The *KIND* Dataset: A Social Collaboration Approach for Nuanced Dialect Data Collection

Asma Z. Yamani[1], Raghad Alziyady[1], Reem AlYami[5], Salma A. Albelali[1,6], Leina Abouhagar[1], Jawharah Almulhim[7], Amjad Alsulami[1], Motaz Alfarraj[1,3,4], and Rabeah Al-Zaidy[1,2]

[1]Department of Information and Computer Science, King Fahd University of Petroleum & Minerals, Saudi Arabia
[2]Center for Integrative Petroleum Research, CIPR, King Fahd University of Petroleum & Minerals, Saudi Arabia
[3]Department of Electrical Engineering, King Fahd University of Petroleum & Minerals, Saudi Arabia
[4]SDAIA-KFUPM Joint Research Center for AI, King Fahd University of Petroleum & Minerals, Saudi Arabia
[5]Preparatory Year Program,King Fahd University of Petroleum & Minerals, Saudi Arabia
[6]Department of Computer Science, Imam Abdulrahman Bin Faisal University, Saudi Arabia
[7]Saudi Data & AI Authority, Saudi Arabia
{g201906630,g202009020,reem.yami,g201907430,g202101130,motaz}@kfupm.edu.sa
{linah.a.abuhajar,rabeah.alzaidy}@gmail.com, jalmulhim@sdaia.gov.sa

Abstract

Nuanced dialects are a linguistic variant that pose several challenges for NLP models and techniques. One of the main challenges is the limited amount of datasets to enable extensive research and experimentation. We propose an approach for efficiently collecting nuanced dialectal datasets that are not only of high quality, but are versatile enough to be multipurpose as well. To test our approach we collect the KIND corpus, which is a collection of fine-grained Arabic dialect data. The data is short texts, and unlike many nuanced dialectal datasets, it is curated manually through social collaboration efforts as opposed to being crawled from social media. The collaborative approach is incentivized through educational gamification and competitions for which the community itself benefits from the open source dataset. Our approach aims to achieve: (1) coverage of dialects from under-represented groups and fine-grained dialectal varieties, (2) provide aligned parallel corpora for translation between Modern Standard Arabic (MSA) and multiple dialects to enable translation and comparison studies, (3) promote innovative approaches for nuanced dialect data collection. We explain the steps for the competition as well as the resulting datasets and the competing data collection systems. The KIND dataset is shared with the research community.

1 Introduction

The Arabic language is one of the most spoken languages in the world with over 400 million speakers from more than 30 countries (Wikipedia, 2023) and has gained wide attention in natural language processing advancements recently. Since most linguistic technologies rely on high quality training data, Arabic data collection is, consequently, becoming the focus of an increasing number of studies. Although a wide range of these studies propose effective approaches for Arabic data collection, the need for large scale, high-quality datasets of nuanced dialect variations is constantly increasing with the demands of domain specific applications as well as large language models.

A main challenge facing NLP technologies in Arabic is the diversity of Arabic dialects, with more than 30 modern dialects across the Arab region, and over 20 documented dialects in Saudi Arabia, the largest country in the Arabian peninsula (Wikipedia, 2023; Aldarsoni, 2013). This challenge is most pronounced in downstream applications that involve machine translation from dialectal Arabic to other languages. The multitude of nuanced or explicit varying synonyms and hyponyms in Arabic dialects have high impact on the quality of translation models. Several approaches were proposed to take advantage of MSA as the universal formal Arabic and ground other dialects to it through translation or similarity analyses. As part of these studies datasets of parallel dialectal and MSA texts were proposed (Harrat et al., 2015; Salloum and Habash, 2011; Zbib et al., 2012). Although very useful, these datasets remain limited in size and dialectal coverage. To address this gap we propose, KIND: King Fahd University of Petroleum and Minerals (KFUPM) *In Your Dialect*

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 32–43

approach, a multiple-tasks competition for obtaining short texts of parallel corpora of fine-grain Arabic dialectal data and question and answer pairs.

The approach aims not only to promote innovative approaches for data collection, but also to raise awareness about the significance of representation in spoken language-based technologies. The competition is accompanied by social educational initiatives to raise awareness about linguistic technologies and to encourage the public to participate in linguistic data collection competitions. Social engagement was solicited by demonstrating the effect of sharing their dialectal utterances on the quality of technologies that will result from their aggregate contributions.

Our approach is comprised of two general tasks. The first is a data collection task where participants compete to enter the largest amount of entries for (1) translation of an MSA sentence to their dialect, and (2) answering an open-ended question in their dialect. The resulting data from this task is denoted the KIND corpus. The second task is to solicit innovative systems that can compete with our approach for task 1. Our approach follows a hackathon format for developing dialectal Arabic data collection systems.

In this paper we describe the approach for collecting the KIND dataset and the quality requirements of the submissions. We demonstrate the effectiveness of our approach by describing the resulting corpora of high–quality training data. The dataset is suitable for training language models, machine translation tasks, as well as Q&A tasks with the respective dialects labeled to a fine granularity level. We make this dataset publicly available to the community along with the labels for each dialect.

