ZAEBUC-Spoken: A Multilingual Multidialectal Arabic-English Speech Corpus

Injy Hamed,[1,2] **Fadhl Eryani,**[1,3] **David Palfreyman,**[4,5] **Nizar Habash**[1]
[1]New York University Abu Dhabi, [2]University of Stuttgart,
[3]University of Tübingen, [4]United Arab Emirates University, [5]Zayed University, Abu Dhabi, UAE
{injy.hamed, fadhl.eryani, nizar.habash}@nyu.edu, dpalf@uaeu.ac.ae

Abstract

We present ZAEBUC-Spoken, a multilingual multidialectal Arabic-English speech corpus. The corpus comprises twelve hours of Zoom meetings involving multiple speakers role-playing a work situation where Students brainstorm ideas for a certain topic and then discuss it with an Interlocutor. The meetings cover different topics and are divided into phases with different language setups. The corpus presents a challenging set for automatic speech recognition (ASR), including two languages (Arabic and English) with Arabic spoken in multiple variants (Modern Standard Arabic, Gulf Arabic, and Egyptian Arabic) and English used with various accents. Adding to the complexity of the corpus, there is also code-switching between these languages and dialects. As part of our work, we take inspiration from established sets of transcription guidelines to present a set of guidelines handling issues of conversational speech, code-switching and orthography of both languages. We further enrich the corpus with two layers of annotations; (1) dialectness level annotation for the portion of the corpus where mixing occurs between different variants of Arabic, and (2) automatic morphological annotations, including tokenization, lemmatization, and part-of-speech tagging.

Keywords: Speech Corpus, Arabic, English, Multilinguality, Arabic Dialects, Code-switching

1. Introduction

Remarkable strides have been recently made in language technologies for distinct, standardized languages. These achievements, however, are not equally met for the vast majority of discourse communities (Ranathunga and de Silva, 2022), including the widespread phenomenon of code-switching (Doğruöz et al., 2021), which involves the mixing between languages. One main bottleneck that hinders advancements for many languages is the lack of data needed for training NLP models, and more essentially, for evaluation.

In the scope of Arabic, which is a diglossic language (Ferguson, 1959), language technologies are usually better suited to the formal language, Modern Standard Arabic (MSA), and less proficient for regional dialects. On top of the challenges posed by diglossia, Arabic speakers code-switch between MSA and dialects as well as between Arabic and other languages. The former code-switching type is usually limited to formal settings. The latter type is prevalent among Arab countries, where code-switching is typically seen between dialectal Arabic and English or French, or both.

The work presented here is part of the Zayed University Arabic-English Bilingual Undergraduate Corpus (ZAEBUC) Project, which is interested in the study of bilingualism. As an extension to the previously collected ZAEBUC corpus (Habash and Palfreyman, 2022; Palfreyman and Habash, 2022), which focused on bilingual writers, we present a new corpus that is focused on the spoken domain, offering a resource that encapsulates interesting challenges and linguistic phenomena of bilingual speakers. ZAEBUC-Spoken corpus is collected through Zoom meetings, with multiple speakers taking part in the conversation. The corpus is multilingual, containing (accented) English, MSA, and two Arabic dialects; Gulf and Egyptian, including speakers from six nationalities. The speakers also code-switch between the four mentioned languages. The corpus includes manual transcriptions of the recordings and dialectness level annotations for the portion containing code-switching between Arabic variants, in addition to automatic morphological annotations. ZAEBUC-Spoken corpus offers a challenging set to ASR systems given its spontaneous conversational speech nature, as well as an interesting setup to examine the interaction between diverse bilingual speakers. We make the corpus publicly available.[1]

Next, we discuss related work in §2. In §3 and §4, we elaborate on the data collection process and the translation guidelines. §5 presents overall corpus statistics. In §6, we provide code-switching analyses, including dialectness level statistics. §7 presents the morphological analysis based on the automatic annotations.

2. Related Work

In this section we discuss related work with regards to dialectal Arabic speech corpora, Arabic code-switched speech corpora, and morphologically annotated dialectal Arabic corpora.

[1]http://www.zaebuc.org

LREC-COLING 2024, pages 17770–17782
20-25 May, 2024.

