ASOS at NADI 2024 shared task: Bridging Dialectness Estimation and MSA Machine Translation for Arabic Language Enhancement

Omer Nacar[1,*], Serry Sibaee[1], Abdullah I. Alharbi[2], Lahouari Ghouti[1], Anis Koubaa[1]
[1]Robotics and Internet-of-Things Lab, Prince Sultan University, Riyadh 12435, Saudi Arabia
[2]Faculty of Computing and Information Technology in Rabigh
King Abdulaziz University, Jeddah, Saudi Arabia
{onajar, ssibaee, lghouti, akoubaa}@psu.edu.sa, aamalharbe@kau.edu.sa
***Correspondence:** onajar@psu.edu.sa

Abstract

This study undertakes a comprehensive investigation of transformer-based models to advance Arabic language processing, focusing on two pivotal aspects: the estimation of Arabic Level of Dialectness and dialectal sentence-level machine translation into Modern Standard Arabic. We conducted various evaluations of different sentence transformers across a proposed regression model, showing that the MARBERT transformer-based proposed regression model achieved the best root mean square error of 0.1403 for Arabic Level of Dialectness estimation. In parallel, we developed bi-directional translation models between Modern Standard Arabic and four specific Arabic dialects—Egyptian, Emirati, Jordanian, and Palestinian—by fine-tuning and evaluating different sequence-to-sequence transformers. This approach significantly improved translation quality, achieving a BLEU score of 0.1713. We also enhanced our evaluation capabilities by integrating MSA predictions from the machine translation model into our Arabic Level of Dialectness estimation framework, forming a comprehensive pipeline that not only demonstrates the effectiveness of our methodologies but also establishes a new benchmark in the deployment of advanced Arabic NLP technologies.

1 Introduction

Dialectal Arabic (DA) is essential in daily interactions for over 422 million Arabic speakers, significantly influencing casual conversations and digital communications, including social media (Elnagar et al., 2021; Diab and Habash, 2007). Despite Modern Standard Arabic (MSA) being used formally in education, official, and media contexts, DA's regional variations and frequent code-switching with MSA present challenges for text analysis and computational processing, highlighting DA's cultural and linguistic importance (Harrat et al., 2018). The rise of digital communication has increased the use of DA, driving significant NLP research to better understand and process it, thus bridging the gap between MSA's formal structure and DA's nuances (Shoufan and Alameri, 2015).

This paper addresses two crucial aspects of Arabic NLP: the estimation of the level of dialectness (ALDi) and the translation of Arabic dialect to MSA. Building on the shared tasks provided by (Abdul-Mageed et al., 2024) and previous years' tasks (Abdul-Mageed et al., 2022, 2023), we explore innovative methodologies to enhance the computational handling of Arabic's dual nature. Integrating ALDi estimation with dialect-to-MSA translation not only advances our understanding of Arabic's linguistic duality but also improves practical applications in processing its informal variants.

Recent advancements in machine translation focus on translating Arabic dialects to MSA. However, much work, such as (Al-Ibrahim and Duwairi, 2020; Baniata et al., 2018), has centered on single dialects using limited datasets, limiting their effectiveness across the Arabic dialect spectrum. Notably, (Hamed et al., 2022) enhanced NMT for the Algerian dialect using transductive transfer learning. Additionally, (Al-Khalifa et al., 2022) developed Turjuman, leveraging the AraT5 model for multi-dialectal translation, demonstrating the benefits of scalable frameworks.

In parallel, the conceptualization of ALDi marks a significant shift in addressing dialect variation in Arabic text. Previous works have focused on dialect estimation at both the sentence level (Zaidan and Callison-Burch, 2014; Elfardy and Diab, 2013) and token level (Solorio et al., 2014; Molina et al., 2019), often viewing the distinction between MSA and DA as binary. Earlier studies (Habash et al., 2008; Zaidan and Callison-Burch, 2011) recognized the spectrum of dialectness and emphasized standard annotation guidelines for identifying dialect switching. Recently, (Keleg et al., 2023) ad-

Proceedings of The Second Arabic Natural Language Processing Conference, pages 748–753

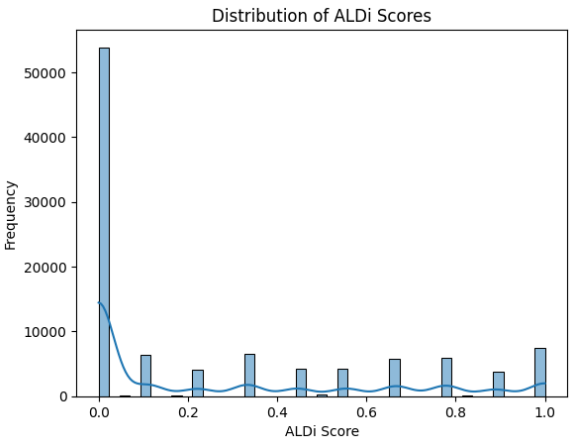

Figure 1: Aldi Estimation Training dataset distribution

vanced this understanding by releasing a dataset of 127,835 Arabic comments with ALDi labels and proposing a method for estimating ALDi that can generalize to other dialect corpora.

