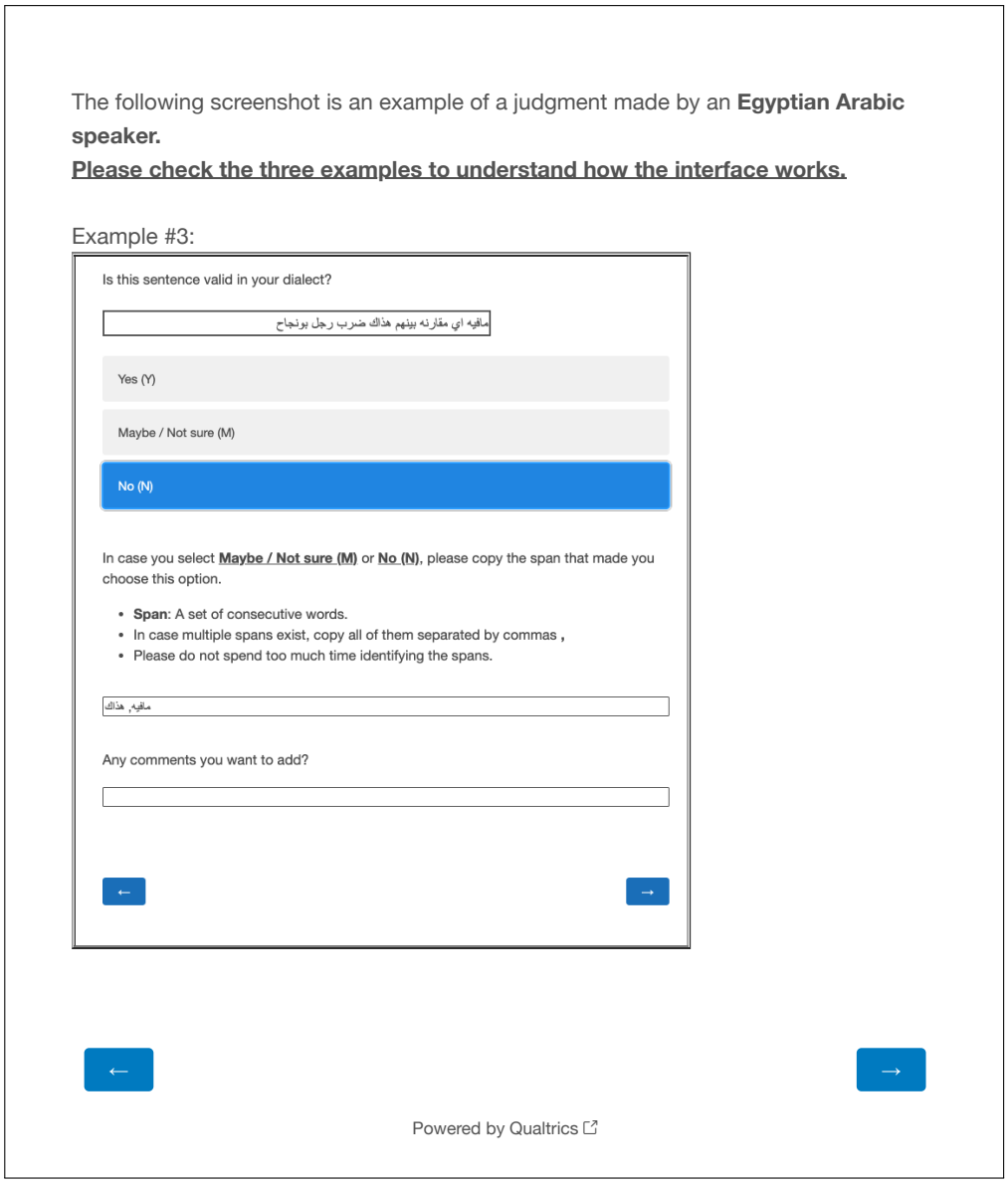


Figure D3: Screenshot of the third example shown to the participants of the error analysis survey.

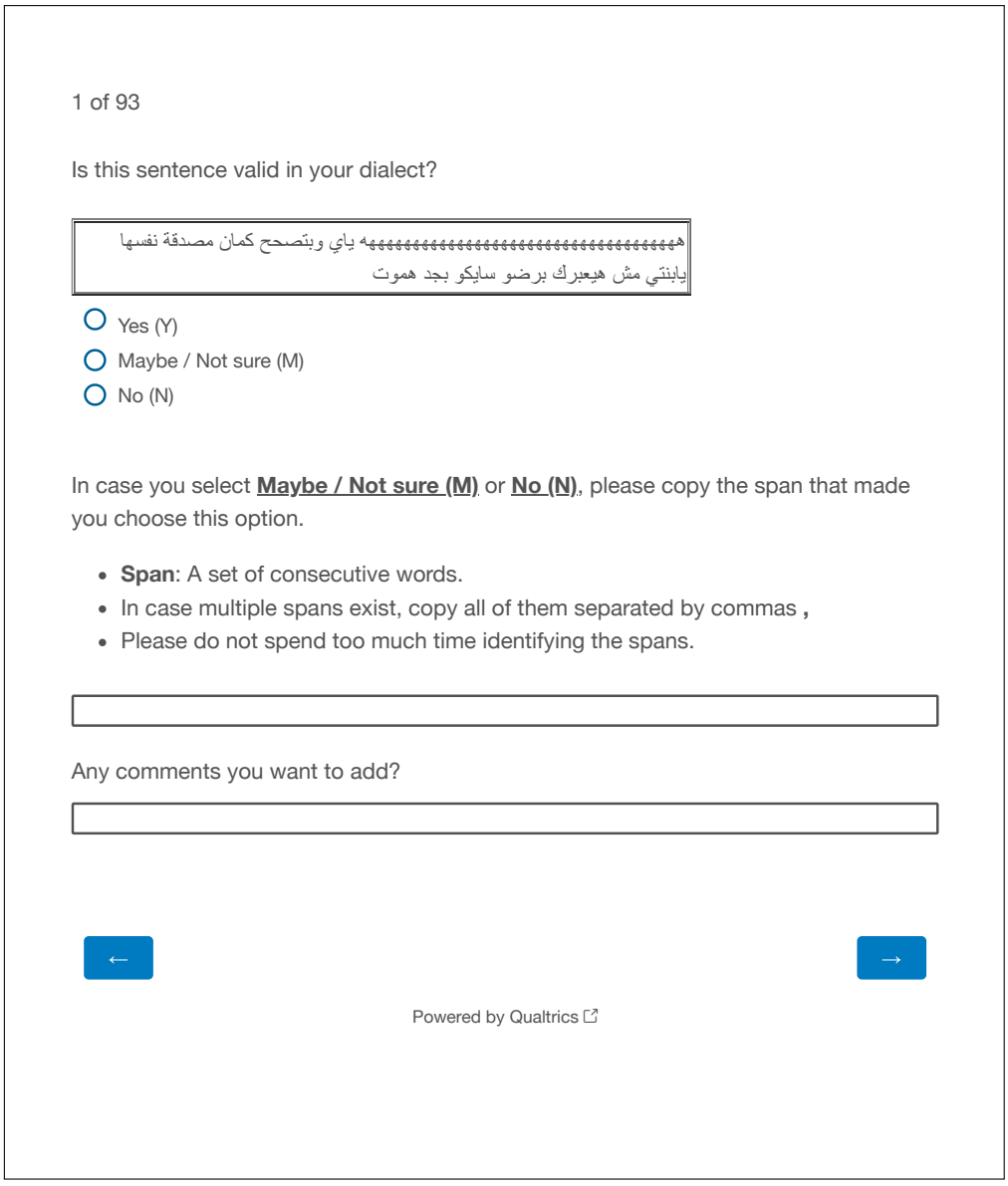


Figure D4: Screenshot demonstrating how the sentences were shown to the participants in the error analysis survey.

E English Translation of the Annotation Guidelines Used in chapter 6 and chapter 7

Please see the following instructions if you need to remember the labeling rules.

Q1) The possibility that the tweets were written by a speaker in your native dialect

- A group of tweets will appear.
- You are asked to check whether each Tweet can be written by a speaker in your native Arabic dialect.
- If you are not sure whether the tweet is correct in your dialect, select **Not Sure** .