With regards to **dialectal Arabic speech corpora**, numerous efforts provided dialectal Arabic resources for speech recognition. The MGB-2 Arabic challenge dataset (Ali et al., 2016) comprising of 1,200 hours of speech gathered from Aljazeera Arabic TV channel contains a portion of dialectal Arabic. While the majority of the corpus is MSA speech, a subset of the corpus (estimated 30%) contains dialectal Arabic speech, including Egyptian, Gulf, Levantine, and North African dialects. Almeman et al. (2013) also collected the Multi Dialect Arabic Speech Parallel Corpora, containing around 32 hours of speech covering MSA as well as Gulf, Egyptian and Levantine dialects. Other efforts have targeted specific dialects. For Egyptian Arabic, several corpora are available, including CALLHOME Egyptian Arabic corpus (Gadalla et al., 1997) and MBG-3 (Ali et al., 2017). For Gulf Arabic, the Gulf Arabic Conversational Telephone Speech corpus (Appen, 2006a,b) consists of around 46 hours of spontaneous Gulf Arabic speech obtained from telephone conversations. Elmahdy et al. (2014) also present the 15-hour QA corpus, collected from TV series and talk show programs. Other corpora have covered other dialects including Moroccan Arabic (Ali et al., 2019) and Levantine Arabic (BBN Technologies et al., 2005; Appen, 2007; Maamouri et al., 2007).

From the perspective of **code-switched Arabic speech corpora**, the presented corpus offers an interesting setup, having two different scenarios of code-switching produced by Arabic speakers; code-switching between Arabic and foreign languages and code-switching between Arabic variants. Similar to our corpus, other researchers have also collected code-switched corpora, however, the corpora usually focus on either of the two code-switching scenarios. Ismail (2015) gathered 89 minutes of speech containing Saudi Arabic-English code-switching through informal dinner gatherings. Amazouz et al. (2018) collected a 7.5 hour corpus containing code-switched Algerian Arabic-French gathered from informal conversations as well as read speech of books and movie transcripts. Hamed et al. (2020, 2022) collected the ArzEn corpus containing 12 hours of code-switched Egyptian Arabic-English speech through informal interviews, and commissioned their English translations. Chowdhury et al. (2021) presented the Economic and Social Commission for West Asia (ESCWA) corpus containing 2.8 hours of United Nations meetings. The corpus contains code-switching between Arabic (including different dialects) and English/French. In Mubarak et al. (2021), 2,000 hours of speech were collected from Aljazeera news channel, where 0.4% of the corpus (∼6,000 utterances) have code-switching between Arabic and English/French. In the direction of covering code-switching between dialectal Arabic and MSA with language identification, Chowdhury et al. (2020) annotate a 2-hour subset from ADI-5 development set in the MGB-3 challenge (Ali et al., 2017) containing Egyptian Arabic-MSA code-switching for word-level language identification.

With regards to **morphologically annotated dialectal Arabic corpora**, the amount of available resources varies across dialects. Egyptian Arabic, receiving significant attention, is supported by several corpora, including the Egyptian Colloquial Arabic Lexicon (Kilany et al., 2002) and the Egyptian Arabic Treebank (Maamouri et al., 2012; Maamouri et al., 2014). Less corpora are available for Gulf Arabic. Khalifa et al. (2016, 2018) collected the Gumar corpus from 1,200 forum novels and extended a subset of the corpus with morphological annotations as well as orthographic modifications following the Conventional Orthography for Dialectal Arabic (CODA) guidelines (Habash et al., 2018). Habash and Palfreyman (2022) also collected the ZAEBUC corpus, an Arabic-English bilingual writer corpus containing short essays written by first-year university students along with assigned writing proficiency ratings. The corpus was manually corrected for spelling and grammar errors, and annotated for morphological tokens, part-of-speech (POS) tags, and lemmas. Other corpora exist for other dialects including Jordanian Arabic (Maamouri et al., 2006), Palestinian Arabic (Jarrar et al., 2016), as well as Moroccan and Sanaani Yemeni Arabic (Al-Shargi et al., 2016). In the context of code-switching, previous efforts also targeted collecting code-switched Egyptian Arabic-English corpora with morphological annotations including morphological segmentations, POS tags and lemmatization (Balabel et al., 2020; Gaser et al., 2022).

In comparison to previously mentioned corpora, our **contribution** stands out on several dimensions. Firstly, the corpus is obtained through Zoom meetings, which opens up possibilities for extending the corpus to other tasks, such as meeting summarization. The corpus also contains multiple Arabic variants (MSA, Gulf Arabic, and Egyptian Arabic), accented English, as well as code-switching across the languages and dialects. We also provide automatic morphological annotations, including tokenization, lemmatization and POS tagging. Unlike the previously mentioned corpora where such annotations were provided for textual data, spontaneous speech data introduces challenges to POS tagging, as the normal flow of sentences may be broken due to repetitions and/or corrections. We currently only provide automatic annotation, which we plan on extending with manual revisions. Moreover, as part of our work, we present our transcription guidelines handling issues arising due to the spontaneous nature of the corpus as well as code-switching.

