DiaSet: An Annotated Dataset of Arabic Conversations

Abraham Israeli,[1,2] Aviv Naaman,[1] Rawaa Makhoul,[1] Guy Maduel,[1] Amir Ejmail,[1] Julian Jubran,[1] Dana karain,[1] Dina Lisnyansky,[1] Shai Fine,[1], Kfir Bar[1]

[1]The Data Science Institute, Reichman University, Israel
[2]Department of Software and Information System Engineering
Ben-Gurion University of the Negev, Israel
{aviv.naaman,makhoul.rawaa,guy.maduel}@post.runi.ac.il,
{julian.jubran,dana.qaraeen,dina.lis}@post.runi.ac.il,
{shai.fine,kfir.bar}@runi.ac.il
ameerejmail44@gmail.com
isabrah@post.bgu.ac.il

Abstract

We introduce DiaSet, a novel dataset of dialectical Arabic speech, manually transcribed and annotated for two specific downstream tasks: sentiment analysis and named entity recognition. The dataset encapsulates the Palestine dialect, predominantly spoken in Palestine, Israel, and Jordan. Our dataset incorporates authentic conversations between YouTube influencers and their respective guests. Furthermore, we have enriched the dataset with simulated conversations initiated by inviting participants from various locales within the said regions. The participants were encouraged to engage in dialogues with our interviewer. Overall, DiaSet consists of 644.8K tokens and 23.2K annotated instances. Uniform writing standards were upheld during the transcription process. Additionally, we established baseline models by leveraging some of the pre-existing Arabic BERT language models, showcasing the potential applications and efficiencies of our dataset. We make DiaSet publicly available for further research.

Keywords: Arabic NLP, Spoken Arabic, Dialectical Arabic, NLP Resource

1. Introduction

Arabic is the fifth most spoken language worldwide, with over 374 million native speakers. Arabic is the official language in over 22 countries, ranking third behind French and English, which are official in 29 and 58 countries, respectively (WorldData, 2023).

While Arabic natural language processing (NLP) was relatively under-explored just half a decade ago, there has been a notable surge in research in this area in recent years (Antoun et al., 2020; Guellil et al., 2021a; Darwish et al., 2021; Habash, 2022). Nowadays, Arabic is considered a well-studied language across various computational domains. Some domains for example are morphological analysis (Obeid et al., 2020), named entity recognition (Ali et al., 2020; Qu et al., 2023), sentiment analysis (Al-Ayyoub et al., 2019; Farha and Magdy, 2019; Israeli et al., 2021), and large language models (LLMs) (Antoun et al., 2020; Abdul-Mageed et al., 2020; Lan et al., 2020; Inoue et al., 2021a). Unsurprisingly, as the Arabic NLP research progresses, we observe a rise in the Arabic-based corpora introduced to the community. Naturally, most of these corpora are generated from textual data, such as news articles, and posts on social media.

Although spoken data (i.e., data from audio sources) are essential for many NLP tasks, only 16.8% of the available Arabic NLP corpora come from audio sources (Alyafeai et al., 2021). The primary reason for this low number is technological maturity. While the performance of voice-to-text algorithms has significantly improved over the years, converting voice to textual data still necessitates manual post-processing, particularly in languages other than English. Specifically in Arabic, there is a difference between the spoken dialect and the written Modern Standard Arabic (MSA) language. This phenomenon is commonly known as diglossia. People's dialects vary based on locality and ethnicity. As a result, transcribing Arabic is often considered more challenging than other languages.[1] In this work, we focus on spoken Arabic as we introduce a large corpus of dialectical Arabic speech, transcribed from two audio sources.

Every human language varies in style, accent, terminology, and slang, depending on the time and place of its usage. However, the range of spoken dialects in Arabic is unique compared to other languages. Originating from the Arabian Peninsula, dozens of Arabic dialects exist, each often treated as its own distinct sub-language. Naturally, most studies on Arabic, which employ computational tools, focus on MSA since it has the highest number of available resources (Darwish et al., 2021). The Arabic dialectical resources that do exist typically focus on the most dominant Arabic dialects: Egyptian (21.56%) and Moroccan (7.43%). The

[1] Historically, spoken Arabic has not been used for writing. MSA is the default way to be expressed in Arabic.

LREC-COLING 2024, pages 4865–4876
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

South Levantine dialects, which are mostly used in Jordan, Israel, and Palestine, have only a limited number of such resources (Alyafeai et al., 2021).

In this paper, we concentrate on the Palestinian dialect (hereafter referred to as PAL). This dialect, a product of the rich cultural mosaic of the Levant, exhibits several linguistic deviations from MSA.

In Table 1 we present five examples of sentences from the new dataset we introduce in this work. These examples highlight some of the unique characteristics, differences, and challenges of PAL. In the first example, we illustrate how words and their order in the sentence differ from MSA. The first sentence would have been written as الاشياء التي أنا اخترتها بنفسي فقط in MSA. In the third example, we highlight a typical PAL term. While the term الآن is commonly used in MSA to refer to the current time (i.e., "now" in English), in this example, the word اسّا ("issa") is used, which is different than the equivalents in other Arabic dialects: هلّأ ("hallaʾ") and دلوقت ("dilwaʾt") in Lebanese and Egyptian respectively. Similarly, in the last example, the term هاى ("hāy") is used while referring to the word "this", unique to PAL.[2]

PAL is influenced by history, culture, and national identity. An example of this can be found in the second row of Table 1. In this example, the phrase عرب الداخل ("Arabs of the interior") is used while referring to the Arabs who stayed within the Israeli territory after the 1948 Arab–Israeli war.