Note: Please consider the dialects of cities and regions of your country as part of your dialect.

Choices:

- **Yes** (the sentence is written in colloquial or formal Arabic or a mixture of both, spoken by speakers of your dialect, the dialect of your city or country)
- **Not sure**
- **No** (there is no one from my country who has not been influenced by other dialects and writes like this)

Illustrative examples of some sentences that are not acceptable in the dialects of different Arab countries:

Country	Tweet	Explanation of choosing No (just for clarification)
Syria	اللي عاليمينن نايس اللي عالشمال ما بحبوو كلشششش	"كلشششش" is a word that is not used in Syria.
Saudi Arabia	وانا بشهد ياعيني على الحجازيات وبعدها الجنوبيات باليبه بس جمال صحباتي حروب بيقولو للقمر قوم لاقعد مطرحك ولا الغامديه قشطه على عسل صافي #بنات_السعوديه_الاكثر_جمالا	"بيقولو، لاقعد مطرحك" are words that are not used in Saudi Arabia.
Egypt	! بجد الناس اللي بتنسى بسرعة بجد كبيبف تعملو ها ! ؟	"كبيبف تعملو ها" is a not an expression used in Egypt.
Lebanon	خرج حدا يشرح يا جماعة ؟ على فكرة وحشتوني اوي اوي	"على فكرة وحشتوني اوي" is an expression that is not used in Lebanon
Algeria	بالطيف مقتلي خير بلاد تسخف حالها علاه هكا ملا حالة والله	"علاه هكا، ملا حالة" are terms not used in Algeria.

Figure E5: English translation of the first page of annotation guidelines of chapter 6.

Q2) Determining the level of colloquial/dialectal language

For tweets that could have been written by a speaker in your native dialect, please rate the level of dialectness of each tweet.

Explanation of some of the characteristics of the different levels

- **Sound MSA:** Tweets written in fluent MSA.
- **Formal Colloquial or Colloquial-influenced MSA:** Tweets written in a language close to MSA but using some colloquial expressions (lexemes/ morphemes).
- **Natural/Ordinary Colloquial:** Tweets written in a colloquial language that is accepted and understood by all members of society of all ages and social/educational levels.
- **Informal (or Vulgar) Colloquial:** Tweets written in a colloquial language having expressions that are not accepted or understood by all members of society. It does not have to be vulgar or weak.

As an Arabic speaker on X (formerly Twitter), I might use standard Arabic to talk to an official account of an Arab embassy, semi-standard colloquial Arabic to offer condolences to a friend, normal (casual) Arabic to express an opinion on a societal issue, and informal Arabic to joke with a close friend.

An illustrative example only at the single word level:
The following verbs are used to express feelings of happiness and pleasure.

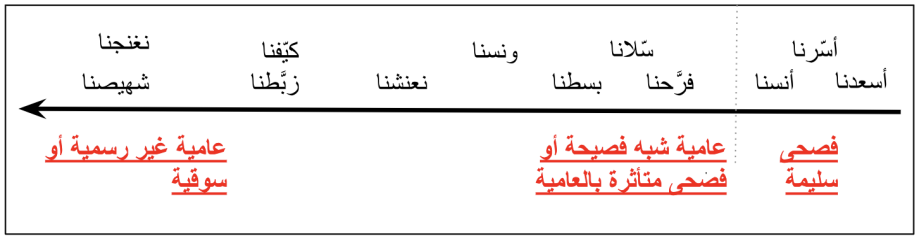


Figure E6: English translation of the second page of annotation guidelines of chapter 6.

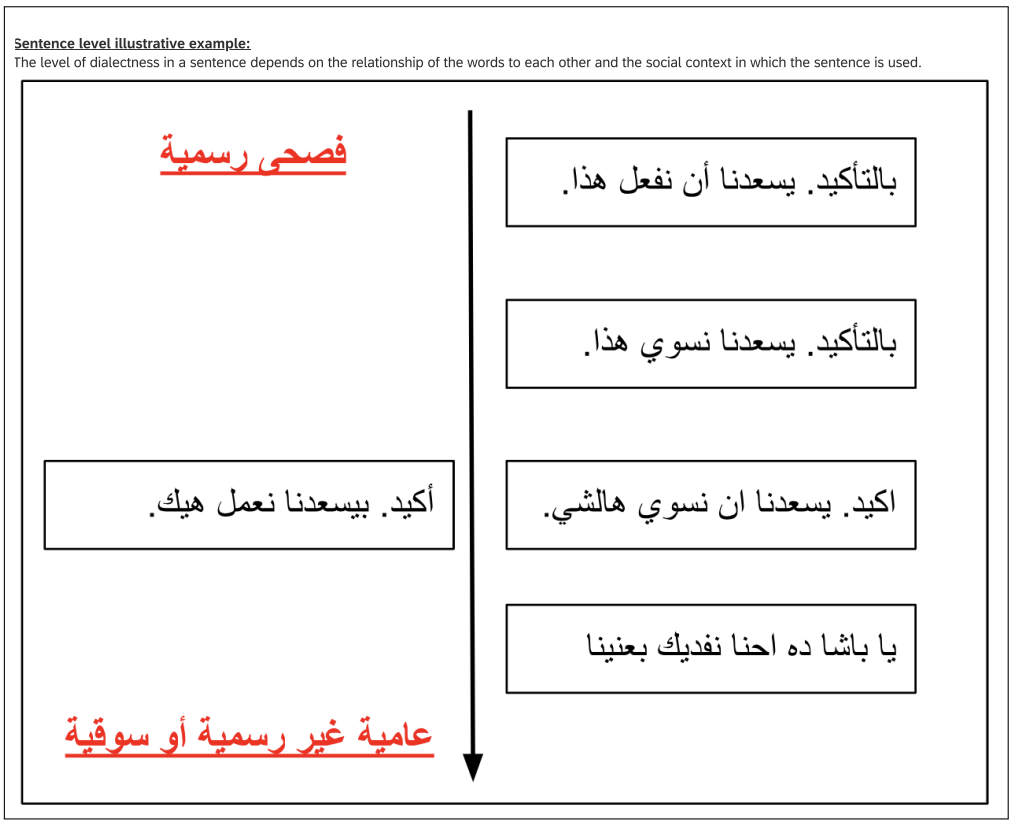


Figure E7: English translation of the third page of annotation guidelines of chapter 6.

F MLADI's Country-level Overlap

I compute the percentage of samples within our dataset that are manually labeled as valid in multiple country-level dialects by annotators from these countries, thereby extending the analysis in chapter 6 by covering two additional country-level dialects. Only 249 sentences ($\approx 25\%$) are single-label as per Figure F8, compared to the $\approx 30\%$ reported for 9 country-level dialects on NADI 2024's development set (see chapter 6). This indicated that incorporating more country-level dialects would still increase the already high percentage of multi-label samples.

I also show the cross-country overlap in Figure F9. While it is clear that countries within the same region overlap more with each other, a substantial overlap also exists with countries from other regions. However, I showed in §7.3.4 of chapter 7 that sentences valid in two country-level dialects can still be perceived differently by speakers of these dialects. Hence, relying solely on the number of sentences shared between two country-level dialects is insufficient to identify the proximity between these dialects.

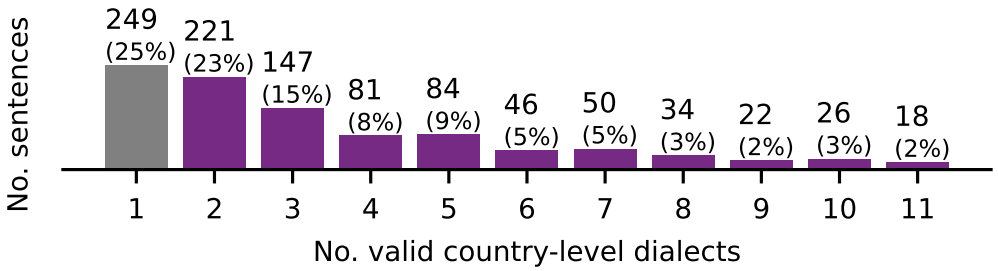


Figure F8: The histogram of the number of dialects in which a sentence is valid on the country-level dialects.