We summarize the contributions of this paper as follows:

1. Propose the design and process of a nuanced dialect data collection system that addresses coverage of dialects from under-represented Arabic speaking groups in addition to fine-grained dialectal varieties.
2. An open-source corpora of aligned parallel texts for translation between Modern Standard Arabic (MSA) to multiple nuanced dialects and between the dialects as well as an Arabic dialect Q&A dataset.
3. A collection of proposed systems for nuanced dialectal data collection.

The remainder of this paper is organized as follows. Section 2 provides an overview of the related literature. Section 3 describes the design and process of our collection approach. Section 4 describes the results of the first task of the collection approach. Section 5 describes the results of the second task of the approach. In section 6 we describe the resources resulting from this study. In section 7 we conclude.

2 Related Work

2.1 Arabic Dialect Datasets

The emergence of different social media platforms increased the use of informal forms of a language. That showed a discrepancy in the levels of support for basic tasks in language technologies for different languages. For example, the lack of keyboard support and spell checking for low-resource languages, although there is a desire among the speakers of these regional languages to use these digital services (Soria et al., 2018; Ruder, 2020).

The ability to thoroughly and effectively evaluate and asses the performance of a system is paramount for the development of advanced NLP technologies. The availability of benchmarks and standardized datasets for quality assessment is essential for this evaluation process. For many languages, including Arabic, the availability of these benchmark datasets is minimal compared to other languages such as English (Zampieri and Nakov, 2021). In English, there are various benchmark datasets to perform different NLP tasks, for instance, *SuperGLUE* and *SQuAD*; the former provides nine natural language understanding tasks, and the latter provides question-answering task (Wang et al., 2019; Rajpurkar et al., 2018). However, when looking at Arabic dialects, corpora, and annotated corpora remain minimal compared to MSA (Althobaiti, 2020; Zampieri and Nakov, 2021). Although various efforts focused on dialectal Arabic and building resources for it (Abdul-Mageed et al., 2020; Bouamor et al., 2018, 2019; Diab et al., 2014; Zaidan and Callison-Burch, 2014), there still remain nuanced dialects of many groups that are still either under-represented or not represented at all. In this work we aim to propose an approach that is capable of leveraging a single user prompt/entry to serve as a training record for as many NLP tasks as possible without compromising functionality.

In the general sense, Arabic dialect datasets continue to exhibit limitations concerning their size,

scope, and the extent of annotation when compared to MSA and other languages, as highlighted in previous studies (Althobaiti, 2020; Zampieri and Nakov, 2021). For instance, the MADAR dataset, which covers dialectal variations across 25 cities, offers valuable insights; however, it is noteworthy that this dataset is primarily a translation from another language within the travel domain. Consequently, the source origin imposes constraints on the cultural and domain diversity represented in the text (Takezawa et al., 2007; Bouamor et al., 2019). Therefore our approach aims to incorporate semantic cultural relevance in the design process of the data collection.

2.2 Arabic Dialect Granularity Levels

Arabic is one of the low-resource languages with rich morphology. It has different varieties; formal Arabic MSA is taught in schools and used in formal venues, whereas informal Arabic is used in daily life interactions. The differences between Arabic nuanced dialects and MSA pose a serious challenge when working on Arabic varieties (Zampieri and Nakov, 2021). The difference between MSA-Dialectal Arabic and Dialectal Arabic-Dialectal Arabic reduces the potential effectiveness of utilizing the resources available for a specific variety to investigate another one, be it another dialect or MSA. (Zampieri and Nakov, 2021).

In the literature, Arabic dialects are typically divided based on a geographical dimension with different levels of granularity: region, country, and city level. The regional level represents different regions in the Arab world consisting of a set of countries. Note that grouping the dialects of those different countries on a regional level does not imply that the group of dialects is entirely homogeneous linguistically (Habash, 2010).

Previous work focuses on those two levels of granularity region (Zaidan and Callison-Burch, 2014; Zampieri et al., 2018). Recently, there has been more work on the country-level dialect that focuses on a specific country and all the sub-dialects spoken in that country. Current work on the country-level dialect focuses on a specific task (AlYami and Al-Zaidy, 2022; Yang et al., 2020; Farha and Magdy, 2019; Habash et al., 2019) or studies MSA and few dialects (AlYami and AlZaidy, 2020; Alshargi et al., 2019; Khalifa et al., 2016).

Other work investigates the city-level dialect of specific cities in a country. Most of the work on this level utilizes social media posts coming from a specific city as the original city dialect (Bouamor et al., 2019; Abdul-Mageed et al., 2019, 2018). However, the social fabric in major cities consists of residents speaking different dialects, which causes a problem at this granularity level. Hence, relying on social media locations for collecting data for users from a specific location does not ensure that the user speaks the predominantly spoken dialect of that location. This work focuses on the location-level and individual-level dialects by allowing users to specify their individual dialect. The dataset is representative of 29 nuanced dialects from Saudi Arabia, city-level dialects of 9 cities from 3 Arab countries and 18 country-level dialects in the Arab world.

3 Our Approach

In this section we describe our proposed approach for nuanced dialect data collection.