PAL stands out for its diversity and variation in both meaning and pronunciation from region to region (Jarrar et al., 2017). Differences exist in the dialects of the north/center/south and even between neighboring cities. PAL is special and unique as it consists of several sub-dialects that vary in many aspects. Bedouin, Druze, rural, and urban are among the sub-dialects we observe. The latter varies phonologically among the major cities such as Jerusalem, Nazareth, Haifa, and Nablus. This is evident in the vocabulary, semantics, and pronunciation of words. These disparities encompass phonological, morphological, and lexical aspects, which warrant specific attention in NLP endeavors. We make sure to include the many sub-dialects of PAL while working on this research.

In this paper, we present DiaSet, a novel dialectical Arabic dataset, comprised of manually transcribed speech. A portion of the transcriptions was manually annotated for classic NLP tasks: sentiment analysis (SA) and named entities recognition (NER). We make DiaSet publicly available for the community.[3] To the best of our knowledge, DiaSet is the first, most thorough, and largest spoken Arabic dataset in PAL. The significance lies in the unique data collection method, as it specifically focuses on audio data derived from human conversations. Using the annotated part of DiaSet, we present baseline results for SA and NER, the first models to be published specifically for PAL. To summarize, our main contribution is twofold:

i Corpus of Arabic speech transcriptions. We introduce a meticulously transcribed dataset in PAL. The dataset consists of two main sources: (a) YouTube podcasts and (b) Interviews, which we recorded for this research. Both sources feature conversations between two speakers: the interviewer and the interviewee. Both participants are native PAL speakers. In total, we manually transcribed 83 hours from YouTube podcasts and 15 hours from dedicated interviews.

ii Annotated dataset. We manually annotate 23.2K text units, taken from the transcribed interviews. Each text unit is annotated for two downstream tasks: SA, and NER. We present two NLP models (i.e., sentiment analysis and named entity recognition) based on the annotated dataset and compare the results with some existing baselines.

The remainder of the paper is organized as follows: in Section 2 we provide a brief review of some related works. Section 3 describes DiaSet and the data creation process. In Sections 4 and 5 we describe the computational methods we use for modeling the annotated data from DiaSet, followed by the results we obtained. In Section 6 we outline the central limitations of this work and in Section 7 we conclude with a discussion and suggest some future research directions. Finally, in Section 8 ethical considerations related to our work are discussed.

2. Related Work

Arabic NLP has experienced a surge of interest and research in recent years (Shoufan and Alameri, 2015; Habash, 2022). Numerous studies have focused on enhancing NLP methodologies for the Arabic language, which includes a variety of elements such as sentiment analysis (SA), machine translation (MT), named entity recognition (NER), and part-of-speech (POS) tagging (Antoun et al., 2020; Abdelali et al., 2016; Soliman et al., 2017). This extensive attention underscores the escalating interest and prospective influence of Arabic NLP across a multitude of domains. A pivotal role given that Arabic is the fourth most popular language on the Web (Guellil et al., 2021b).

The Palestinian dialect (PAL) Despite progress in Arabic NLP, significant challenges remain in

[2] For example, the word هَيدي is used in Lebanese to indicate the term "this".

[3] https://idc-dsi.github.io/DiaCorpus/, under 'Tagged dataset'.

	Text	Translated Text
1	بس الأشياء يلي انا اخترتها بحالي.	Only the things I chose myself.
2	وكان لازم في إجا ثمن كثير كبير معه إللي هو يعني بندمش عليه بندمش على الطريق الطويل إللي مرقتها عشان أوصل لوين أنا اليوم أعرف أعبر عن نفسي اليوم بغنية شوي شوي إللي هي بتيجي امبلا تشرح للناس إنه أوكي خذوا نبذة شوي تعالوا نحكي شوي عن شو أنا مرقت تعال أمرقو شوي أعطيكوا أفتحلكو شباك عن عرب الداخل كيف نحنا عايشين على غسل الدماغ إللي نحن منمرقه ونحنا مجبورين نعمل نحنا مجبورين ننسى غسل دماغ عن عينينا.	And it was with a heavy price that I regret. I regret the long way I took in order to get where I am today. I know now to tell myself to slow down, in a way that tells people that, Ok, take a small look and let's talk a little about what I went through, come and I'll open a window for you on what the 48 Arabs have been through, how we live and wipe our tears after what happened, and we are obligated to forget the tears.
3	حلو، اسا ماريا إنت تعلمتِ قصتك حلوة. إنت جاي من بير السبع. تعلمت بمدرسة يهودية وبس إنت ربيتِ يعني أهلك عرب من شفاعمرو أصلهن	Nice, Maria, now your story is sweet. You come from Beer Sheva and studied in a Jewish school. But you were educated, I mean your family are Arabs originally from Shefaraam.
4	فا أكثر من مرة كانوا ينادوني أمي حسين بس هم إللي بيتمسخروا يعني عارفين إنه ماسميش أم حسين إسمي أم محمود فصاروا يقولولي أم محمود, بالضمه حتى.	They used to shout "Um-Hussein" to me more than once, those who mock me. They know my name isn't Um-Hussein, my name is Um Mahmoud, so they started to call me Um-Mahmoud, even in public.
5	وطلعوا على بيروت وصارت الحرب الأهلية ببيروت وبشي مرحلة بالكتاب بتحكي إنه هي لما رجعت على دير ميماس كانت دير ميماس كأنه ولا كأنه في حرب أهلية ولا كأنه عم بيصير إشي والعالم بعده مكمل كل شي تمام، وهناك أنا فهمت إنه هاي هي اعبلين كمان إعبلين هي محل.	They went to Beirut and then the Civil War started and things became bad. It says in the book when she came back to Deir Mimas, it was like there was nothing going on, no war and everybody was minding their business. That's how I understood that Eiblin is also like that. Eiblin is also a place.