Theoretically, the dataset is uniformly representative of the 14 different countries to which the samples were geolocated. However, it was found that the precision of their geolocation methodology varies for the different countries, and is the lowest for the countries of the Maghreb region (49.3% for Tunisia, 57.3% for Morocco, and 65.3% for Algeria). Hence, we think that further investigations are required before using these percentages as proxies for proximity between dialects.

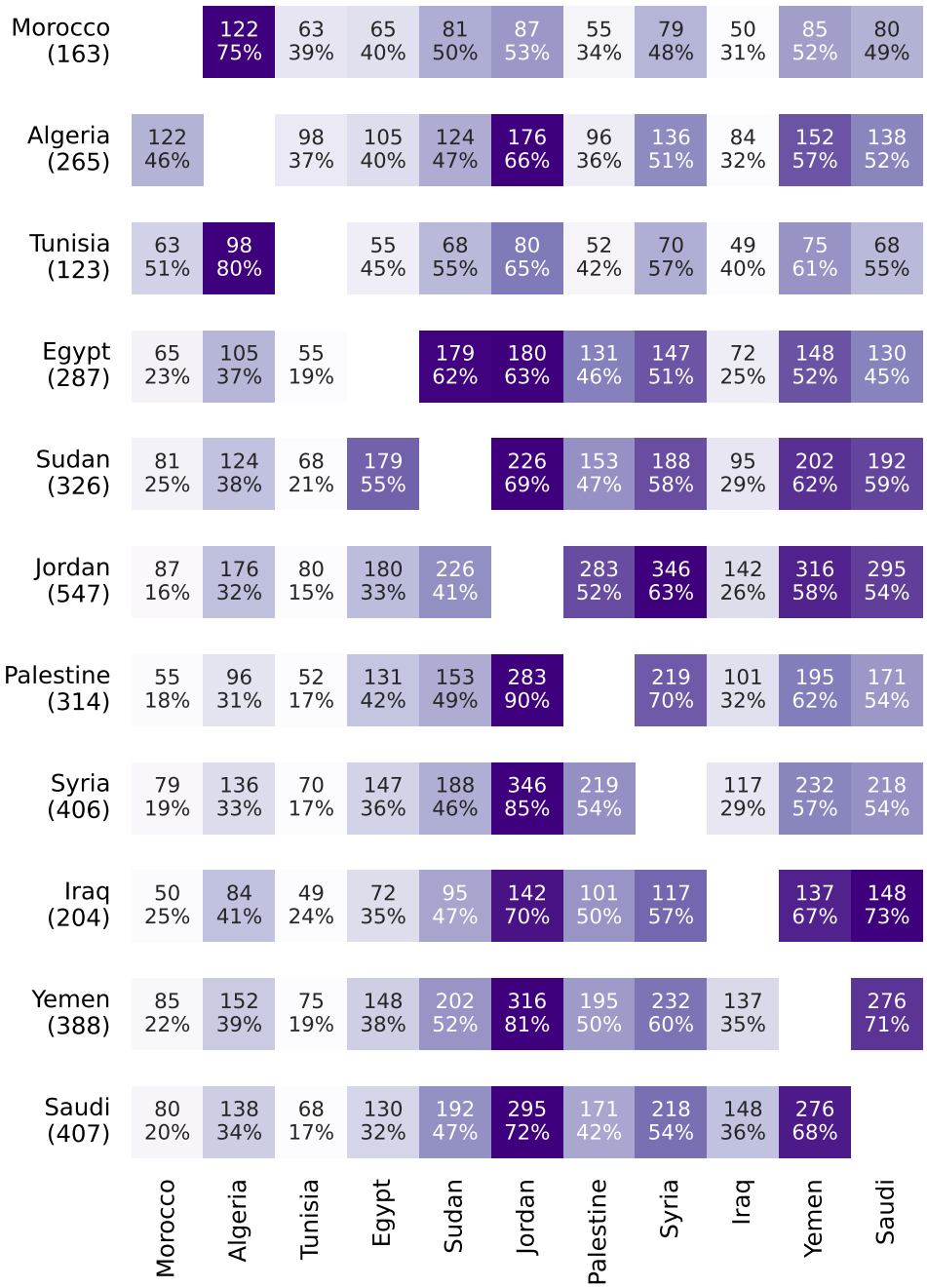


Figure F9: The percentage and number of each row country's valid samples that are also valid in the column country. The total number of samples labeled as valid in each country is presented below each country's row label. **Note:** Each row's colormap range is independent from the other rows.

Bibliography

AAlAbdulsalam, A. K. (2022). SQU-CS @ NADI 2022: Dialectal Arabic identification using one-vs-one classification with TF-IDF weights computed on character n-grams. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 436–441, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., and Darwish, K. (2021). QADI: Arabic dialect identification in the wild. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abdul-Mageed, M., Alhuzali, H., and Elaraby, M. (2018). You tweet what you speak: A city-level dataset of Arabic dialects. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Abdul-Mageed, M. and Diab, M. (2011). Subjectivity and sentiment annotation of Modern Standard Arabic newswire. In Ide, N., Meyers, A., Pradhan, S., and Tomanek, K., editors, Proceedings of the 5th Linguistic Annotation Workshop, pages 110–118, Portland, Oregon, USA. Association for Computational Linguistics.

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2021a). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural

Language Processing (Volume 1: Long Papers), pages 7088–7105, Online. Association for Computational Linguistics.

Abdul-Mageed, M., Elmadany, A., Zhang, C., Nagoudi, E. M. B., Bouamor, H., and Habash, N. (2023). NADI 2023: The fourth nuanced Arabic dialect identification shared task. In Sawaf, H., El-Beltagy, S., Zaghouani, W., Magdy, W., Abdelali, A., Tomeh, N., Abu Farha, I., Habash, N., Khalifa, S., Keleg, A., Haddad, H., Zitouni, I., Mrini, K., and Almatham, R., editors, Proceedings of ArabicNLP 2023, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Abdul-Mageed, M., Keleg, A., Elmadany, A., Zhang, C., Hamed, I., Magdy, W., Bouamor, H., and Habash, N. (2024). NADI 2024: The fifth nuanced Arabic dialect identification shared task. In Habash, N., Bouamor, H., Eskander, R., Tomeh, N., Abu Farha, I., Abdelali, A., Touileb, S., Hamed, I., Onaizan, Y., Alhafni, B., Antoun, W., Khalifa, S., Haddad, H., Zitouni, I., AlKhamissi, B., Almatham, R., and Mrini, K., editors, Proceedings of the Second Arabic Natural Language Processing Conference, pages 709–728, Bangkok, Thailand. Association for Computational Linguistics.

Abdul-Mageed, M., Zhang, C., Bouamor, H., and Habash, N. (2020a). NADI 2020: The first nuanced Arabic dialect identification shared task. In Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., and Zaghouani, W., editors, Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. (2021b). NADI 2021: The second nuanced Arabic dialect identification shared task. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 244–259, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. (2022). NADI 2022: The third nuanced Arabic dialect identification shared task. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 85–97, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abdul-Mageed, M., Zhang, C., Elmadany, A., and Ungar, L. (2020b). Toward micro-dialect identification in diaglossic and code-switched environments. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5855–5876, Online. Association for Computational Linguistics.

Abdullah, B. M., Baas, M., Möbius, B., and Klakow, D. (2025). Voice conversion improves cross-domain robustness for spoken Arabic dialect identification.

Abu Farha, I. and Magdy, W. (2020). From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In Al-Khalifa, H., Magdy, W., Darwish, K., Elsayed, T., and Mubarak, H., editors, Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 32–39, Marseille, France. European Language Resource Association.

Abu Farha, I. and Magdy, W. (2022). The effect of Arabic dialect familiarity on data annotation. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 399–408, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Abu Farha, I., Oprea, S. V., Wilson, S., and Magdy, W. (2022). SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In Emerson, G., Schluter, N., Stanovsky, G., Kumar, R., Palmer, A., Schneider, N., Singh, S., and Ratan, S., editors, Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 802–814, Seattle, United States. Association for Computational Linguistics.

Abu Farha, I., Zaghouani, W., and Magdy, W. (2021). Overview of the WANLP 2021 shared task on sarcasm and sentiment detection in Arabic. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 296–305, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Abu Kwaik, K. and Saad, M. (2019). ArbDialectID at MADAR shared task 1: Language modelling and ensemble learning for fine grained Arabic dialect identification. In

El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 254–258, Florence, Italy. Association for Computational Linguistics.