3.1 Overview

The data collection approach is comprised of two general tasks. The goal of the first task is to collect quality short-texts representing nuanced dialects that are both versatile in nature and large in quantity. Since MSA texts are available in abundance, due to it's common use in digitized content, it has been widely studied leading to NLP systems obtaining high accuracies for the MSA variety. Given that MSA is the variety from which all Arabic dialects are derived from, similarities and differences between Arabic dialects and MSA has always been of interest to both linguists and NLP researchers. For that reason our approach is designed to collect data that enables further research and modeling of these similarities and differences. Additionally, our system includes an approach to incorporate the semantic-level cultural nuances of the collected dialects.

The goal of the second task is to promote the collection of additional nuanced-dialect data collection systems. Social media content has been the predominant source of dialectal data. Although social media content has proven effective for improving NLP performance on dialects to a great extent, nuances in dialects still remain a major challenge to most dialectal Arabic NLP systems unlike MSA. The goal of this task is to contribute to the quantity aspect of manually entered nuanced-dialect data by developing more systems similar in goal to the one

we develop for our first task.

3.2 Nuanced Dialect Short-Texts Collection

This approach aims to collect as large a volume as possible of nuanced dialectal Arabic data. The approach is designed as a data marathon competition, where the competitors goal is to respond to as many prompts as possible, in their own dialect within a fixed time-frame with as few errors as possible. Winners are the the top ranked teams with highest volumes of entries.

Two methods were followed to collect the two distinct corpora. Their description is as follows:

Aligned Parallel Dataset This collection method is designed to allow participants to translate sentences from MSA to their local dialects. The MSA sentences are sampled from subset of $11,670$ sentences from an existing well-known MSA dataset, namely the MADAR dataset. The participants translate it to the dialect they registered as their own when joining the competition.

Q&A Dataset This method allows participants to answer open-ended questions. The set of questions are updated regularly for the competition participants, where they answer them in an open-ended fashion. Questions are either constructed by the authors or collected from QA websites such as Quora. The total number of questions used is 796 and will be released with the dataset.

The competition was implemented using a web application designed to receive submissions for the competition and was built to be highly usable even by non-technology-savvy people. The designed collection tool consists of two stages: the registration, where the participants register themselves in the competition with their information, a dialect they speak in with native fluency, and either creating or joining a team. Individual participation was allowed (with a team of one), and up to five members could be included in the team. The second stage is the submission page, where the participants can choose between the tasks of either translating or question answering. Gamification elements were integrated in the design to encourage the participants, such as different game levels with different progress bar colors for the team and each team member. Participants were also encouraged to report any inappropriate sentence or questions, and the reported sentence or questions would be reviewed within 24 hours and removed if necessary. Participants have the option to skip any question they did not want to answer or translate.

3.3 Innovating Data Collection Methods

This approach aims to collect systems that are used to collect nuance dialect data. It follows a typical hackathon format. The competition elicits creative ideas to collect nuanced dialectal Arabic data. The competition was launched to the public on 26^{th} of February 2022. It consisted of two stages: in the first stage, which lasted for 12 days, participants were asked to submit their team's information and a brief description of their proposed idea. A total of 57 submissions were received from 173 participants. It ranged from ideas to extracting dialectal data of social media content, games, and crowdsourcing techniques. In total, 24 teams were nominated to move to the second stage based on the relevance criteria reviewed by 2 evaluators. Nominees from this stage were provided the opportunity to attend two workshops; the first was titled "Automatic Data Collection and Annotation" and presented different existing methods for collecting and annotating data along with special challenges that face collecting dialectal Arabic. The second was titled "Designing Inclusive Applications and Platforms" and focused on the usability of web and mobile apps in addition to tips and tools for presenting Hackathon ideas. They also joined the competition discord account, where they received mentoring from experts. Out of 24 nominees, 19 did proceed to make the final submission, which consists of the prototype of the solution, a short video explaining the idea of the proposed solution, and a time sheet to realistically complete such a project.

4 Data Collection Results

In this section we describe the resulting datasets and methods.

4.1 Nuanced Dialect Short Texts

The data collected from this approach covered 21 dialects from Arabic-speaking countries. The number of dialects, denoted n, is as follows. For Saudi Arabia, 29 Saudi dialects were collected, i.e. $n = 29$. Since no official definition for Saudi dialects exists, in this study we mainly adopt the taxonomy used in an online linguistic effort, معجم اللهجات المحكية في المملكة العربية السعودية (Aldarsoni, 2013). For Yemen, Jordan, and Syria, we collect city-level dialects for major cities where,

Table 1: Samples of MADAR sentences provided for translation and open-ended questions created by the Data Marathon team

sentences for translation	open-ended questions
غالباً ما أخذ مقاس اثنتي وعشرين في اليابان	إذا اضطررت للانتقال لبلد آخر، فما الأشياء التي ستفتقدها ببلدك الآن؟ ولماذا؟
I often wear a size 22 in Japan	*If you had to move to another country, what things would you miss about your country now? And why?*
الوقت انتهى	بماذا تتميز مدينتك؟
The time is up	*What distinguishes your city?*
هل هذه متجه إلى فندق جراند	كيف تبدو أجواء رمضان في منطقتك؟
Is this heading to the Grand Hotel?	*What does Ramadan look like in your area?*
نريد مائدة بجانب النافذة	ما أثر تقديم الهدية للآخرين
We want a table by the window	*What is the effect of giving a gift to others?*
هل لديكم أية جولات سياحية استطلاعية بالحافلة حول المدينة	ما الذي يحفزك للاستيقاظ كل يوم؟
Do you have any guided bus tours around the city?	*What motivates you to wake up every day?*