Table 1: Examples of transcribed texts from DiaSet. In the rightmost column, we present English translations, which were manually prepared by an Arabic native speaker. We highlight certain words and phrases that are unique to the Palestinian dialect.

addressing the disparities among dialects (Jarrar et al., 2017). Notably, the PAL dialect group has been comparatively under-researched, especially compared to the Egyptian dialect, which has attracted a considerable amount of interest (Shoufan and Alameri, 2015; Jeblee et al., 2014). The PAL dialect presents unique characteristics and challenges related to phonetics, morphology, and vocabulary (Jarrar et al., 2014; Baimukan et al., 2022). Some of the characteristics can be attributed to the ethnic and geographic diversity of PAL speakers, resulting in geographical and ethnical sub-dialects, such as urban, rural, Bedouin, and Druze (Jarrar et al., 2017).

According to Jarrar et al. (2017), Urban dialect itself varies phonologically among major cities that are divided into three primary geographical regions, each with its distinct dialect: (i) Gaza. The dialect of Gaza inhabitants is influenced by the Egyptian dialect due to its geographical proximity to Egypt. Additionally, the migration of some Jaffa residents to Gaza has led to variations in phonetics, morphology, and semantics (Cotter, 2013);
(ii) Arab citizens of Israel. This dialect refers to the one spoken by the Arabs residing within the State of Israel. Their dialect is influenced by the Hebrew language, as evidenced by words such as حَسَّمُه ("ḥassamo"), meaning "He blocked him" and أبرُص ("ubroṣ"), meaning "Make a U-turn" (Amara, 2017); and
(iii) The West Bank. Arabs residing in the West Bank under the governance of the Palestinian Authority also encounter numerous influences from other languages and dialects. Notably, there's a prevalent use of the Jordanian dialect and English (Dibas et al., 2022).

In light of these challenges and divisions, we find that scholarly research in this field lacks focus on the subtle differences between these sub-dialects and how they impact daily communication and interaction among people. Understanding these nuances can have a significant impact on the development of applications that rely on natural language processing. In this context, our work aims to contribute to a deeper understanding of PAL and its sub-dialects and to enhance further research in the field of NLP.

Arabic speech dataset. A significant challenge encountered in this endeavor is the scarcity of appropriate resources, particularly speech datasets in PAL. Most of the existing SA research in Arabic has predominantly focused on written texts. However, spoken Arabic recordings, especially those representing dialectical Arabic variations such as PAL, remain conspicuously limited. Most of the research based on speech datasets was conducted on MSA or dialectical Arabic, but not on PAL (Eldesouki et al., 2019; Al-Azani and El-Alfy, 2019). As a result, our research takes on a pioneering role in this domain by introducing and curating an extensive transcribed PAL corpus. This allows for a more in-depth exploration of various NLP tasks specific to PAL.

SA and NER in Arabic SA and NER tasks in Arabic have garnered significant scholarly focus in recent years (Farha et al., 2021; Shaalan, 2014; Al-Ayyoub et al., 2019; Farha and Magdy, 2019; Ali et al., 2020). Despite this, the literature largely omits detailed consideration of Arabic's dialectal diversity, notably the Palestinian dialect. Boudad et al. (2018) and Inoue et al. (2021b) provide valuable frameworks for SA in MSA and Arabic dialects, yet neither explicitly addresses PAL. This is also evident in studies such as those suggested by Abu Farha and Magdy (2019) that tackle SA in Arabic without specific reference to PAL. Similarly, research introduced by Jarrar et al. (2017) and Al-Mutlaq (2017) explores NER in MSA without extending to PAL nuances.

Arabic LLMs Since the release of BERT by Devlin et al. (2018), large language models (LLMs) have been intensively developed (Wei et al., 2022). Recent years have seen the introduction of Arabic LLMs, which have achieved state-of-the-art (SOTA) results in many NLP tasks. We delve deeper into the primary Arabic LLMs utilized in our research in Section 4.

1. AraBERT (Antoun et al., 2020). This first Arabic transformer language model was originally trained on a ∼24GB worth of texts, corresponding to 2.5B tokens (about 70 million sentences). Most of the texts are new articles collected from several media outlets originating in different geographies. We use the bert-base-arabertv02 model, which is available on Hugging Face.

YouTube Podcasts		Personal Interviews	
Word	Translation	Word	Translation
إنه	That is	شو	What
بدي	I want	هيك	Like that
إشي	Something	اسا	Now
عشان	Because	زي	Like something
شوي	Little	سنين	Years

Table 2: Common words in DiaSet. Examples of the most common words in the YouTube podcasts (left) and the personal interviews (right). The words are taken from the top-100 most common words in each data source.

2. GigaBERT (Lan et al., 2020). A bilingual BERT model trained on texts written in either Arabic or English, consisting of 10.4B tokens collected from known media outlets as well as Wikipedia. The authors of GigaBERT employed a data augmentation process, leveraging English-Arabic dictionaries to artificially create additional training data (Conneau et al., 2019). We use the GigaBERT-v4-Arabic-and-English model, which is available on Hugging Face.

3. MARBERT (Abdul-Mageed et al., 2020). This Arabic BERT model, trained on ∼128GB worth of text, corresponds to 15.6B tokens. MARBERT relies solely on Twitter data from various Arabic dialects. We use the MARBERT model, which is available on Hugging Face.

4. CAMeLBERT (Inoue et al., 2021a) is the latest pre-trained Arabic language model, that achieves SOTA results in multiple NLP tasks. The model has already been fine-tuned for several downstream Arabic NLP tasks, including the two tags we experiment with in this paper.

3. Data

In this section, we present our methodology for the creation of DiaSet. An overview of the data creation flow is outlined in Figure 1. The figure consists of four phases (A-D). Phases A, B-C, and D are detailly explained in Sections 3.1, 3.2, and 3.3 respectively.

3.1. Data Collection

We collect data from two audio sources: (a) YouTube podcasts in spoken PAL; and (b) Simulated interviews in spoken Arabic by PAL native speakers (see Figure 1, Phase A). General statistics describing the YouTube podcasts and the simulated interviews are presented in Table 3.