Building on these foundations, this study introduces a transformative approach that bridges ALDi estimation with DA-to-MSA translation. Our contributions are threefold:

ALDi Estimation Model, introducing a transformer-based regression model that sets new benchmarks in ALDi estimation, offering a sophisticated tool for analyzing dialectal nuances in Arabic text.

Dialect-Specific Translation Models, developing advanced bi-directional translation models (MSA2Dialect and Dialect2MSA), fine-tuned to handle the specificities of individual Arabic dialects, enhancing translation accuracy.

Evaluation Pipeline, establishing a novel evaluation pipeline that utilizes machine translation outputs to refine ALDi estimations, illustrating the potential of integrating these two critical aspects of Arabic NLP.

The remainder of this paper details our work, starting with the dataset and our methodological approach, followed by the system setup, results with discussion, and finally a conclusion reflecting on the impact with future directions of our research.

2 Dataset

For the first subtask focused on the ALDi estimation, the dataset utilized consisted solely of the training set provided by NADI 2024 (Abdul-Mageed et al., 2023). This dataset consists of 100k sentences and it is specifically tailored to evaluate the ALDi across various Arabic texts. A distribution plot of the ALDi scores from this training dataset is included in Figure 2, which illustrates the frequency of scores across different intervals. The distribution of ALDi scores, as depicted in Figure 2, shows a significant concentration at lower scores with less frequent occurrences across higher values, indicating a prevalence of text closely aligned with MSA.

Dataset	Egy	Emi	Jor	Pal
Mono	20k	20k	20k	20k
MADAR	18k	-	4k	4k
Total	38k	20k	24k	24k

Table 1: Distribution of data by dialect source, showing numbers from Monolingual and MADAR datasets.

For the dialect to MSA subtask, a more involved dataset preparation strategy was employed. To create a robust and dialect-specific training environment, 20,000 sentences from a monolingual MSA dataset provided by NADI 2024 (Abdul-Mageed et al., 2023) were generated for each of the target dialects: Egyptian, Emirati, Jordanian, and Palestinian by using dialect-specific trained machine translation models. Additionally, the MADAR dataset (Bouamor et al., 2018) was utilized selectively for each dialect to include region-specific variations. This ensured that each dialect's dataset was finely tuned for optimal translation performance. The composition of these datasets is summarized in Table 1, showing the distribution of data, which was critical for the success of the translation task.

3 System

Our framework consists of an integrated pipeline developed to advance the machine translation of Arabic dialects to MSA and to estimate the Level of Dialectness. The system is described in Figure 2 showing the integration of the machine translation model with the ALDi estimation regression model. As shown in Figure 2, the initial phase of our pipeline involves translating MSA to specific Arabic dialects using the monolingual MSA dataset, which contains 20,000 sentences for each target dialect. This is part of a data augmentation strategy intended to enhance the training datasets for subsequent translation tasks.

Following this, we integrate dialect-specific datasets from the MADAR corpus (Bouamor et al., 2018), which provides additional contextual diver-