$n = 6, 4, 7$ for the countries, respectively. As for the remaining countries, country-level dialects are defined with $n = 18$ for 18 countries. The lists of all dialects in their Arabic names are provided in Table **??**, Table 10, and Table 9 in the appendix.

The data collection duration was from 26^{nd} of February 2022 until 21^{st} of March 2022. It was highly publicized on social media by Arabic NLP experts, several local university accounts, and several local NLP enthusiasts. Data was collected from 560 participants from 14 countries grouped under 422 teams. Over these teams, 354 teams were teams of individuals, 34 teams were a team of 2, 11 teams were a team of 3, 11 were a team of 4, and 12 were a team of 5.

A total number submission of 55,484 was received. The number of submissions for the Saudi dialects is in Table 2. We received more than 5 submissions for 19 out of the 29 targeted Saudi dialects. We received more than 5 submission for two of the Yemeni regional dialects, three of the Jordanian regional dialects, three of the Syrian regional dialects, and 10 of the remaining 18 Arabic countries, in Table 3.

The winners of the Data Marathon were announced on 24^{th} March 2022 and belonged to *Yemen-Ta'izz* dialect (code 600) with 7413 submissions, *Saudi Arabia-Ghamid and Zahran* (code 21) with 6328 submissions, and *Saudi Arabia-Al Qassim* (code 2) with 6134 submissions. The three places keep their order whether we consider all submissions or only submissions with lengths more than 10 characters.

4.2 The *KIND* Dataset

In order to prepare the data for public use, all submissions are anonymized by releasing only 3 fields per submission.

Table 2: Saudi dialect submissions by sentence length.

Dialect Code	1-10	11-25	26-50	51-100	100 <	All Submissions
0	1525	4447	3653	1074	220	10919
1	57	259	258	75	16	665
2	478	3178	2536	394	31	6617
3	31	117	113	11	3	275
6	51	66	31	34	31	213
7	2	16	14	2	3	37
8	16	25	20	2	2	65
11	26	100	80	44	18	268
12	474	1881	1278	349	86	4068
13	0	1	0	0	0	1
16	0	0	1	0	0	1
18	50	141	118	73	21	403
19	291	798	501	131	11	1732
20	220	1320	1082	226	69	2917
21	767	4341	2581	241	11	7941
22	12	23	12	5	1	53
23	27	151	161	53	27	419
24	53	202	179	61	15	510
25	0	0	2	3	1	6
26	38	166	149	46	24	423
28	209	1115	805	203	46	2378
Total per length range	4327	18347	13574	3027	636	39911

Table 3: Arabic dialects submissions by sentence length. (without Saudi Dialects)

Dialect Code	1-10	11-25	26-50	51-100	100 <	All Submissions
100	453	1744	1425	317	68	4007
200	81	280	116	10	0	487
300	9	41	64	27	6	147
400	11	119	136	39	4	309
500	2	13	21	14	0	50
600	674	3694	2596	707	463	8134
601	19	77	79	13	3	191
701	28	29	20	10	6	93
703	1	7	8	16	21	53
705	53	553	555	122	10	1293
800	8	25	29	2	0	64
801	2	18	16	7	2	45
803	0	5	1	0	0	6
1000	37	62	57	16	3	175
1200	31	150	96	24	3	304
1300	14	75	46	23	9	167
1500	4	8	4	2	2	20
2000	1	15	12	0	0	28
Total per length range	1428	6915	5281	1349	600	15573

Table 4: Sample of submissions received by open-ended questions.