Abu Kwaik, K., Saad, M., Chatzikyriakidis, S., and Dobnik, S. (2018). Shami: A corpus of Levantine Arabic dialects. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Adouane, W. and Dobnik, S. (2017). Identification of languages in Algerian Arabic multilingual documents. In Habash, N., Diab, M., Darwish, K., El-Hajj, W., Al-Khalifa, H., Bouamor, H., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Third Arabic Natural Language Processing Workshop, pages 1–8, Valencia, Spain. Association for Computational Linguistics.

Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., and Abdullah, S. (2018). Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis. Journal of Information Science, 44(3):345–362.

Al-Obaidi, A. Y. and Samawi, V. W. (2016). Opinion mining: Analysis of comments written in Arabic colloquial. In Proceedings of the World Congress on Engineering and Computer Science, volume 1.

Al-Sabbagh, R. and Girju, R. (2012). YADAC: Yet another dialectal Arabic corpus. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2882–2889, Istanbul, Turkey. European Language Resources Association (ELRA).

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). Morphologically annotated corpora and morphological analyzers for Moroccan and sanaani yemeni Arabic. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Tenth International Conference on Language Resources and

Evaluation (LREC'16), pages 1300–1306, Portorož, Slovenia. European Language Resources Association (ELRA).

Alakrot, A., Murray, L., and Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in Arabic. Procedia Computer Science, 142:174–181. Arabic Computational Linguistics.

Alghamdi, A. (2021). العرنجية - بلسان عربي هجين Arangia - with a Hybrid Arabic Tongue. Takween.

Alharbi, B., Alamro, H., Alshehri, M., Khayyat, Z., Kalkatawi, M., Jaber, I. I., and Zhang, X. (2021). ASAD: A twitter-based benchmark Arabic sentiment analysis dataset.

Alharbi, S., Alowisheq, A., Tüske, Z., Darwish, K., Alrajeh, A., Alrowithi, A., Tamran, A. B., Ibrahim, A., Aloraini, R., Alnajim, R., Alkahtani, R., Almuasaad, R., Alrasheed, S., Alsubaie, S., and Alonaizan, Y. (2024). SADA: Saudi audio dataset for Arabic. In ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 10286–10290.

AlKhamissi, B., Gabr, M., ElNokrashy, M., and Essam, K. (2021). Adapting MARBERT for improved Arabic dialect identification: Submission to the NADI 2021 shared task. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 260–264, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Almeman, K. and Lee, M. G. (2013). Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. 2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA), pages 1–6.

Aloraini, A., Poesio, M., and Alhelbawy, A. (2020). The QMUL/HRBDT contribution to the NADI Arabic dialect identification shared task. In Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., and Zaghouani, W., editors, Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 295–301, Barcelona, Spain (Online). Association for Computational Linguistics.

Alorifi, F. (2008). Automatic Identification of Arabic Dialects Using Hidden Markov Models. PhD thesis, University of Pittsburgh.

Alsarsour, I., Mohamed, E., Suwaileh, R., and Elsayed, T. (2018). DART: A large dataset of dialectal Arabic tweets. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Alshabanah, A. and Annavaram, M. (2025). On using Arabic language dialects in recommendation systems. In Chiruzzo, L., Ritter, A., and Wang, L., editors, Findings of the Association for Computational Linguistics: NAACL 2025, pages 2178–2186, Albuquerque, New Mexico. Association for Computational Linguistics.

Alshargi, F., Dibas, S., Alkhereyf, S., Faraj, R., Abdulkareem, B., Yagi, S., Kacha, O., Habash, N., and Rambow, O. (2019). Morphologically annotated corpora for seven Arabic dialects: Taizi, Sanaani, Najdi, Jordanian, Syrian, Iraqi and Moroccan. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 137–147, Florence, Italy. Association for Computational Linguistics.

Alshutayri, A. (2017). Exploring Twitter as a source of an Arabic dialect corpus. International Journal of Computational Linguistics (IJCL), 8(2):37–44.

Alsudais, A., Alotaibi, W., and Alomary, F. (2022). Similarities between Arabic dialects: Investigating geographical proximity. Information Processing & Management, 59(1):102770.

Altaher, Y., Fadel, A., Alotaibi, M., Alyazidi, Z., et al. (2022). Masader Plus: A new interface for exploring +500 Arabic NLP datasets. arXiv preprint arXiv:2208.00932.

Altanir, M. D. (2017). ألفاظ عامية فصيحة Standard Colloquial Words. Dar El-Shorouk.

Althobaiti, M. J. (2020). Automatic Arabic dialect identification systems for written texts: A survey.

Althobaiti, M. J. (2022). Creation of annotated country-level dialectal Arabic resources: An unsupervised approach. Natural Language Engineering, 28(5):607–648.

Alturayeif, N. S., Luqman, H. A., and Ahmed, M. A. K. (2022). Mawqif: A multi-label Arabic dataset for target-specific stance detection. In Bouamor, H., Al-Khalifa,

H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 174–184, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alyafeai, Z., Masoud, M., Ghaleb, M., and Al-shaibani, M. S. (2021). Masader: Metadata sourcing for Arabic text and speech data resources.

Aminian, M., Ghoneim, M., and Diab, M. (2015). Unsupervised false friend disambiguation using contextual word clusters and parallel word alignments. In Wu, D., Carpuat, M., Agirre, E., and Aranberri, N., editors, Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 39–48, Denver, Colorado, USA. Association for Computational Linguistics.

Aoun, J., Benmamoun, E., and Sportiche, D. (1994). Agreement, word order, and conjunction in some varieties of Arabic. Linguistic Inquiry, 25(2):195–220.

Attieh, J. and Hassan, F. (2022). Arabic dialect identification and sentiment classification using transformer-based models. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 485–490, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Badawi, A.-S. M. (1973). مستويات العربية المعاصرة في مصر. Levels of Contemporary Arabic in Egypt. Dar Al-Maarif.

Bafna, N. and Wiesner, M. (2025). LID models are actually accent classifiers: Implications and solutions for LID on accented speech.

Baimukan, N., Bouamor, H., and Habash, N. (2022). Hierarchical aggregation of dialectal data for Arabic dialect identification. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 4586–4596, Marseille, France. European Language Resources Association.

Baldwin, T. and Lui, M. (2010). Language identification: The long and the short of the matter. In Kaplan, R., Burstein, J., Harper, M., and Penn, G., editors, Human

Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 229–237, Los Angeles, California. Association for Computational Linguistics.

Bassiouney, R. (2009). Code-switching, page 28–87. Edinburgh University Press.

Bayrak, G. and Issifu, A. M. (2022). Domain-adapted BERT-based models for nuanced Arabic dialect identification and tweet sentiment analysis. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 425–430, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Behnstedt, P. and Woidich, M. (2013). Arabic Dialectology. In The Oxford Handbook of Arabic Linguistics. Oxford University Press.

Bergman, A. and Diab, M. (2022). Towards responsible natural language annotation for the varieties of Arabic. In Muresan, S., Nakov, P., and Villavicencio, A., editors, Findings of the Association for Computational Linguistics: ACL 2022, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Bernier-Colborne, G., Goutte, C., and Leger, S. (2023). Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In Scherrer, Y., Jauhiainen, T., Ljubešić, N., Nakov, P., Tiedemann, J., and Zampieri, M., editors, Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken Arabic dialect identification using phonotactic modeling. In Rosner, M. and Wintner, S., editors, Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pages 53–61, Athens, Greece. Association for Computational Linguistics.

Blanc, H. (1960). Style Variations in Spoken Arabic: A Sample of Interdialectical Educated Conversation. Contributions to Arabic linguistics. 1. Harvard University Press.

Blodgett, S. L. and O'Connor, B. (2017). Racial disparity in natural language processing: A case study of social media african-american english.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1240–1245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Bouamor, H., Hassan, S., and Habash, N. (2019). The MADAR shared task on Arabic fine-grained dialect identification. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Business for Social Responsibility (2022). Human rights due diligence of meta's impacts in Israel and Palestine in may 2021. https://about.fb.com/wp-content/uploads/2022/09/Human-Rights-Due-Diligence-of-Metas-Impacts-in-Israel-and-Palestine-in-May-2021.pdf.