Overall, DiaSet consists of 644.8K tokens and 62.4K unique tokens. It contains unique dialectical Arabic words that are a result of the way we built the

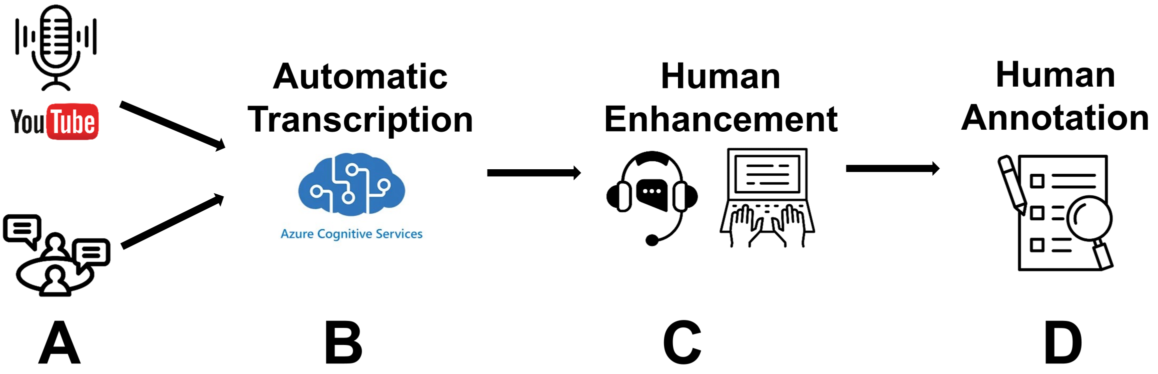


Figure 1: Data creation flow. The letters represent the four different phases of the process.

corpus. For comparison, the largest PAL corpus (Jarrar et al., 2017) contains 56K tokens and 16.4K unique tokens. A short list of interesting common words in DiaSet is presented in Table 2. These terms are distinct, as they are not found in MSA but are unique to dialectical Arabic. Certain words are particularly used among PAL speakers.

YouTube podcasts We collected 83 recorded hours, taken from 55 podcast episodes on YouTube. The podcasts feature dialogues between two native Arabic speakers. We selected these podcasts after a methodical search of YouTube for Arabic-speaking podcasts in PAL. We use the following criteria while selecting these videos: (i) Sufficient number of long enough episodes;[4] (ii) High recording quality; (iii) Variation of subjects discussed; and (iv) The videos have been uploaded under a Creative Commons license at the time of download.[5]

Simulated interviews We conducted personal interviews with native Arabic speakers in PAL. The interviews were conducted by a dedicated interviewer hired specifically for this research. The face-to-face interviews took place in a dedicated office equipped with professional recording equipment.[6] We interviewed 46 people at an average interviewing length of 21 minutes. Interviewees volunteered to participate in the project after signing a participation form. The form signed by the interviewees is presented in the Appendix, Figure 4. All of the interviewees were students between the ages of 18 and 28. We ensured a relatively balanced gender distribution with 26 (56%) males and 20 (44%) females.

All interviewees are native Arabic speakers, fluent in PAL. However, as we mention in Section 1, PAL is diverse and consists of many sub-dialects. Thus, we ensured that the backgrounds of the interviewees were as diverse as possible. The majority of interviewees (60%) are from the northern parts of Israel (e.g., Haifa) while 15% are from the Jerusalem-Ramallah district. The remaining interviewees come from various villages, including Bedouins from the Negev and Druze. In total, we collected 15 recorded hours in this process.

3.2. Transcription

We take a two-step approach to transcribe the audio files we collected. We first use an off-the-shelf automatic transcription system and later manually correct the resulting transcriptions (phases B and C respectively in Figure 1).

Automatic transcription We use the Azure transcription endpoint which is part of the Azure Cognitive Services (Tajane et al., 2018). This service had been chosen following a thorough check of three alternative transcribing systems. Based on our internal evaluation, the Azure engine achieves a 79% word error rate accuracy, significantly outperforming the other alternatives.

The Azure transcription system segments the input data into *text units* (usually a single sentence or a paragraph of 1-3 sentences) based on the discerned speech pauses detected within the audio data. Each text unit consists of 23.3 tokens on average. We use these text units in further phases of the research.

Human enhancement While the Azure system provides the most accurate transcription results, human refinement and validation are still necessary to ensure reliability. Hence, we hired human transcribers to enhance the textual output of the Azure system (see phase C in Figure 1). We make sure that each text unit is enhanced and correctly written

[4] At least 30 episodes, 30 minutes minimum each.

[5] Creative Commons license allows the usage of podcasts' content for any purpose, including academic research.

[6] The recording equipment was made available on a loan basis through a collaborative partnership with a local radio station.

	YouTube Podcasts **(55)**				Personal Interviews **(46)**			
	Total	Mean	Median	STD	Total	Mean	Median	STD
Length (HH:MM)	83:08	01:26	01:34	00:21	15:05	00:21	00:16	00:12
Textual Units	23.3K	435.2	412	203.2	10.3K	223.9	192	63.9
Tokens	540.8K	9.83K	10.88K	3.06K	104.0K	2.26K	1.63K	882.7

Table 3: Data statistics for the YouTube Podcasts and the personal interviews. Numbers in parentheses of the headers indicate the number of episodes. We follow the HH:MM (hours:minutes) length format in the first row.

by two sequential people. In this phase, we use the Microsoft Azure AI Video Indexer,[7] which allows a suitable environment for editing the output content of the transcription system by multiple users.