Dialect Code	ماهو روتين يوم عيد الفطر عند عائلتكم ؟ What is your family's Eid al-Fitr routine?
0	جمعه الاهل والاطفال وتوزيع الهدايا Family and children gathering and gift distribution
0	نتجمع ونحتفل وناكل شوكليت ونشوف كل اللي نحبهم We gather, celebrate, eat chocolate, and see everyone we love
0	نجتمع سوى ونتقهوى ونسلم على بعض ونلبس ثيابنا الجديدة ثم كلن يروح غرفته ويرقد We gather together, have coffee, say hello to each other, put on our new clothes, and then we all go to our rooms and sleep.
3	اسير على الربع وما اخلي حوي الا ادخله و اعيد على اهله Visit friends and not leave a neighborhood I did not enter and wish it's residents a happy Eid.
12	احنا في العيد نزور الكباريه الصباح مرا بدري ونتجمع هناك و نفطر ونشرب قهوه ، عاد بالليل تصير الفعاليات اكتر شي We, on Eid, visit the elders early morning and gather there and have breakfast and drink coffee, then most festivities happen at night.
12	تتجمع كل العيلة بصباح العيد ونقعد نشرب شاهي ونهرج لين الظهرية بعدها نتغدا وكل واحد تمشي بيتو بعدها We gather, the entire family, Eid morning and sit to drink tea and chat until noon then we have lunch and everyone goes home after that
19	نرقد We sleep.
19	نصلي صلاة العيد ونطلع نعايد لين الظهر وبعدها نغفي ونرجع نكشخ المغرب نكمل نعايد We pray the Eid prayer and go out for Eid celebrations until noon, then we go sleep and at Maghrib time again we dress up and continue our Eid celebrations.
21	نتجمع مع صدقاننا من صلاة الفجر ونفطر We gather with our friends at Fajr prayer and have breakfast.
24	نسلم على بعض صلاة الصبح ونفطر ونرجع نرقد We greet each other at Fajr prayer, have breakfast, then go back to sleep.
28	ان العايله تجتمع The family gathers.
100	لازم نعمل محشي و كحك و بتيفور We have to make *mah'shi, ka'ak* and *betefour*
100	الفرحه مع الاهل و ناكل مع بعد و نلعب مع بعد و نصلي مع بعد Joy with the family, we eat together and play together and pray together.
200	نوضوا الصباح نروحوا نصليوا ونرجعوا نفطروا مع بعض مبعد نروحوا نعيدوا على باقي العايلة We do ablution in the morning, then we go to prayer, we come back and have breakfast together, then we go to greet the rest of the family
600	نجهز الكعك و المعماله للجهال ونسير المجلس ونتصل ونبارك للعيد We prepare *ka'ak* and *ja'ala* for the children then we go to the *majlis* and celebrate Eid
600	نفرح طماش ما نرقدشمن الفرح We are elated so happy we cannot sleep
703	بعد صلاة العيد يرجعوا عالبيت ومنعمل فطور لكل العيلة وبعدها منفوت بغيبوبة نوم والعصر او الظهر منقوم وبيبدأ عيدنا ومنلبس اواعي العيد After the Eid prayer, they return home and we prepare breakfast for the whole family, and then we fall into a sleepy coma, and the afternoon or noon prayer begins, and our Eid begins, and we wear Eid clothes.
1200	نقعدو واعيين لعند الصبح وبعدها قبل الصلاة نحشو نتبحو ونطلعو نلبسو العربي ونمشو للصلاة وبعد ما نروحو نلقو العصيدة واتية ناكلوها ونبدو في مشاوير لمعايدة We stay up until morning then before the prayer, we go in and shower then go out and dress up and go to the prayer then when we are back we find *asida* ready, we eat it then start Eid visiting errands.

Table 5: Percentage of submissions generated from open-ended questions per submission length for dialects with more than 5 submissions.

Dialect Code	11-25	26-50	51-100	Longer than 100
0	32%	34%	60%	**87%**
1	34%	54%	**81%**	**100%**
2	4%	9%	25%	**90%**
3	29%	42%	64%	**100%**
6	71%	**97%**	**100%**	**100%**
7	13%	14%	**100%**	**100%**
8	76%	40%	0%	**100%**
11	56%	73%	**95%**	**100%**
12	16%	24%	63%	**93%**
18	35%	59%	**92%**	**100%**
19	31%	37%	69%	**100%**
20	10%	21%	51%	**94%**
21	6%	6%	17%	45%
22	43%	50%	**80%**	**100%**
23	43%	63%	**85%**	**100%**
24	47%	49%	70%	**100%**
25	-	**100%**	**100%**	**100%**
26	12%	8%	57%	**100%**
28	6%	13%	55%	**87%**

1. *dialect code* which is the label that indicates the specific dialect the text belongs to.

2. *sentenceOriginID* which is the identifier used to reference either the MSA sentence to it's source in MADAR dataset, this ranges (1000000-2000000), or the reference to link the question to the constructed question dataset, this ranges (2000000-3000000).

3. *textString* contains the submitted sentence.

Additional processing of the submissions includes trimming, and removal of duplicate translations of a sentence or answer submitted by the same participant. This reduced the total number of entries from $55,484$ to $54,883$. It is worth noting the final dataset does not include duplicates from the same dialect and source; however, it does include duplicates of different sources of the same dialect, e.g., two different questions have the same answers.

4.3 Discussion

The resulting datasets demonstrates the efficacy of our collection approach that relies on gamification combined with awareness-raising on the importance of inclusiveness and availability of open-

Table 6: Sample of submissions received by translation of *May I ask about your name?* from Arabic MSA to other dialects.

Dialect code	هل لي أن أسأل عن اسمك ؟
0	اقدر اعرف اسمك May I know your name?
1	هالحين وش اسمك Now what's your name?
20	بسألش وش اسمش لو سمحت I want to ask you, what is your name please?
21	اقدر اسال عن اسمك Can I ask about your name?
100	ممكن اسألك اسمك ايه؟ Is it possible to ask what your name is?
600	؟ ككا قولي مو اسمك Tell me, what is your name?

source resources. The collected data size in Table 7, shows the number of unique entries when considering the uniqueness on the dialect level. Both approaches resulted in a large number of texts for the duration. The sentence translation approach generated a larger number of entries, as was expected since minimal effort is required to simply restate existing content as opposed to question answering that requires the participant to generate new content. Also, the sentnece translation system was available to the users 3 more days than the question answering system.