Chambers, J. K. and Trudgill, P. (1998). Dialectology. Cambridge Textbooks in Linguistics. Cambridge University Press, 2 edition.

Charfi, A., Zaghouani, W., Mehdi, S. H., and Mohamed, E. (2019). A fine-grained annotated multi-dialectal Arabic corpus. In Mitkov, R. and Angelova, G., editors, Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 198–204, Varna, Bulgaria. INCOMA Ltd.

Chen, G., Liu, F., Meng, Z., and Liang, S. (2022). Revisiting parameter-efficient tuning: Are we really there yet? In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, Proceedings of the 2022 Conference on Empirical Methods in Natural Language

Processing, pages 2612–2626, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chiang, D., Diab, M., Habash, N., Rambow, O., and Shareef, S. (2006). Parsing Arabic dialects. In McCarthy, D. and Wintner, S., editors, 11th Conference of the European Chapter of the Association for Computational Linguistics, pages 369–376, Trento, Italy. Association for Computational Linguistics.

Chowdhury, S. A., Mubarak, H., Abdelali, A., Jung, S.-g., Jansen, B. J., and Salminen, J. (2020). A multi-platform Arabic news comment dataset for offensive language detection. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6203–6212, Marseille, France. European Language Resources Association.

Condon, S., Arehart, M., Parvaz, D., Sanders, G., Doran, C., and Aberdeen, J. (2012). Evaluation of 2-way iraqi arabic—english speech translation systems using automated metrics. Machine Translation, 26(1–2):159–176.

Condon, S., Parvaz, D., Aberdeen, J., Doran, C., Freeman, A., and Awad, M. (2010). Evaluation of machine translation errors in English and iraqi Arabic. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

Condon, S., Parvaz, D., Aberdeen, J., Doran, C., Freeman, A., and Awad, M. (2011). Machine translation errors: English and iraqi arabic. ACM Transactions on Asian Language Information Processing, 10(1).

Condon, S., Sanders, G. A., Parvaz, D., Rubenstein, A., Doran, C., Aberdeen, J., and Oshika, B. (2009). Normalization for automated metrics: English and Arabic speech translation. In Proceedings of Machine Translation Summit XII: Papers, Ottawa, Canada.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., edi-

tors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 241–245, Reykjavik, Iceland. European Language Resources Association (ELRA).

Cotterell, R., Renduchintala, A., Saphra, N., and Callison-Burch, C. (2014). An Algerian Arabic-French code-switched corpus. In Workshop on free/open-source Arabic corpora and corpora processing tools workshop programme, page 34.

Cowell, M. W., of Languages, G. U. I., and Linguistics. (1964). A reference grammar of Syrian Arabic: based on the dialect of Damascus. Arabic series / Georgetown University. Institute of Languages and Linguistics; no. 7. Georgetown University Press, Washington.

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., Kublik, S., Amer, M., Aryabumi, V., Campos, J. A., Tan, Y.-C., Kocmi, T., Strub, F., Grinsztajn, N., Flet-Berliac, Y., Locatelli, A., Lin, H., Talupuru, D., Venkitesh, B., Cairuz, D., Yang, B., Chung, T., Ko, W.-Y., Shi, S. S., Shukayev, A., Bae, S., Piktus, A., Castagné, R., Cruz-Salinas, F., Kim, E., Crawhall-Stein, L., Morisot, A., Roy, S., Blunsom, P., Zhang, I., Gomez, A., Frosst, N., Fadaee, M., Ermis, B., Üstün, A., and Hooker, S. (2024). Aya Expanse: Combining research breakthroughs for a new multilingual frontier.

Darwish, K. and Magdy, W. (2014). Arabic information retrieval. Foundations and Trends® in Information Retrieval, 7(4):239–342.

Darwish, K., Sajjad, H., and Mubarak, H. (2014). Verifiably effective Arabic dialect identification. In Moschitti, A., Pang, B., and Daelemans, W., editors, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1465–1468, Doha, Qatar. Association for Computational Linguistics.

Das, A. and Gambäck, B. (2014). Identifying languages at the word level in code-mixed Indian social media text. In Sharma, D. M., Sangal, R., and Pawar, J. D., editors, Proceedings of the 11th International Conference on Natural Language Processing, pages 378–387, Goa, India. NLP Association of India.

Demszky, D., Sharma, D., Clark, J., Prabhakaran, V., and Eisenstein, J. (2021). Learning to recognize dialect features. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y.,

editors, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2315–2338, Online. Association for Computational Linguistics.

Denny, F. M. (1989). Qur'ān recitation: A tradition of oral performance and transmission. Oral tradition, 4(1-2):5–26.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Dhaou, G. and Lejeune, G. (2020). Comparison between voting classifier and deep learning methods for Arabic dialect identification. In Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., and Zaghouani, W., editors, Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 243–249, Barcelona, Spain (Online). Association for Computational Linguistics.

Diab, M., Hacioglu, K., and Jurafsky, D. (2004). Automatic tagging of Arabic text: From raw text to base phrase chunks. In Proceedings of HLT-NAACL 2004: Short Papers, pages 149–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Diab, M. T. (2004). An unsupervised approach for bootstrapping Arabic sense tagging. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, pages 43–50, Geneva, Switzerland. COLING.

Drozdík, L. (2006). Prestigious oral Arabic as a linguistic model in the instruction of Arabic. Asian and African Studies, 15(1):3–17.

El-Beltagy, S. R. (2016). NileULex: A phrase and word level sentiment lexicon for Egyptian and Modern Standard Arabic. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2900–2905, Portorož, Slovenia. European Language Resources Association (ELRA).

El-Haj, M. (2020). Habibi - a multi dialect multi national Arabic song lyrics corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 1318–1326, Marseille, France. European Language Resources Association.

El-Haj, M., Rayson, P., and Aboelezz, M. (2018). Arabic dialect identification in the context of bivalency and code-switching. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

El Mekki, A., El Mahdaouy, A., Essefar, K., El Mamoun, N., Berrada, I., and Khoumsi, A. (2021). BERT-based multi-task model for country and province level MSA and dialectal Arabic identification. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 271–275, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

El-Yasin, M. K. (1985). Basic word order in Classical Arabic and Jordanian Arabic. Lingua, 65(1):107–122.

Elfardy, H. and Diab, M. (2012). Simplified guidelines for the creation of large scale dialectal Arabic annotations. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 371–378, Istanbul, Turkey. European Language Resources Association (ELRA).

Elfardy, H. and Diab, M. (2013). Sentence level dialect identification in Arabic. In Schuetze, H., Fung, P., and Poesio, M., editors, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 456–461, Sofia, Bulgaria. Association for Computational Linguistics.

Elmadany, A., Mubarak, H., and Magdy, W. (2018). An Arabic speech-act and sentiment corpus of tweets. In Proceedings of the Eleventh International Conference on

Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA). The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, OSACT3 ; Conference date: 08-05-2018.

Eltanbouly, S., Bashendy, M., and Elsayed, T. (2019). Simple but not naïve: Fine-grained Arabic dialect identification using only n-grams. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 214–218, Florence, Italy. Association for Computational Linguistics.

Emerson, L. H. S. and Ghanim, M. A. (1943). Aden Arabic: grammar. Al-Maaref Press, Aden.

Ennaji, M., Makhoukh, A., Es-saiydi, H., Moubtassime, M., and Slaoui, S. (2004). A grammar of Moroccan Arabic. Facultê des Lettres et des Sciences Humaines, Universitê Sidi Mohamed Ben Abdellah, Fês.

Ferguson, C. A. (1959). Diglossia. word, 15(2):325–340.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5):378.

Gaanoun, K., Naira, A. M., Allak, A., and Benelallam, I. (2024). DarijaBERT: a step forward in NLP for the written Moroccan dialect. International Journal of Data Science and Analytics.

Gamal-Eldin, S. (1968). A syntactic study of Egyptian colloquial Arabic / By Saad M. Gamal-Eldin. Janua linguarum. Series practica ; 34. Mouton, The Hague.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. Translation studies in Scandinavia, 1:88–95.

Ghoul, D. and Lejeune, G. (2019). MICHAEL: Mining character-level patterns for Arabic dialect identification (MADAR challenge). In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 229–233, Florence, Italy. Association for Computational Linguistics.

Graff, D., Buckwalter, T., Maamouri, M., and Jin, H. (2006). Lexicon development for varieties of spoken colloquial Arabic. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy. European Language Resources Association (ELRA).