Spoken Arabic, in any dialect, poses a challenge for transcription/enhancement as it is fundamentally a spoken language lacking well-defined orthographic standards, as opposed to MSA. To achieve a standardized written dataset, we adopt and follow the $CODA^*$ conventions (Shazal et al., 2020) for general rules and exceptions. $CODA^*$ is an extension of the original $CODA$, suggested by Habash et al., 2012 that is designed for the Egyptian dialect, containing sets of rules, exceptions, and conventions that can be used in the transcription of any Arabic dialect. We modified $CODA^*$ for our internal usage to create a detailed guideline for this part of the project.

To carry out the human enhancement work, we hired a team of six native Arabic speakers; all are graduate or undergraduate students from all genders. They are all very familiar with PAL. Each team member was compensated at an hourly rate of $16. The identical team members are engaged for the subsequent phase of the project, human annotation (see Section 3.3 below).

3.3. Human Annotation

Among other things, DiaSet is valuable for continuous pre-training of Arabic LLMs. However, to derive the most of DiaSet, we manually annotate a portion of the data for two common NLP downstream tasks. We refer to the annotated portion of DiaSet as Anno-DiaSet for the remainder of this paper.

Anno-DiaSet consists of 45 podcasts as well as 26 interviews that are randomly sampled from DiaSet. A total of 23.2K text units were manually annotated in Anno-DiaSet, approximately a quarter from interviews and the rest from podcasts. The creation of text units out of each podcast is described in Section 3.2, in the second paragraph. We annotate the dataset for two NLP tasks, which are:

SA Annotators are asked to label each instance by its textual polarity. We use the standard sentiment tagging scheme of positive/neutral/negative classes (Liu et al., 2010).

NER Annotators are asked to identify named entities in the text and classify them into a single category. We use the basic four NER categories suggested by Sang and De Meulder, 2003, that is, PER (person, e.g., Mahmoud Darwish), LOC (location, e.g., Ramallah), ORG (organization, e.g., Al Jazeera), and MISC (miscellaneous, e.g., Ramadan). Named entities are extensively used in DiaSet. Table 1 highlights some of these cases. For instance, Deir Mimas and Eiblin (row 5) are LOC entities, Um-Hussein (row 4) is a PER entity, and 48 Arabs (row 2) is a MISC entity.

To ensure a high-quality annotation of Anno-DiaSet we create a detailed guidelines document that contains explanations, examples, and special cases per annotation task.[8] In the first phase, *all* annotators independently annotated a small set of 100 instances to train and calibrate the guidelines. The guidelines were adjusted to handle cases of annotator disagreements. Only when we reached an acceptable inter-annotation agreement level we asked the annotators to label the entire Anno-DiaSet.

Each instance assigned remains active until tagged by two annotators, once completed, the instance is out of the tagging queue and is ready to be reviewed and analyzed. In case of disagreement, a third annotator is assigned the adjudication task, where they are asked to label only the cases on which the two annotators disagree to have a final decision for each instance. Over the NER task, the pairwise F1 agreement level between annotators (Deleger et al., 2012; Brandsen et al., 2020) is 0.72. Over the sentiment annotation task, the average agreement between annotators is measured to be 68%, corresponding to a kappa (Cohen, 1960) value of 0.38. Importantly, a significant portion (94.6%) of the disagreement cases in the sentiment annotations are between the positive-neutral and the negative-neutral labels rather than

[7] Azure AI Video Indexer: https://vi.microsoft.com/

[8] The guidelines are available on the following Google Doc: https://tinyurl.com/4xes4j6j

the positive-negative labels.

For this part of the project, we use 'Label-Studio' (Tkachenko et al., 2020-2022), an open-source data labeling platform that allows multi-dimensional tagging by a large group of annotators.

We observe the following distribution of the sentiment tag in Anno-DiaSet: 22%, 53%, and 25% for the positive, neutral, and negative labels, respectively. Over the NER tag, 9.7K (42%) of the textual units contain at least a single named entity. We annotated 24,405 named entities in Anno-DiaSet , by the following distribution: 8.6K (35%) as 'LOC', 8.2K (34%) as 'MISC' 4.5K (18%) as 'PER', and 3.1K as 'ORG' (13%).

4. Computational Approaches

To validate the Anno-DiaSet and its usability, we trained multiple machine-learning models and compared their performance with some existing baseline models. As Anno-DiaSet is annotated over two different tasks, we experiment with multiple Arabic models for each task independently.

To get the most out of Anno-DiaSet, we do not split the corpus into train and test sets for evaluation but rather use a five-fold cross-validation approach. We use standard classification evaluation metrics (e.g., F1-score) per fold and report the average value over the five folds.

Specifically, we fine-tune different Arabic BERT (Devlin et al., 2018) models on the two tasks independently for the duration of five epochs. For a detailed explanation of the Arabic LLMs we use, please see Section 2.

SA We fine-tune four Arabic BERT models: CAMelBERT (Inoue et al., 2021a), AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2020), and GigaBERT (Lan et al., 2020). In addition, we implement a bag-of-words (BOW) model, to train a logistic regression classifier. We compare the results of these models with the CAMeL-Lab dialectical sentiment model, available on Hugging-Face.[9]

NER The NER task is to assign a named entity tag per token. A single named entity could span over subsequent tokens within a sentence. We use the IOB format (i.e., Inside, Outside, Beginning) tagging scheme (Lample et al., 2016). We identify 1,200 entity mentions on which at least two annotators disagree on the assigned entity type. Following Mollá et al., 2006 and Su and Yu, 2023, we preserve the multiple labels in such cases.

We fine-tune four Arabic BERT models: CAMel-BERT (Inoue et al., 2021a), AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2020) ,and GigaBERT(Lan et al., 2020). We compare the performance of the models with the dialectical Arabic NER model suggested by CAMel-Lab (Inoue et al., 2021a) available online.[10]

5. Results and Analysis

Sentiment analysis results The obtained results by each model of the SA task are summarized in Table 4. The fine-tuned MARBERT model outperforms other alternatives, over most of the evaluation measures we use. Overall, when comparing the aggregated measures (i.e., Macro F1 and Weighted F1), MARBERT performs best.