Samples of responses to open ended questions are in Table 4, while samples of responses of translation are in Table 6. It is notable that the sentence translation approach captures dialectal markers which is a main challenge for nuanced dialects. Additionally the texts are collected with their MSA translation and other dialects, that are essential for studies addressing translation-based solutions to modelling nuanced dialects. The open-ended question answering approach is capable of capturing both the syntactic aspects of the dialect as well as the semantics associated with the culture of the speakers of that specific dialect.

In terms of the length of submissions, it is noted that the percentage of submissions on the longer length side are the ones generated by the question answering approach. In Table 5, focusing on sentences of Saudi Dialects, in 17 out of 18 investigated dialects, submission longer than 100 characters came mainly (more than 85%) from open-ended questions. In contrast, submissions shorter than 50 characters came mainly from translated sentences. Same pattern applies to non-Saudi dialectal Arabic sentences, where translation sentences were responsible for 75% of the submissions shorter than 50 characters, whereas open-ended questions were responsible for 77% of the submissions longer than 50 characters. This observed too in Table 4 and Table 6, where samples of the open ended question mainly consists of longer sentences whereas for translation we are bound by the length of the original sentence. To generate longer sentences for translation, the approach requires using a dataset with longer texts to prompt the participants in the translation task.

5 Hackathon Results

In this section, we shed light on the top projects that received the highest scores from the hackathon judging committee per the evaluation criteria, which seek to balance technical knowledge with originality, creativity, and relevance to the hackathon's objectives.

5.1 Lesan

Lesan is a volunteer platform designed to enrich the Arabic dialect content, focusing mainly on voice-recorded sentences in various dialects. The volunteers start by choosing a dialect to voice record Arabic text written in the selected dialect and complete their daily rounds, where each user has a daily target of 10 rounds per day. The users of Lesan have the choice to record an existing text in the platform or type and record their own new text. Gamification elements such as trophies and leaderboards are used to motivate users by creating a competitive atmosphere. Moreover, Lesan provides an "Open Library" that contains high-quality dialectal Arabic datasets that are available, reliable, and ready to use.

5.2 Teach us your Dialects

Participants proposed a game in which players guess the meaning of a word given in one of the Arabic dialects with the help of an image indicating the meaning. The players' answers can be written or recorded in their voices. In this project, the database is populated by the players themselves, where any player can add a word in a specific dialect, and it will be added to the database if it is approved by at least ten other players who speak the same dialect or live in the same place. Each

Table 7: Entries uniqueness.

Number of submission	Unique per (dialect, source, participent)	Unique per (dialect, source)	Unique per (dialect)
Translation	40481	40119	39957
Open ended Questions	15003	14759	14338

new word has to be recorded in voice to make pronunciation easier. To motivate players, they can see a leaderboard of the players' points and ranks by their countries. Moreover, The ranking of each country is displayed as well, and it depends on the number of words shared by players from the country.

5.3 Faseeh (Fluent)

A video game that asses a person's level of knowledge of Arabic dialects. As a first step, data is collected from Twitter automatically using a scrapper. Data is then cleaned and stored in a database for the game, where five tweets will be displayed for each player from the database. In Faseeh game, a tweet will pop up on the screen where the player must answer the following questions in a row: (1) Is there a text in the tweet that indicates a specific dialect? (Yes/No), (2) If yes, what is the dialect classification?(Egyptian - Gulf etc.), (3) If yes, why was this dialect chosen?, (4) What is the text/word indicating the selected dialect?.

Once the player finishes a tweet, another tweet will pop up until the round (five tweets) is completed. Eventually, the Arabic dialect test result in points will be displayed to the user based on the number of correct answers. Verification of answers is done in two approaches: automatically by knowing the geographical location of the source of the tweet and manually by operators who are experts in each of the existing dialects and can evaluate the answers.

5.4 Nutq (Pronunciation)

The project's main interface is divided into several sections for Saudi dialects, including Northern, Southern, Hijazi, Najdi. etc. The application consists of 3 stages to collect data for each dialect in a funny way to grab the user's attention. In the first stage, the application shows an image to users where they can choose the appropriate word from their own perspective. As the second stage, If none of the options matches the word in their dialect, they can add their own synonym for the word. In the third stage, users can go the extra mile and add more words along with their meanings in the game dictionary and get simple financial rewards.

5.5 Evaluation and results

Each of the 19 submissions received three evaluations on the premises of:

1. Creativity level of the idea.
2. Technical quality and suitability of techniques and methods used.
3. Potential to Grow.
4. The collected data using the provided technology in terms of quantity and quality.

The results were aggregated, and the announcement of the winners was on the 24^{th} of March 2022. The first three places consist of teams Lesan, Teach Us Your Dialect, and Fsaeeh, respectively.

6 Challenges and Recommendations

In this section we describe challenges to the collection approach and provide recommendations for organizing future hackathons.