Găman, M., Chifu, A.-G., Domingues, W., and Ionescu, R. T. (2023). Frecdo: A large corpus for french cross-domain dialect identification. Procedia Comput. Sci., 225(C):366–373.

Habash, N. (2010). Introduction to Arabic natural language processing, volume 3 of Synthesis Lectures on Human Language Technologies. Morgan and Claypool Publishers, 1st edition.

Habash, N., Diab, M., and Rambow, O. (2012a). Conventional orthography for dialectal Arabic. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 711–718, Istanbul, Turkey. European Language Resources Association (ELRA).

Habash, N., Eryani, F., Khalifa, S., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., Al-Shargi, F., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified guidelines and resources for Arabic dialect orthography. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Habash, N., Eskander, R., and Hawwari, A. (2012b). A morphological analyzer for Egyptian Arabic. In Cahill, L. and Albright, A., editors, Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Habash, N. and Rambow, O. (2005). Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In Knight, K., Ng, H. T., and Oflazer, K., editors, Proceedings of the 43rd Annual Meeting of the Association

for Computational Linguistics (ACL'05), pages 573–580, Ann Arbor, Michigan. Association for Computational Linguistics.

Habash, N. and Rambow, O. (2006). MAGEAD: A morphological analyzer and generator for the Arabic dialects. In Calzolari, N., Cardie, C., and Isabelle, P., editors, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 681–688, Sydney, Australia. Association for Computational Linguistics.

Habash, N., Rambow, O., Diab, M., and Kanjawi-Faraj, R. (2008). Guidelines for annotation of Arabic dialectness. In Proceedings of the LREC Workshop on HLT & NLP within the Arabic world, pages 49–53.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration, pages 15–22. Springer Netherlands, Dordrecht.

Hamed, I., Eryani, F., Palfreyman, D., and Habash, N. (2024). ZAEBUC-spoken: A multilingual multidialectal Arabic-English speech corpus. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 17770–17782, Torino, Italia. ELRA and ICCL.

Hamed, I., Habash, N., Abdennadher, S., and Vu, N. T. (2022). ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 119–130, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hamed, I., Sabty, C., Abdennadher, S., Vu, N. T., Solorio, T., and Habash, N. (2025). A survey of code-switched Arabic NLP: Progress, challenges, and future directions. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, Proceedings of the 31st International Conference on Computational Linguistics, pages 4561–4585, Abu Dhabi, UAE. Association for Computational Linguistics.

Hamed, I., Vu, N. T., and Abdennadher, S. (2020). ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In Calzolari, N., Béchet, F., Blache, P.,

Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 4237–4246, Marseille, France. European Language Resources Association.

Harrat, S., Meftouh, K., and Smaïli, K. (2017). Creating Parallel Arabic Dialect Corpus: Pitfalls to Avoid. In 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING), Budapest, Hungary.

Heylighen, F. and Dewaele, J.-M. (1999). Formality of language: definition, measurement and behavioral determinants. Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel, 4.

Ho, T. (2006). Tatoeba: Collection of sentences and translations. Available online, Accessed: 10 September 2023.

Holbrook, A. W. (1942). An elementary grammar of Egyptian colloquial Arabic. Al-Ettamad Printing Office, Cairo, first edition. edition.

Holes, C. (1995). Modern Arabic: Structures, Functions, and Varieties. London : Longman.

Holes, C. (2013). Word order and textual function in gulf arabic. In Information structure in spoken Arabic, pages 79–92. Routledge.

Huang, F. (2015). Improved Arabic dialect classification with social media data. In Màrquez, L., Callison-Burch, C., and Su, J., editors, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2118–2126, Lisbon, Portugal. Association for Computational Linguistics.

Issa, E., AlShakhori1, M., Al-Bahrani, R., and Hahn-Powell, G. (2021). Country-level Arabic dialect identification using RNNs with and without linguistic features. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 276–281, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Jamal, S., .Kassem, A. M., Mohamed, O., and Ashraf, A. (2022). On the Arabic dialects' identification: Overcoming challenges of geographical similarities between Arabic

dialects and imbalanced datasets. In Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 458–463, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019). Automatic language identification in texts: A survey. The Journal of Artificial Intelligence Research, 65:675–782. Copyright - © 2019. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the associated terms available at https://www.jair.org/index.php/jair/about; Last updated - 2024-01-17.

Jones, N. (2015). Artificial-intelligence institute launches free science search engine. Nature.

Kanjirangat, V., Samardzic, T., Dolamic, L., and Rinaldi, F. (2024). NLP_DI at NADI 2024 shared task: Multi-label Arabic Dialect Classifications with an Unsupervised Cross-Encoder. In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024).

Kanjirangat, V., Samardzic, T., Rinaldi, F., and Dolamic, L. (2022). Early guessing for dialect identification. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6417–6426, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karoui, A., Kammoun, R., Gharbi, F., Laouirine, I., and Bougares, F. (2024). ELYADATA at NADI 2024 shared task: Arabic Dialect Identification with Similarity-Induced Mono-to-Multi Label Transformation. In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024).

Kaye, A. S. and Rosenhouse, J. (1997). Arabic dialects and maltese. In Hetzron, R., editor, The Semitic Languages, Routledge Language Family Series, pages 263–311. Routledge, London & New York.

Kchaou, S., Bougares, F., and Hadrich-Belguith, L. (2019). LIUM-MIRACL participation in the MADAR Arabic dialect identification shared task. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing

Workshop, pages 219–223, Florence, Italy. Association for Computational Linguistics.

Keleg, A., Goldwater, S., and Magdy, W. (2023). ALDi: Quantifying the Arabic level of dialectness of text. In Bouamor, H., Pino, J., and Bali, K., editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10597–10611, Singapore. Association for Computational Linguistics.

Keleg, A., Goldwater, S., and Magdy, W. (2025). Revisiting common assumptions about Arabic dialects in NLP. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3309–3327, Vienna, Austria. Association for Computational Linguistics.

Keleg, A. and Magdy, W. (2023). Arabic dialect identification under scrutiny: Limitations of single-label classification. In Sawaf, H., El-Beltagy, S., Zaghouani, W., Magdy, W., Abdelali, A., Tomeh, N., Abu Farha, I., Habash, N., Khalifa, S., Keleg, A., Haddad, H., Zitouni, I., Mrini, K., and Almatham, R., editors, Proceedings of ArabicNLP 2023, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Keleg, A., Magdy, W., and Goldwater, S. (2024). Estimating the level of dialectness predicts inter-annotator agreement in multi-dialect Arabic datasets. In Ku, L.-W., Martins, A., and Srikumar, V., editors, Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 766–777, Bangkok, Thailand. Association for Computational Linguistics.

Kementchedjhieva, Y. (2016). Code-switching as strategically employed in political discourse. Lifespans and Styles, 2(1):3–9.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Al Kaabi, M. (2018). A morphologically annotated corpus of emirati Arabic. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Khered, A., Abdelhalim, I. A., and Batista-Navarro, R. (2022). Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. In

Bouamor, H., Al-Khalifa, H., Darwish, K., Rambow, O., Bougares, F., Abdelali, A., Tomeh, N., Khalifa, S., and Zaghouani, W., editors, Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP), pages 479–484, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kholy, A. E. and Habash, N. (2012). Orthographic and morphological processing for English-Arabic statistical machine translation. Machine Translation, 26(1/2):25–45.

Kiesling, S. F. (2011). Linguistic Variation and Change, volume 1. Edinburgh University Press.

Koppel, M. and Ordan, N. (2011). Translationese and its dialects. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Krippendorff, K. (2004). Content Analysis, an Introduction to Its Methodology. Thousand Oaks, CA: Sage Publications, second edition.

Kwaik, K. A., Saad, M., Chatzikyriakidis, S., and Dobnik, S. (2018). A lexical distance study of arabic dialects. Procedia Computer Science, 142:2–13. Arabic Computational Linguistics.

Lahiri, S. (2016). Squinky! a corpus of sentence-level formality, informativeness, and implicature.

Lahlali, M. (2011). The arab spring and the discourse of desperation: shifting from an authoritarian discourse to a "democratic one". The Journ. of Arab Media and Society.—Cairo: American Univ. in Cairo, (13).