3. Data Collection

The recordings were collected through Zoom meetings in which two Students and an Interlocutor simulated a work situation relevant to the students' major. Topics included *employee health and wellbeing*, *studying abroad*, *arts and design*, *SWOT analysis*, *advertising*, and *tourism*. The Students were asked to prepare ideas to present to an Arabic-speaking or English-speaking Interlocutor of senior status (e.g., manager, dean, etc., depending on the topic). Afterwards, the Interlocutor, a person unknown to the Students, joined them to hear about and discuss their ideas. The interactions were set up by one of the authors, referred to as the Moderator. Each meeting consists of four phases:

1. ***Phase 1*:** The Moderator introduces the task to both Students, showing an email from the Interlocutor requesting their ideas for a specific purpose. This phase is conducted in English, except where Students are asked to read aloud a task which is written in MSA for an Arabic-speaking Interlocutor.
2. ***Phase 2:*** The two Students discuss the task together. They are allowed to converse in any language, as they prefer. This phase usually contains a mix of Gulf Arabic and English.
3. ***Phase 3*:** The Students present their ideas to the Interlocutor, who stimulates further discussion of the task at hand. There are two options for this phase, where the Interlocutor talks in either English or MSA. In the first case, the Students use the English language; in the second case, the Students are allowed to choose between MSA and dialectal Arabic. In the case of Arabic-speaking Interlocutors, we observe code-switching between Arabic and English in addition to code-switching between MSA and dialects which arises due to the Egyptian Interlocutors primarily speaking in MSA with slight use of Egyptian Arabic and Students using a mixture of MSA and Gulf Arabic.
4. ***Phase 4*:** The Moderator ends the meeting. This phase is conducted in English.

A total of 14 meetings were conducted, with equal distribution among both setups; Arabic-speaking and English-speaking Interlocutors. Overall, 16 Students took part in the recordings, where each Student participated in a maximum of two meetings; one with an Arabic-speaking and the other with an English-speaking Interlocutor.[2] We also include a 15^{th} recording as part of the corpus, which was a pilot recording collected as part of the development process, only consisting of *Phase 2*. The Zoom meetings are audio recorded, with the audio input of all participants saved in a single audio file. We also obtain separate single-channel recordings for participants' audio streams, where the audio from each participant is saved as a separate file. This is obtained for 11 out of the 15 recordings, due to technical reasons. This setup allows researchers working on speech recognition to utilize the corpus as needed with regards to overlapping speech.

Overview of the Participants: Across all meetings, there is one English-speaking Moderator. All Students are Emirati and all but one are female. The Students are enrolled in different majors, with four Students in *International Studies*, three in *Communication and Media*, two in *Finance*, two in *Animation Design* and the remaining Students each in *Psychology*; *Public Health and Nutrition*; *Multimedia Design*; *Interior Design*; *Marketing and Entrepreneurship*; *Human Resources*; and *Accounting*. With regards to the Interlocutors, there are two Egyptian Arabic speakers and six English speakers coming from different countries: United Kingdom (3), Greece (1), Austria (1), and China (1), introducing a range of different English accents.

4. Transcription

The collected recordings were manually transcribed. The transcribers were allowed to choose between using Praat (Paul Boersma and David Weenink, 2022) or ELAN (Brugman et al., 2004). In both cases, four tiers were used, with each corresponding to one of the four speakers (Moderator, two Students, and Interlocutor), as shown in Figure 1. The annotation team on the project comprised of three Arabic-English bilingual annotators: *A1* (Emirati Arabic native speaker), *A2* (Egyptian Arabic native speaker), and *A3* (Yemeni Arabic native speaker with extensive residency in the UAE).

All recordings were initially transcribed by annotator *A1*. The annotation at this round also involved segmentation of utterances, where an utterance is defined as a segment of speech that is grammatically and semantically complete (except for unfinished or interrupted utterances). The preferable minimum length of an utterance is 10 seconds, and the maximum is 30 seconds unless it is not possible to divide the segment. To ensure transcription quality, each recording underwent two rounds of revision performed by annotators *A3* followed by *A2*. The final checks demonstrate good quality of transcriptions and minimal errors. The recordings and transcriptions are made available, in addition to ASR files following Kaldi (Povey et al., 2011) data preparation format to further facilitate the use of the corpus for speech recognition purposes.[1]

[2] Information about the recordings and participants are released as part of the corpus metadata.