6.1 Target Dataset

In order for the collection approach to effectively achieve a high yield of quality data, it is recommended to have a clear specification of the dataset to be collected, specifically a predetermined purpose or use for the data. Although this ensures a consistent collection process, however, during collection it is equally beneficial to adapt to patterns in user behavior to maximize the outcome of user participation. For example, the first collection task was initiated with the purpose of curating a parallel corpora for different dialects. It was noticed during user submission that many of the MSA sentences that were presented to the users for translation were irrelevant to the culture and could not capture cultural and context-rich dialectal data. The sentences were parts of conversations in the hotel, airports, and restaurants and were a direct translation from non-Arabic sentences. Therefor the second approach was introduced proposing the

use of open-ended QA for collecting dialectal data to enable culture and context-rich sentences while relaxing the alignment requirement to have totally different responses for the same question.

6.2 Target Participants

The target audience should be clearly defined as it is a focal point in the competition design process. Since we are interested in collecting data from under-represented groups, it was imperative to design a system that is easy to use by non-technical groups who typically have low online presence. To reach our target audience we used social media outlets that have high visibility in the region, using simple video advertisements to convey the purpose of the data collection and the potential of the benefits to society as a whole for a non-technical user. Incentives are used to maximize user participation in social collaborative efforts, such as monetary awards in our competition. In cases of limited funding, we recommend to emphasize the social media campaign and raising awareness efforts, as many educational sessions targeting college students and the general public contributed greatly to the high amount of participation.

6.3 Technical Resources

Storage and database size limitations dictate the limits of the collection process. The participant solicitation must be guided by the volumes of data received during collection to ensure system stability. Additionally, available personnel to provide technical support and monitor entries to perform corrections or incorporate user feedback, is a challenge. In our case, the authors along with student volunteers from the university were responsible for these tasks.

7 Resources

The dataset is released for the research community at: https://huggingface.co/KIND-Dataset. The repository holds both the Data Marathon submissions and the open-ended question dataset.

8 Conclusion

In this paper we describe the design and process of collecting a multi-dialect Arabic dataset as well as the resulting systems and data. Similar in concept to ACL shared tasks, the KIND competition aims to encourage innovative contributions towards high-quality data collection. The competition resulted in a corpus of over 50k high-quality texts labeled with fine-grained Arabic dialects. As well as over 20 approaches for Arabic dialect crowdsourcing techniques. The resulting data is made public for the research community. As future work, the authors aim to propose new competitions for domain specific as well as NLP-task specific data collection for Arabic dialects.

Limitations

There are several limitations in the published dataset, that open doors for further investigation: (1) The first is the presence of white dialect submissions, as most of the participants were 35 of age or younger and live in big cities not in their hometown, we can find that some of the dialect has softened from how the original dialect sounds. (2) As dialect classification is a multi-label problem, submissions could be mapped to more dialects than what is reported.(3) Further cleaning is required, although there were minimal spam submissions from our observations, there still could be submissions that do not answer to the question or translated the intended sentence. (4) Not all intended city-level (or tribe-level) dialects in Saudi Arabia were covered as intended. The dataset does lack submissions from dialect belonging mainly to the Northern Regions of Saudi Arabia. Also, not all Country level dialects were covered, especially for dialects of North African countries.

Ethics Statement

All participation in the competition was voluntary and participants waived their copyrights to the submitted data before participation. All information related to the participants identity was removed. The dataset is not comprehensive of all Arabic dialects and should not be treated in such manner.

Acknowledgements

The authors would like to thank Ruba Alzahrani, Mais Alheraki, and Nadeen AlAmoudi for their assistance with various competition organization tasks.

References

Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. 2018. You tweet what you speak: A

city-level dataset of Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop (WANLP 2020)*, Barcelona, Spain.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Arun Rajendran, and Lyle Ungar. 2019. Dianet: Bert and hierarchical attention multi-task learning of fine-grained dialect.

Sulaiman Aldarsoni. 2013. معجم اللهجات المحكية في المملكة العربية السعودية.

Faisal Alshargi, Shahd Dibas, Sakhar Alkhereyf, Reem Faraj, Basmah Abdulkareem, Sane Yagi, Ouafaa Kacha, Nizar Habash, and Owen Rambow. 2019. Morphologically annotated corpora for seven Arabic dialects: Taizi, sanaani, najdi, jordanian, syrian, iraqi and Moroccan. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 137–147, Florence, Italy. Association for Computational Linguistics.

Maha J. Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey.

R. AlYami and R. AlZaidy. 2020. Arabic dialect identification in social media. In *2020 3rd International Conference on Computer Applications Information Security (ICCAIS)*, pages 1–2.

Reem AlYami and Rabah Al-Zaidy. 2022. Weakly and semi-supervised learning for Arabic text classification using monodialectal language models. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 260–272, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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)*.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Mona Diab, Mohamed Al-Badrashiny, Maryam Aminian, Mohammed Attia, Heba Elfardy, Nizar Habash, Abdelati Hawwari, Wael Salloum, Pradeep Dasigi, and Ramy Eskander. 2014. Tharwa: A large scale dialectal Arabic - Standard Arabic - English lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3782–3789, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

N.Y. Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis digital library of engineering and computer science. Morgan & Claypool Publishers.