Lahrouchi, M. (2018). The Amazigh influence on Moroccan Arabic: Phonological and morphological borrowing. The International Journal of Arabic Linguistics, 4(1):39–58.

Lichouri, M., Lounnas, K., Zahaf, B. N., and Rabiai, M. A. (2024). dzNLP at NADI 2024 Shared Task: Multi-Classifier Ensemble with Weighted Voting and TF-IDF Features. In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024).

Lopetegui, J. A., Riabi, A., and Seddah, D. (2025). Common ground, diverse roots: The difficulty of classifying common examples in Spanish varieties. In Scherrer, Y., Jauhiainen, T., Ljubešić, N., Nakov, P., Tiedemann, J., and Zampieri, M., editors, Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects, pages 168–181, Abu Dhabi, UAE. Association for Computational Linguistics.

Lulu, L. and Elnagar, A. (2018). Automatic Arabic dialect classification using deep learning models. Procedia Computer Science, 142:262–269. Arabic Computational Linguistics.

Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., and Tiedemann, J. (2016). Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In Nakov, P., Zampieri, M., Tan, L., Ljubešić, N., Tiedemann, J., and Malmasi, S., editors, Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

McNeil, K. (2018). Tunisian Arabic Corpus: Creating a Written Corpus of an 'Unwritten' Language, page 30–55. Edinburgh University Press.

McNeil, K. (2022). 'we don't speak the same language:'language choice and identity on a Tunisian internet forum. International Journal of the Sociology of Language, 2022(278):51–80.

McNeil, K. and Faiza, M. (2010-). Tunisian Arabic corpus (TAC): 895,000 words. Available online, Accessed: 10 September 2023.

Meermeier, R., Colbath, S., and Lillie, M. (2018). Portable speech-to-speech translation on an android smartphone: The MFLTS system. In Campbell, J., Yanishevsky, A., Doyon, J., and Jones, D., editors, Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track), pages 303–308, Boston, MA. Association for Machine Translation in the Americas.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In Zhao, H., editor, Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation, pages 26–34, Shanghai, China.

Messaoudi, A., Cheikhrouhou, A., Haddad, H., Ferchichi, N., BenHajhmida, M., Korched, A., Naski, M., Ghriss, F., and Kerkeni, A. (2022). TunBERT: Pretrained contextualized text representation for Tunisian dialect. In Bennour, A., Ensari, T., Kessentini, Y., and Eom, S., editors, Intelligent Systems and Pattern Recognition, pages 278–290, Cham. Springer International Publishing.

Mill, J. S. (1865). Dissertations and discussions : political, philosophical, and historical. W. V. Spencer, Boston.

Mohamed Eida, M., Nassar, M., and Dunn, J. (2024). How well do tweets represent sub-dialects of Egyptian Arabic? In Scherrer, Y., Jauhiainen, T., Ljubešić, N., Zampieri, M., Nakov, P., and Tiedemann, J., editors, Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 41–55, Mexico City, Mexico. Association for Computational Linguistics.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In Diab, M., Fung, P., Ghoneim, M., Hirschberg, J., and Solorio, T., editors, Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Mosquera, A. and Moreda, P. (2012). A qualitative analysis of informality levels in web 2.0 texts: The facebook case study. In Proceedings of the LREC workshop:@ NLP can u tag# user generated content, pages 23–29.

Mubarak, H. (2018). Dial2MSA: A tweets corpus for converting dialectal arabic to modern standard arabic. In Al-Khalifa, H., University, K. S., Magdy, K. W., of Edinburgh, U., Darwish, U. K., Institute, Q. C. R., Elsayed, Q. T., University, Q., and Qatar, editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

Mubarak, H. and Darwish, K. (2014). Using Twitter to collect a multi-dialectal corpus of Arabic. In Habash, N. and Vogel, S., editors, Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), pages 1–7, Doha, Qatar. Association for Computational Linguistics.

Mubarak, H. and Darwish, K. (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.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on Arabic social media. In Waseem, Z., Chung, W. H. K., Hovy, D., and Tetreault, J., editors, Proceedings of the First Workshop on Abusive Language Online, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Nabil, M., Aly, M., and Atiya, A. (2015). ASTD: Arabic sentiment tweets dataset. In Màrquez, L., Callison-Burch, C., and Su, J., editors, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

Nacar, O., Sibaee, S., I. Alharbi, A., Ghouti, L., and Koubaa, A. (2024). ASOS at NADI 2024 shared task: Bridging Dialectness Estimation and MSA Machine Translation for Arabic Language Enhancement. In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024).

Nayel, H., Hassan, A., Sobhi, M., and El-Sawy, A. (2021). Machine learning-based approach for Arabic dialect identification. In Habash, N., Bouamor, H., Hajj, H., Magdy, W., Zaghouani, W., Bougares, F., Tomeh, N., Abu Farha, I., and Touileb, S., editors, Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 287–290, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nieder, J. and List, J.-M. (2024). A computational model for the assessment of mutual intelligibility among closely related languages. In Hahn, M., Sorokin, A., Kumar, R., Shcherbakov, A., Otmakhova, Y., Yang, J., Serikov, O., Rani, P., Ponti, E. M., Muradoğlu, S., Gao, R., Cotterell, R., and Vylomova, E., editors, Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP, pages 37–43, St. Julian's, Malta. Association for Computational Linguistics.

Nigatu, H. H., Tonja, A. L., Rosman, B., Solorio, T., and Choudhury, M. (2024). The zeno's paradox of 'low-resource' languages. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 7022–7032, Marseille, France. European Language Resources Association.

Olsen, H., Touileb, S., and Velldal, E. (2023). Arabic dialect identification: An in-depth error analysis on the MADAR parallel corpus. In Sawaf, H., El-Beltagy, S., Zaghouani, W., Magdy, W., Abdelali, A., Tomeh, N., Abu Farha, I., Habash, N., Khalifa, S., Keleg, A., Haddad, H., Zitouni, I., Mrini, K., and Almatham, R., editors, Proceedings of ArabicNLP 2023, pages 370–384, Singapore (Hybrid). Association for Computational Linguistics.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Parkinson, D. B. (1991). Searching for modern Fus-ha: Real-life formal Arabic. Al-'Arabiyya, 24:31–64.

Pavlick, E. and Tetreault, J. (2016). An empirical analysis of formality in online communication. Transactions of the Association for Computational Linguistics, 4:61–74.

Peterson, K., Hohensee, M., and Xia, F. (2011). Email formality in the workplace: A case study on the Enron corpus. In Nagarajan, M. and Gamon, M., editors, Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 86–95, Portland, Oregon. Association for Computational Linguistics.

Přibáň, P. and Taylor, S. (2019). ZCU-NLP at MADAR 2019: Recognizing Arabic dialects. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural

Language Processing Workshop, pages 208–213, Florence, Italy. Association for Computational Linguistics.

Ragab, A., Seelawi, H., Samir, M., Mattar, A., Al-Bataineh, H., Zaghloul, M., Mustafa, A., Talafha, B., Freihat, A. A., and Al-Natsheh, H. (2019). Mawdoo3 AI at MADAR shared task: Arabic fine-grained dialect identification with ensemble learning. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 244–248, Florence, Italy. Association for Computational Linguistics.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In Walker, M., Ji, H., and Stent, A., editors, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Refaee, E. and Rieser, V. (2014). An Arabic Twitter corpus for subjectivity and sentiment analysis. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 2268–2273, Reykjavik, Iceland. European Language Resources Association (ELRA).

Robinson, N. R., Abdelmoneim, S., Marchisio, K., and Ruder, S. (2025). AL-QASIDA: Analyzing LLM quality and accuracy systematically in dialectal Arabic. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, Findings of the Association for Computational Linguistics: ACL 2025, pages 22048–22065, Vienna, Austria. Association for Computational Linguistics.

Rosenthal, S., Farra, N., and Nakov, P. (2017). SemEval-2017 task 4: Sentiment analysis in Twitter. In Bethard, S., Carpuat, M., Apidianaki, M., Mohammad, S. M., Cer, D., and Jurgens, D., editors, Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Sajjad, H., Abdelali, A., Durrani, N., and Dalvi, F. (2020). AraBench: Benchmarking

dialectal Arabic-English machine translation. In Scott, D., Bel, N., and Zong, C., editors, Proceedings of the 28th International Conference on Computational Linguistics, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sakr, A., Torki, M., and El-Makky, N. (2024). AlexUNLP-STM at NADI 2024 shared task: Quantifying the Arabic Dialect Spectrum with Contrastive Learning, Weighted Sampling, and BERT-based Regression Ensemble. In Proceedings of the Second Arabic Natural Language Processing Conference (ArabicNLP 2024).