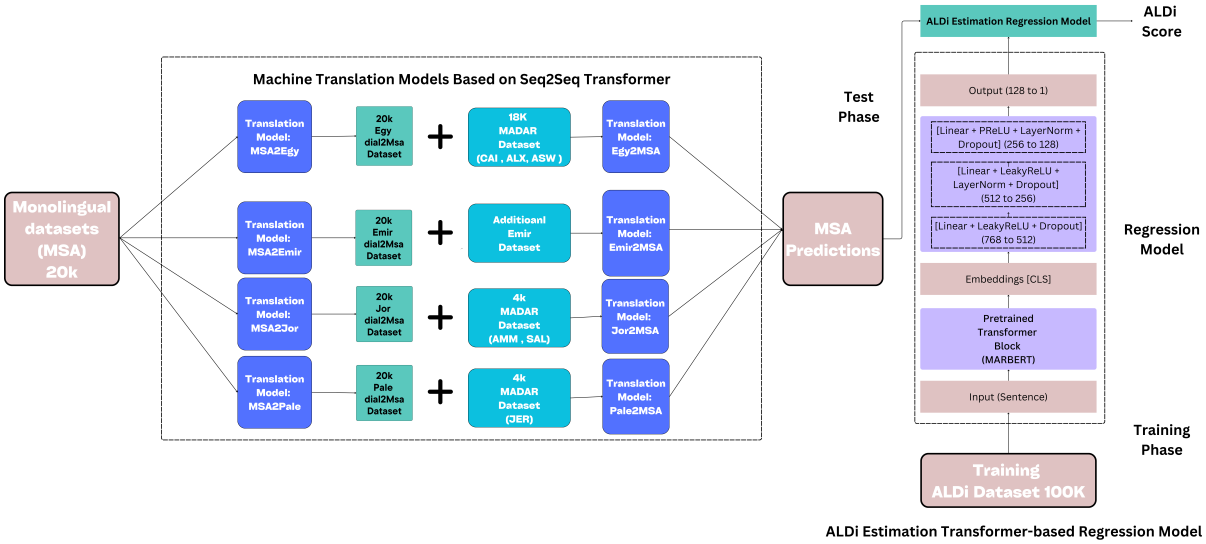

Figure 2: Comprehensive Pipeline Showcasing ALDi Estimation and Dialect-to-MSA Translation Tasks

sity and linguistic features pertinent to regional Arabic dialects like Egyptian (Cairo, Alexandria, Aswan), Jordanian (Amman, Salt), and Palestinian (Jerusalem). Each of these enriched datasets is then used to fine-tune separate DA-MSA models for each dialect using the AraT5-V2 (Elmadany et al., 2022) model, thereby generating MSA predictions.

The output from our machine translation models, termed 'MSA predictions' feeds into the ALDi estimation model as a test phase. The regression model is a MARBERT transformer based model (Mageed et al., 2021), where the architecture pretrained specifically for the Arabic language. The proposed regression model as shown in Figure 2 is designed to estimate the level of dialectness as a single percentage value, reflecting how closely a sentence aligns with MSA. The pipeline enhances Arabic dialect understanding by automating parallel corpora construction and serving as a critical evaluation framework for machine translation models and ALDi estimation approaches.

4 Result

This section demonstrates the results of our comprehensive experiments conducted to evaluate the effectiveness of our machine translation models from various Arabic dialects to MSA and vice versa, as well as the performance of our ALDi estimation model.

MSA to Dialect Models, the MSA to Dialect (MSA2Dia) models were trained using a monolingual dataset to increase the volume of the training data for the dialect to MSA training. The validation BLEU scores for each dialect are shown in Figure 3.

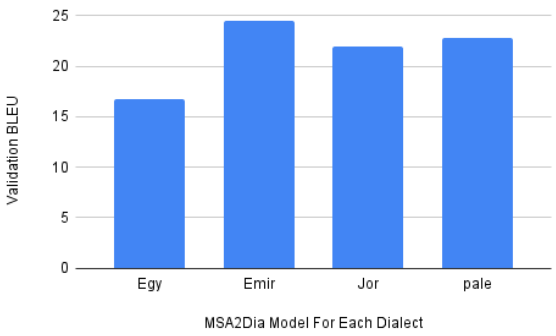

Figure 3: Validation BLEU of MSA2Dia Model For Each Dialect

The results depicted in Figure 3 showcase the performance of the MSA to Dialect (MSA2Dia) models across various dialects. The Emirati dialect model achieved the highest validation BLEU score, closely followed by the Palestinian and Jordanian models, with the Egyptian model trailing. The models' strong BLEU scores indicate that each dialect's training data was well-represented, which enabled the models to provide dialect-specific translations from Modern Standard Arabic (MSA).

Dialect to MSA Models, the Dialect to MSA (Dia2MSA) translation models discussed here represent an effort to convert various Arabic dialects back to Modern Standard Arabic (MSA). This process involved using a generated monolingual dataset combined with existing datasets as shown

No. of Submission	Model	Overall Dev BLEU	Egy	Emir	Jor	Pal
1	AraT5 V2 Model	0.2874	0.1415	0.6851	0.1520	0.1665
2	Turjman Model	0.2112	0.1293	0.4662	0.1095	0.1350
3	mt5 Model	0.1134	0.1002	0.2107	0.0662	0.0809

Table 2: Dialect to MSA (Dia2MSA) development results.