Salima Harrat, Karima Meftouh, Mourad Abbas, Salma Jamoussi, Motaz Saad, and Kamel Smaili. 2015. Cross-dialectal arabic processing. In *Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I 16*, pages 620–632. Springer.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of Gulf Arabic. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4282–4289, Portorož, Slovenia. European Language Resources Association (ELRA).

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Sebastian Ruder. 2020. Why You Should Do NLP Beyond English. `http://ruder.io/nlp-beyond-english`.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, pages 10–21.

Claudia Soria, Valeria Quochi, and Irene Russo. 2018. The DLDP survey on digital use and usability of EU regional and minority languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing,*

Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006, pages 303–324.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Wikipedia. 2023. List of countries and territories where arabic is an official language.

Qiang Yang, Hind Alamro, Somayah Albaradei, Adil Salhi, Xiaoting Lv, Changsheng Ma, Manal Alshehri, Inji Jaber, Faroug Tifratene, Wei Wang, Takashi Gojobori, Carlos M. Duarte, Xin Gao, and Xiangliang Zhang. 2020. Senwave: Monitoring the global sentiments under the covid-19 pandemic.

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

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *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.

Marcos Zampieri and Preslav Nakov. 2021. *Similar Languages, Varieties, and Dialects: A Computational Perspective*. Studies in Natural Language Processing. Cambridge University Press.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

A Appendix

A.1 Dialect Codes

Table 8: City/Tribe level Saudi Dialects.

Dialect Code	City/Tribe Dialect
0	حاضرة نجد (الرياض، قرى بني حنيفة، الوشم) Najd - urban
1	بادية نجد (عتيبة، مطير، قحطان، سبيع..) Najd - suburbs
2	أهل القصيم Al Qassim
3	أهل وادي الدواسر Wadi Al-Dawasir
4	حوطة بني تميم Hotat Bani Tamim
5	أهل الأفلاج Al-Aflaj
6	أهل الدلم والخرج Ad-Dilam and Al-Kharj
7	شمر Shammar
8	عنزة Anaza
9	الشرارات Al-Shararat
10	الحويطات Al-Howaitat
11	الحجاز Hejaz/Hijaz
12	عوائل الحجاز Hijaz families
13	أهل العلا Al-Ula
14	الرشايدة Al-Rashaida
15	خيبر Khaybar/Khaibar
16	جهينة Juhaina
17	العجمان Al-Ajman
18	الهواجر Al-Hawajir
19	الأحساء Al-Ahsa
20	القطيف Al-Qatif
21	غامد وزهران Ghamid and Zahran
22	بني شهر وباللحمر ،باللسمر Bani-Shehr, Ballahmar, Ballasmar

Dialect Code	City/Tribe Dialect
23	شهران العريضة Shahran Alaridha
24	تهامة (رجال ألمع، الأزد، قحطان) Tihama
25	فيفا Faifa/Fifa/Fayfa
26	جازان Jazan
27	المهرة Al-Mahra
28	بني يام (نجران) Bani-Yam (Najran)

Table 9: Country level Arabic Dialects.

Dialect Code	Country Dialect
100	Egypt مصر
200	Algeria الجزائر
300	Sudan السودان
400	Iraq العراق
500	Morocco المغرب
900	Somalia الصومال
1000	Tunisia تونس
1100	Emirates(UAE) الإمارات
1200	Libya ليبيا
1300	Palestine فلسطين
1400	Oman عمان
1500	Kuwait الكويت
1600	Mauritania موريتانيا
1700	Qatar قطر
1800	Djibouti جيبوتي
1900	Comoros جزر القمر
2000	Lebanon لبنان
2100	Bahrain البحرين

Table 10: City level (non-Saudi) Arabic Dialects.

Dialect Code	City Dialect
600	اليمن ـ اللهجة التعزية Yemen-Ta'izz
601	اليمن ـ اللهجة الصنعانية Yemen-Sana'a
602	اليمن ـ اللهجة الحضرمية Yemen-Hadhramut
603	اليمن ـ اللهجة اليافعية Yemen-Yafea
604	اليمن ـ اللهجة العدينية Yemen-Adeeni
605	اليمن ـ اللهجة العدنية Yemen-Aden
700	سوريا ـ لهجة إدلب Syria-Idlib
701	سوريا ـ لهجة حلب Syria-Aleppo
702	سوريا ـ لهجة حمص Syria-Homs
703	سوريا ـ لهجة دمشق(شامي) Syria-Damascus (Shami)
704	سوريا ـ لهجة درعا Syria-Daraa
705	سوريا ـ لهجة حماه Syria-Hama
706	سوريا ـ لهجة اللاذقيةوالساحل Syria-Latakia and coast
801	الأردن ـ لهجة الفلاحين Jordan-Fallahin
802	الأردن ـ لهجة الشمال الأردنية Jordan-Northern
803	الأردن ـ لهجة الجنوب الأردنية Jordan-Southern
804	الأردن ـ اللهجة البدوية Jordan-Bedouin