Salama, A., Bouamor, H., Mohit, B., and Oflazer, K. (2014). YouDACC: the Youtube dialectal Arabic comment corpus. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1246–1251, Reykjavik, Iceland. European Language Resources Association (ELRA).

Salameh, M., Bouamor, H., and Habash, N. (2018). Fine-grained Arabic dialect identification. In Bender, E. M., Derczynski, L., and Isabelle, P., editors, Proceedings of the 27th International Conference on Computational Linguistics, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Salloum, W. S. (2018). Machine Translation of Arabic Dialects. PhD thesis, Columbia University.

Samih, Y., Mubarak, H., Abdelali, A., Attia, M., Eldesouki, M., and Darwish, K. (2019). QC-GO submission for MADAR shared task: Arabic fine-grained dialect identification. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 290–294, Florence, Italy. Association for Computational Linguistics.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2022). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Schechter Vera, H., Dua, S., Zhang, B., Salz, D., Mullins, R., Raghuram Panyam, S., Smoot, S., Naim, I., Zou, J., Chen, F., Cer, D., Lisak, A., Choi, M., Gonzalez, L., Sanseviero, O., Cameron, G., Ballantyne, I., Black, K., Chen, K., Wang, W., Li, Z., Martins, G., Lee, J., Sherwood, M., Ji, J., Wu, R., Zheng, J., Singh, J., Sharma, A., Sreepat, D., Jain, A., Elarabawy, A., Co, A., Doumanoglou, A., Samari, B., Hora, B., Potetz, B., Kim, D., Alfonseca, E., Moiseev, F., Han, F., Palma Gomez, F., Hernández Ábrego, G., Zhang, H., Hui, H., Han, J., Gill, K., Chen, K., Chen, K., Shanbhogue, M., Boratko, M., Suganthan, P., Duddu, S. M. K., Mariserla, S., Ariafar, S., Zhang, S., Zhang, S., Baumgartner, S., Goenka, S., Qiu, S., Dabral, T., Walker, T., Rao, V., Khawaja, W., Zhou, W., Ren, X., Xia, Y., Chen, Y., Chen, Y.-T., Dong, Z., Ding, Z., Visin, F., Liu, G., Zhang, J., Kenealy, K., Casbon, M., Kumar, R., Mesnard, T., Gleicher, Z., Brick, C., Lacombe, O., Roberts, A., Sung, Y., Hoffmann, R., Warkentin, T., Joulin, A., Duerig, T., and Seyedhosseini, M. (2025). EmbeddingGemma: Powerful and lightweight text representations.

Schmitt, G. A. (2020). Relevance of Arabic dialects: A brief discussion. Handbook of the changing world language map, pages 1383–1398.

S'hiri, S. (2013). Speak Arabic Please!: Tunisian Arabic Speakers' Linguistic Accommodation to Middle Easterners, pages 149–174. Taylor and Francis. Publisher Copyright: © 2002 Aleya Rouchdy.

Shlonsky, U. (1997). Clause structure and word order in Hebrew and Arabic : an essay in comparative Semitic syntax. Oxford studies in comparative syntax. Oxford University Press, New York ;.

Shoemark, P., Sur, D., Shrimpton, L., Murray, I., and Goldwater, S. (2017). Aye or naw, whit dae ye hink? Scottish independence and linguistic identity on social media. In Lapata, M., Blunsom, P., and Koller, A., editors, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 1239–1248, Valencia, Spain. Association for Computational Linguistics.

Shon, S., Ali, A., Samih, Y., Mubarak, H., and Glass, J. (2020). ADI17: A fine-grained Arabic dialect identification dataset. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 8244–8248.

Søgaard, A., Ebert, S., Bastings, J., and Filippova, K. (2021). We need to talk about random splits. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1823–1832, Online. Association for Computational Linguistics.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In Diab, M., Hirschberg, J., Fung, P., and Solorio, T., editors, Proceedings of the First Workshop on Computational Approaches to Code Switching, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Talafha, B., Ali, M., Za'ter, M. E., Seelawi, H., Tuffaha, I., Samir, M., Farhan, W., and Al-Natsheh, H. (2020). Multi-dialect Arabic BERT for country-level dialect identification. In Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., and Zaghouani, W., editors, Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 111–118, Barcelona, Spain (Online). Association for Computational Linguistics.

Talafha, B., Fadel, A., Al-Ayyoub, M., Jararweh, Y., AL-Smadi, M., and Juola, P. (2019). Team JUST at the MADAR shared task on Arabic fine-grained dialect identification. In El-Hajj, W., Belguith, L. H., Bougares, F., Magdy, W., Zitouni, I., Tomeh, N., El-Haj, M., and Zaghouani, W., editors, Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 285–289, Florence, Italy. Association for Computational Linguistics.

Terner, O., Bar, K., and Dershowitz, N. (2020). Transliteration of Judeo-Arabic texts into Arabic script using recurrent neural networks. In Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., and Zaghouani, W., editors, Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 85–96, Barcelona, Spain (Online). Association for Computational Linguistics.

Tiedemann, J. and Ljubešić, N. (2012). Efficient discrimination between closely related languages. In Kay, M. and Boitet, C., editors, Proceedings of COLING 2012, pages 2619–2634, Mumbai, India. The COLING 2012 Organizing Committee.

Tobaili, T. (2016). Arabizi identification in Twitter data. In He, H., Lei, T., and Roberts,

W., editors, Proceedings of the ACL 2016 Student Research Workshop, pages 51–57, Berlin, Germany. Association for Computational Linguistics.

Touileb, S. (2020). LTG-ST at NADI shared task 1: Arabic dialect identification using a stacking classifier. In Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., and Zaghouani, W., editors, Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 313–319, Barcelona, Spain (Online). Association for Computational Linguistics.

Uma, A. N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., and Poesio, M. (2021). Learning from disagreement: A survey. Journal of Artificial Intelligence Research, 72:1385–1470.

Wainer, H., Gessaroli, M., and Verdi, M. (2006). Visual revelations. CHANCE, 19(1):49–52.

Walters, L., Chisadza, C., and Clance, M. (2023). The effect of pre-colonial ethnic institutions and european influences on contemporary education in sub-saharan africa. The Journal of Development Studies, 59(10):1469–1490.

Weiss, B., Schlenoff, C., Sanders, G., Steves, M., Condon, S., Phillips, J., and Parvaz, D. (2008). Performance evaluation of speech translation systems. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Younes, J., Achour, H., and Souissi, E. (2015). Constructing linguistic resources for the Tunisian dialect using textual user-generated contents on the social web. In Daniel, F. and Diaz, O., editors, Current Trends in Web Engineering, pages 3–14, Cham. Springer International Publishing.

Zaghouani, W. and Charfi, A. (2018). Arap-tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Zaidan, O. F. and Callison-Burch, C. (2011). The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. Computational Linguistics, 40(1):171–202.

Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., and Aepli, N. (2017). Findings of the VarDial evaluation campaign 2017. In Nakov, P., Zampieri, M., Ljubešić, N., Tiedemann, J., Malmasi, S., and Ali, A., editors, Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J., Scherrer, Y., Samardžić, T., Ljubešić, N., Tiedemann, J., van der Lee, C., Grondelaers, S., Oostdijk, N., Speelman, D., van den Bosch, A., Kumar, R., Lahiri, B., and Jain, M. (2018). Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In Zampieri, M., Nakov, P., Ljubešić, N., Tiedemann, J., Malmasi, S., and Ali, A., editors, Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zampieri, M., North, K., Jauhiainen, T., Felice, M., Kumari, N., Nair, N., and Bangera, Y. M. (2024). Language variety identification with true labels. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 10100–10109, Torino, Italia. ELRA and ICCL.

Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J. (2014). A report on the DSL shared task 2014. In Zampieri, M., Tan, L., Ljubešić, N., and Tiedemann, J., editors, Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of Arabic dialects. In Fosler-Lussier, E., Riloff, E., and Bangalore, S., editors, Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).