We conducted an error analysis for the SA task. Figure 2 is the confusion matrix we get by running the best-performing model on the five cross-validation folds. Interestingly, the negative and positive labels are rarely "mixed up" by the model. In both positive and negative classes, the precision of the model is relatively low (55% and 60%). We argue that this distribution is a result of the skewness of the dataset toward the neutral class.

NER results The results obtained by each model of the NER task are detailed in Table 5. All the fine-tuned BERT models significantly outperform the baseline CAMeLl-Lab pre-trained model. The fine-tuned AraBERT model outperforms other alternatives in all cases. The performance of the models over the 'ORG' category is the lowest compared with other categories. We hypothesize that this is a result of the under-representation of this label in the corpus (13% only).

In Figure 3 we present the NER confusion matrix of the best-performing model. Naturally, the majority of the false-positive and false-negative mistakes are among the 'MISC' category.

Both our sentiment and NER fine-tuned models are available for the research community in the Hugging-Face repository.[11].

6. Limitations

As we elaborate in Section 1, PAL is diverse and consists of many sub-dialects. We put extra effort into making DiaSet diverse by including as many sub-dialects as possible. Still, there are a few underrepresented sub-dialects such as the Bedouins due to lack of online resources and low motivation

[9] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da-sentiment

[10] https://huggingface.co/CAMeL-Lab/bert-base-arabic-camelbert-da-ner

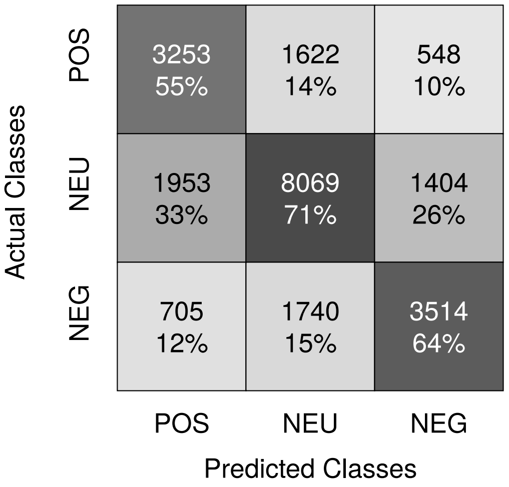
[11] https://huggingface.co/DSI

	Positive			Neutral			Negative			All	
	P	R	F1	P	R	F1	P	R	F1	Macro F1	W-F1
CAMeL-SA	0.50	0.38	0.43	0.66	0.37	0.48	0.38	**0.78**	0.51	0.47	0.47
BOW	0.50	0.39	0.44	0.63	**0.76**	0.69	0.53	0.41	0.46	0.53 ± 0.802	0.57 ± 0.862
CAMeLBERT	0.54	0.55	0.55	**0.71**	0.68	0.69	0.58	0.62	0.60	0.61 ± 0.0086	0.63 ± 0.0081
AraBERT	**0.55**	0.47	0.51	0.69	0.71	0.70	0.57	0.59	0.58	0.60 ± 0.029	0.62 ± 0.0210
GigaBERT	0.54	0.56	0.55	0.69	0.70	0.70	0.60	0.56	0.58	0.61 ± 0.0099	0.63 ± 0.0088
MARBERT	**0.55**	**0.60**	**0.57**	**0.71**	0.71	**0.71**	**0.64**	0.59	**0.62**	**0.63 ± 0.0096**	**0.65 ± 0.0104**

Table 4: Sentiment analysis results. P and R are precision and recall. M-F1 and W-F1 are the macro-F1 and weighted-F1 over the three labels. Results are averaged over the five cross-validation folds. The standard deviation of the overall results is provided in the last two columns. The best results are in boldface. CAMeL-SA is the pretrained sentiment analysis model of CAMeL-Lab (Inoue et al., 2021a).

	PER			LOC			ORG			MISC			All	
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Macro F1	W-F1
CAMeL-NER	0.53	0.40	0.46	0.69	0.41	0.52	0.26	0.12	0.16	0.25	0.03	0.05	0.30	0.28
CAMaEl-BERT	0.77	0.78	0.78	0.77	0.83	0.80	0.59	0.57	0.58	0.69	0.72	0.69	0.72 ± 0.095	0.74 ± 0.076
AraBERT	**0.79**	**0.84**	**0.81**	**0.80**	**0.86**	**0.83**	**0.63**	**0.64**	**0.64**	**0.72**	**0.74**	**0.73**	**0.75 ± 0.096**	**0.77 ± 0.074**
GigaBERT	0.74	0.77	0.75	0.78	0.81	0.80	0.57	0.53	0.54	0.65	0.68	0.67	0.69 ± 0.089	0.71 ± 0.073
MARBERT	0.75	0.80	0.75	0.76	0.80	0.79	0.55	0.50	0.50	0.66	0.70	0.67	0.67 ± 0.061	0.70 ± 0.048

Table 5: NER results. P and R are precision and recall. M-F1 and W-F1 are the macro-F1 and weighted-F1 over the four categories. Results are averaged over the five cross-validation folds. The standard deviation of the overall results is provided in the last two columns. The best results are in boldface. CAMeL-NER is the pretrained NER model of CAMeL-Lab (Inoue et al., 2021a).



Actual Classes	POS	NEU	NEG
POS	3253 55%	1622 14%	548 10%
NEU	1953 33%	8069 71%	1404 26%
NEG	705 12%	1740 15%	3514 64%

Figure 2: Confusion matrix for the best-performing SA model. POS, NEU, and NEG are the positive, neutral, and negative labels, respectively. The percentage number is calculated columnwise.

to participate in the personal interviews we conducted.