No. of Submission	Model	Overall Test BLEU	Egy	Emir	Jor	Pal
1	AraT5 V2 Model	0.1713	0.1482	0.1939	0.1580	0.1838
2	Turjman Model	0.1560	0.1163	0.0966	0.1293	0.1331
3	mt5 Model	0.1377	0.0935	0.0909	0.1164	0.1075

Table 3: Dialect to MSA (Dia2MSA) test results

in Figure 2 in order to fine-tune different sequence-to-sequence models, leading to the results captured in the Tables 2 and 3 which present the BLEU scores for three different models: AraT5 V2, Turjman, and mt5 for the development phase and test phase of the shared task respectively.

As shown in Tables 2 and 3, three models, AraT5 V2[1], Turjman[2], and mt5[3], were tested for their ability to translate Arabic dialects back to Modern Standard Arabic. AraT5 V2 showed the highest proficiency, particularly with the Emirati dialect, with overall BLEU scores of 28.74 and 17.13 on dev and test sets, respectively. Turjman and mt5 models had lower scores, with mt5 struggling across all dialects, indicating limitations in adaptability and generalization capabilities.

ALDi Estimation, we evaluated various transformer-based models on the development dataset using a proposed regression model. The development phase results, as depicted in Table 4, showed that the MARBERT model[4] outperformed others with the lowest Root Mean Square Error (RMSE) of 0.12801, indicating its superior accuracy in estimating dialectal levels. All models, including MARBERT, were fine-tuned without freezing the transformer layers. Training was conducted for 1 epoch with a batch size of 32, utilizing an 80/20 split for training and validation. Early stopping was applied with a patience of 10 steps. Other models like ARBERT[5], AraElectra[6], and AraBert[7] followed, with RMSE scores gradually increasing.

The test phase as shown in Table 4, with the MARBERT-based regression model showing the best performance, resulting in an RMSE of 0.14031, which although slightly higher than in the development phase, still indicates reliability through unseen examples. Comparatively, the ARBERT-based model, which performed better during development than some other models, exhibited a significantly higher RMSE of 0.30607 during testing, suggesting potential issues with generalization or overfitting. The AraElectra model maintained a moderate level of accuracy with an RMSE of 0.18266. The study shows that MARBERT is the most effective model for estimating ALDi, accurately distinguishing MSA from its dialectal variations.

MSA Predictions for ALDi Estimation, we proposed an evaluation framework designed to assess the effectiveness of Dialect to MSA (Dia2MSA) models by leveraging outputs from these models as input for the proposed ALDi estimation model. Figure 4 shows the histogram of the distribution of the ALDi scores for the MSA predictions produced by our best model in Table 4, using a bin size of 30. We used the best-performing model due to its significantly lower RMSE compared to other models.

The distribution depicted in Figure 4 illustrates that the majority of translations by the Dia2MSA models are closely aligned with MSA, as indicated by the distribution primarily clustering around lower ALDi scores, with a significant peak at 0.1. This skew towards lower scores suggests a high level of effectiveness in the Dia2MSA models' ability to produce outputs that mirror the characteristics of MSA.

Error Analysis, an error analysis of the ALDi score estimation model which quantifies dialec-

[1]https://huggingface.co/UBC-NLP/AraT5v2-base-1024
[2]https://huggingface.co/UBC-NLP/turjuman
[3]https://huggingface.co/google/mt5-base
[4]https://huggingface.co/UBC-NLP/MARBERT
[5]https://huggingface.co/UBC-NLP/ARBERT
[6]https://huggingface.co/aubmindlab/araelectra-base-discriminator
[7]https://huggingface.co/aubmindlab/bert-base-arabertv2

No. of Submission	Model	Dev RMSE	Test RMSE
1	MARBERT	0.12801	0.14031
2	ARBERT	0.14972	0.30607
3	AraElectra	0.16987	0.18266
4	AraBert	0.18553	–
5	Paraphrase-multilingual-MiniLM-L12-v2	0.23829	–

Table 4: ALDi Estimation RMSE results on the development and test datasets.

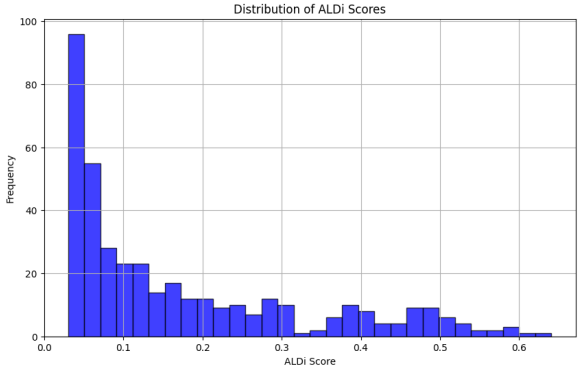

Figure 4: Distribution of ALDi scores from the best-performing model for ALDi estimation.

tal content in Arabic text is implemented. The model was evaluated using RMSE across various segments determined by the data percentiles, revealing how model accuracy varied with different levels of dialectal content. The RMSE distribution is shown in Figure 5, illustrating the model's performance across segmented score ranges.