Anno-DiaSet is annotated over only a portion of DiaSet and by the sentiment and named-entity solely. However, the entire dataset can potentially be annotated and even over a larger choice of labels, designed for other NLP tasks (e.g., coreference resolution). In addition, we believe that by making Anno-DiaSet larger, we would observe better performance over both tasks.

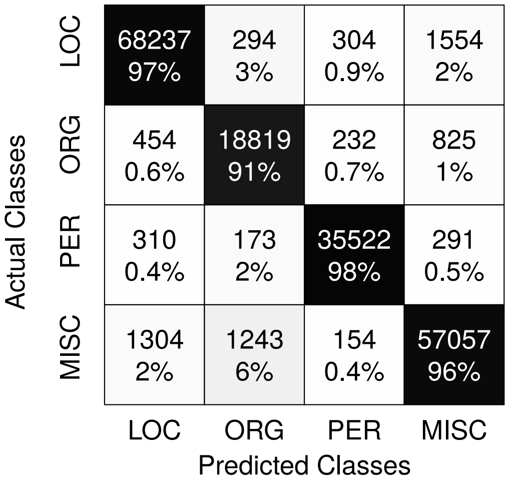
While the data in hand are conversations between speakers, we did not make use of this while modeling. We treat each text unit independently. We believe that taking into account the conversational structure of the data and building models that take this input into account, will probably outperform the results we obtained.

7. Discussion and Conclusions

In this work, we presented DiaSet, a new dataset of Arabic speech transcripts. It is constructed from two sources: YouTube podcasts and simulated interviews. Both sources feature spoken dialectical Arabic, which constitutes the central uniqueness of the dataset. However, the two sources differ by nature. The podcast episodes are longer, contain longer text units, have a larger vocabulary (see Table 3), and include different discussion topics. In this work, we do not treat each source independently but rather both as a whole. We leave this as research direction for future consideration. Overall, DiaSet comprises more than 600K tokens, corresponding to 33.6K textual units (i.e., sentences or paragraphs). Of these, 23.2K have been annotated for two common downstream tasks.

Both the podcasts and simulated interviews feature conversations between two participants, making them valuable for conversational linguistic analysis in Arabic. All the conversations and interviews released in this work are labeled with a speaker index, denoting either the first or the second speaker.

In future work, we aim to augment the dataset with conversations from further sources, showcas-



	LOC	ORG	PER	MISC
LOC	68237 97%	294 3%	304 0.9%	1554 2%
ORG	454 0.6%	18819 91%	232 0.7%	825 1%
PER	310 0.4%	173 2%	35522 98%	291 0.5%
MISC	1304 2%	1243 6%	154 0.4%	57057 96%

Figure 3: Confusion matrix for the best-performing NER model. The percentage number in each cell is calculated columnwise. We focus only on entity labels and remove the 'O' category for better readability of the matrix.

ing dialectical Arabic specific to the pertinent locale. Additionally, we intend to annotate the existing conversations for more downstream tasks. We plan to leverage the most recently published LLM, Jais (Sengupta et al., 2023), to set a new zero-shot baseline for NER and SA on speech transcriptions.

8. Ethical Considerations

DiaSet consists of conversations between humans. We ensure the anonymity of the speakers when sharing DiaSet with the research community, omitting both explicit names and metadata about them. The interviewees in the simulated interviews were asked about the topics they wished to discuss. We ensure that all interviewees agree to share the textual content of their interviews with the research community by signing a participation form. The form includes a declaration that the interviewee may ask to remove their interview content from DiaSet for any reason. Our student annotators were paid $16 per hour, which is considered above average in Israel.

9. Bibliographical References

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2020. Arbert & marbert: Deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

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, Florence, Italy. Association for Computational Linguistics.

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N Al-Kabi. 2019. A comprehensive survey of arabic sentiment analysis. *Information processing & management*, 56(2):320–342.

Sadam Al-Azani and El-Sayed M El-Alfy. 2019. Audio-textual arabic dialect identification for opinion mining videos. In *2019 IEEE symposium series on computational intelligence (SSCI)*, pages 2470–2475. IEEE.

Anhar Al-Mutlaq. 2017. Tebyan: Interactive spelling correction application for quranic verse. In *2017 9th IEEE-GCC Conference and Exhibition (GCCCE)*, pages 1–9. IEEE.

Brahim Ait Ben Ali, Soukaina Mihi, Ismail El Bazi, and Nabil Laachfoubi. 2020. A recent survey of arabic named entity recognition on social media. *Rev. d'Intelligence Artif.*, 34(2):125–135.

Zaid Alyafeai, Maraim Masoud, Mustafa Ghaleb, and Maged S. Al-shaibani. 2021. Masader: Metadata sourcing for arabic text and speech data resources.

Muhammad Amara. 2017. *Arabic in Israel: Language, identity and conflict*. Routledge.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Naaima Boudad, Rdouan Faizi, Rachid Oulad Haj Thami, and Raddouane Chiheb. 2018. Sentiment analysis in arabic: A review of the literature. *Ain Shams Engineering Journal*, 9(4):2479–2490.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

William M Cotter. 2013. Dialect contact and change in gaza city. *Unpublished MA thesis. Colchester, UK: University of Essex*.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R El-Beltagy, Wassim El-Hajj, et al. 2021. A panoramic survey of natural language processing in the arab world. *Communications of the ACM*, 64(4):72–81.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shahd Dibas, Christian Khairallah, Nizar Habash, Omar Fayez Sadi, Tariq Sairafy, Karmel Sarabta, and Abrar Ardah. 2022. Maknuune: A large open palestinian arabic lexicon. *arXiv preprint arXiv:2210.12985*.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Mohamed Eldesouki, Naassih Gopee, Ahmed Ali, and Kareem Darwish. 2019. Farspeech: Arabic natural language processing for live arabic speech. In *INTERSPEECH*, pages 2372–2373.