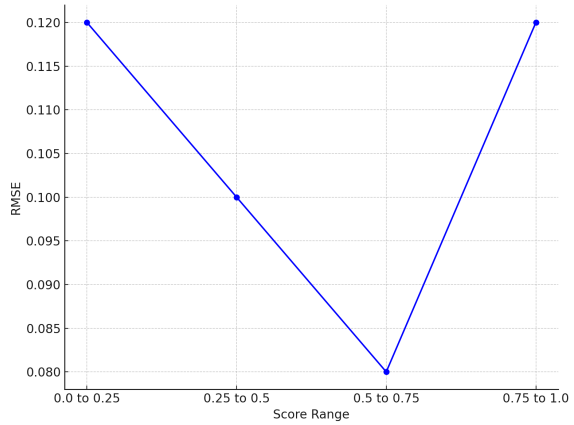

Figure 5: RMSE Distribution Across ALDi Score Segments

The model excels in regions with moderate dialectal content with RMSE of 0.08669 but struggles at the extremes, a challenge likely increased by the lack of representation of higher dialectal scores in the training set. To enhance model performance, we will work on enriching the training dataset with more examples of medium to high dialectal content, retraining the model for better generalization across the dialect spectrum, and employing advanced techniques like transfer learning for improved handling of dialectal nuances.

5 Conclusion

Our work makes a significant contribution to the NADI2024 shared task by employing and evaluating transformer-based models to tackle two critical aspects: the estimation of ALDi and the MT of dialectal sentences into MSA. We rigorously tested various sentence transformers and found that the MARBERT transformer-based regression model performed exceptionally well, achieving an RMSE of 0.1403 in ALDi estimation on unlabeled test data. Additionally, we developed and refined bi-directional translation models between MSA and four specific Arabic dialects. This not only expanded our dataset but also significantly improved translation quality, as evidenced by a BLEU score of 0.1713. Integrating MSA predictions from the MT model into our ALDi estimation framework established a comprehensive pipeline that effectively captured the spectrum of dialectal variation within Arabic text. These advancements set a new benchmark for deploying advanced Arabic NLP technologies, highlighting their potential in nuanced linguistic analyses.

Acknowledgments

The authors thank Prince Sultan University for their support.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The fourth nuanced Arabic dialect identification shared task. In *Proceedings of ArabicNLP 2023*, pages 600–

613, Singapore (Hybrid). Association for Computational Linguistics.

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 (ArabicNLP 2024)*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2022. Nadi 2022: The third nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2210.09582*.

Roqayah Al-Ibrahim and Rehab M Duwairi. 2020. Neural machine translation from jordanian dialect to modern standard arabic. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 173–178. IEEE.

Hend Al-Khalifa, Tamer Elsayed, Hamdy Mubarak, Abdulmohsen Al-Thubaity, Walid Magdy, and Kareem Darwish. 2022. Proceedinsg 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. In *Proceedinsg 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*.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational Intelligence & Neuroscience*.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Mona Diab and Nizar Habash. 2007. Arabic dialect processing tutorial. In *Proceedings of the human language technology conference of the NAACL, companion volume: tutorial abstracts*, pages 5–6.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2022. Arat5: Text-to-text transformers for arabic language generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 628–647.

Ashraf Elnagar, Sane M Yagi, Ali Bou Nassif, Ismail Shahin, and Said A Salloum. 2021. Systematic literature review of dialectal arabic: identification and detection. *IEEE Access*, 9:31010–31042.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. Investigating lexical replacements for arabic-english code-switched data augmentation. *arXiv preprint arXiv:2205.12649*.

Salima Harrat, Karima Meftouh, and Kamel Smaïli. 2018. Maghrebi arabic dialect processing: an overview. *Journal of International Science and General Applications*, 1.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. Aldi: Quantifying the arabic level of dialectness of text. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Muhammad Abdul Mageed, Abdelrahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *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.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2019. Overview for the second shared task on language identification in code-switched data. *arXiv preprint arXiv:1909.13016*.

Abdulhadi Shoufan and Sumaya Alameri. 2015. Natural language processing for dialectical arabic: A survey. In *Proceedings of the second workshop on Arabic natural language processing*, pages 36–48.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the first workshop on computational approaches to code switching*, pages 62–72.

Omar Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.