Ashraf Elnagar, Sane Yagi, Ali Bou Nassif, Ismail Shahin, and Said A. Salloum. 2021. Sentiment analysis in dialectal arabic: A systematic review. In *Advanced Machine Learning Technologies and Applications*, pages 407–417, Cham. Springer International Publishing.

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.

Ibrahim Abu Farha and Walid Magdy. 2022. The effect of arabic dialect familiarity on data annotation. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 399–408.

Ibrahim Abu Farha, Wajdi Zaghouani, and Walid Magdy. 2021. Overview of the wanlp 2021 shared task on sarcasm and sentiment detection in arabic. In *Proceedings of the sixth Arabic natural language processing workshop*, pages 296–305.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021a. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021b. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A morphological analyzer for egyptian arabic. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology*, pages 1–9.

Nizar Y Habash. 2022. *Introduction to Arabic natural language processing*. Springer Nature.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021a. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Online). Association for Computational Linguistics.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021b. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Abraham Israeli, Aviv Naaman, Yotam Nahum, Razan Assi, Shai Fine, and Kfir Bar. 2022. Love me, love me not: Human-directed sentiment analysis in arabic. In *Proceedings of the Third International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2022) co-located with ICNLSP 2022*, pages 22–30.

Abraham Israeli, Yotam Nahum, Shai Fine, and Kfir Bar. 2021. The idc system for sentiment classification and sarcasm detection in arabic. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 370–375.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian Arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27, Doha, Qatar. Association for Computational Linguistics.

Mustafa Jarrar, Mohammed Khalilia, and Sana Ghanem. 2022. Wojood: Nested arabic named entity corpus and recognition using bert. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3626–3636.

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into egyptian arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 196–206.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Wuwei Lan, Yang Chen, Wei Xu, and Alan Ritter. 2020. Gigabert: Zero-shot transfer learning from english to arabic. In *Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP)*.

Bing Liu et al. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Xiaoye Qu, Yingjie Gu, Qingrong Xia, Zechang Li, Zhefeng Wang, and Baoxing Huai. 2023. A survey on arabic named entity recognition: Past, recent advances, and future trends. *arXiv preprint arXiv:2302.03512*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Khaled Shaalan. 2014. A survey of arabic named entity recognition and classification. *Computational Linguistics*, 40(2):469–510.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the fifth arabic natural language processing workshop*, pages 167–177.

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.

Abu Bakr Soliman, Kareem Eissa, and Samhaa R El-Beltagy. 2017. Aravec: A set of arabic word embedding models for use in arabic nlp. *Procedia Computer Science*, 117:256–265.

Jindian Su and Hong Yu. 2023. Unified named entity recognition as multi-label sequence generation. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Kapil Tajane, Saransh Dave, Pankaj Jahagirdar, Abhijeet Ghadge, and Akash Musale. 2018. Ai based chat-bot using azure cognitive services. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–4. IEEE.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

WorldData. 2023. Arabic speaking countries. `https://www.worlddata.info/languages/arabic.php#google_vignette`. Accessed: 2023-08-30.

10. Language Resource References

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical aggregation of dialectal data for arabic dialect identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596.

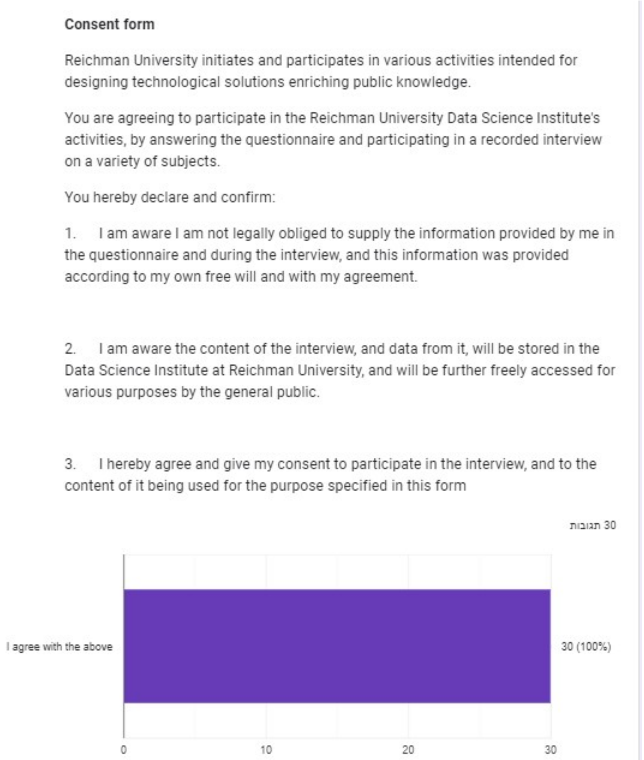
Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4573–4577.

Mustafa Jarrar, Nizar Habash, Diyam Akra, and Nasser Zalmout. 2014. Building a corpus for palestinian Arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27, Doha, Qatar. Association for Computational Linguistics.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2017. Curras: an annotated corpus for the palestinian arabic dialect. *Language Resources and Evaluation*, 51:745–775.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. Camel tools: An open source python toolkit for arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032.

A. Appendix



Consent form

Reichman University initiates and participates in various activities intended for designing technological solutions enriching public knowledge.

You are agreeing to participate in the Reichman University Data Science Institute's activities, by answering the questionnaire and participating in a recorded interview on a variety of subjects.

You hereby declare and confirm:

1. I am aware I am not legally obliged to supply the information provided by me in the questionnaire and during the interview, and this information was provided according to my own free will and with my agreement.

2. I am aware the content of the interview, and data from it, will be stored in the Data Science Institute at Reichman University, and will be further freely accessed for various purposes by the general public.

3. I hereby agree and give my consent to participate in the interview, and to the content of it being used for the purpose specified in this form

Figure 4: The consent form. The interviewees signed the form before participating in the personal interview sessions.