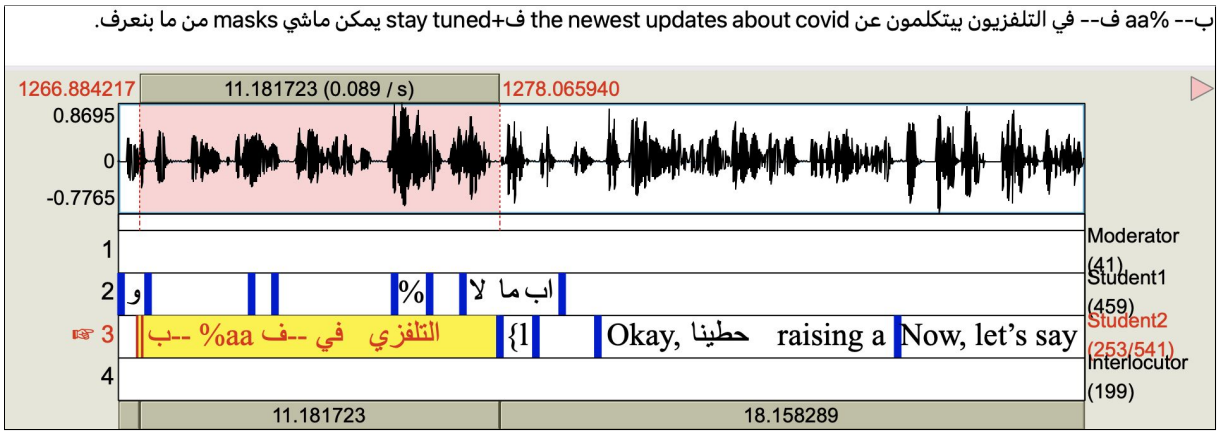

Figure 1: Example of transcription using Praat.

4.1. Transcription Guidelines

In this section, we discuss the transcription guidelines. The guidelines cover four categories; general transcription rules, conversational speech transcription rules, code-switching transcription rules, and orthography rules for English and Arabic. We mostly rely on Callhome (Gadalla et al., 1997) for the general and conversational speech transcription rules, on ArzEn speech corpus (Hamed et al., 2020) for rules related to code-switching, on the SRI Speech-based Collaborative Learning Corpus (SBCLC) (Richey et al., 2016; Richey et al., 2019) for rules related to transcribing English words, and on CODA guidelines (Habash et al., 2018) for dialectal Arabic orthography decisions. We mark transcription decisions that are based on Callhome, SBCLC and ArzEn corpora with H, C , and A, respectively. Afterwards, we discuss CODA guidelines and elaborate on specific decisions that were made within the scope of our corpus.

4.1.1. General Transcription Rules (GR)

[*GR-punctuation*] For punctuation, transcribers are requested to use punctuation as they see fit.

[*GR-numbers*]^H Numbers are written in full text, rather than digits.

[*GR-background*]^H For background noise and typing sounds, the following annotation is used: *[/noise] transcription [noise/]* and *[/typing] transcription [typing/]*, where these tags surround the transcriptions with overlapping background sound. If the sound does not overlap with transcriptions, the annotation is inserted with empty transcription, for example [/noise] [noise/]. We also use the following annotation *[/reading] transcription [reading/]* to denote that a person is reading text aloud.

[*GR-unclear*]^H For unclear words, transcribers are requested to listen to the audio multiple times, including making use of the single channels. The transcription of the unclear words is placed between double parentheses, as *((transcription))*. In case the words are unclear due to corruption in the audio file, this is marked as *(([transcription]))*. In case the transcribers are not able to make a guess, the parentheses are placed with empty transcription.

[*GR-mispronounciation*]^H If a speaker mispronounces a word, the intended word is transcribed, regardless of its pronunciation, and is surrounded by equal sign as *=mispronounced word=*.

[*GR-newTerm*]^H If a speaker comes up with a new term, the term is transcribed between double asterisks as ***term***.

4.1.2. Conversational Speech Rules (CR)

[*CR-repetition*]^H The transcription needs to be an exact reflection of what was said, including repetitions and incomplete words.

[*CR-partial*]^H Partial or incomplete words are marked with '$--$', such as *arch*$--$ and $--$*tective*.

[*CR-flow*]^A Given the spontaneous nature of the corpus, it is common to find disfluencies that break the flow of sentences, including repetitions, corrections, and changing course mid-sentence. We use double dots (..) within text to mark such cases as needed to help with the readability of the transcriptions. We also use (..) at the end of unfinished utterances when the speaker stops mid-sentence.

[*CR-interruption*] We mark short interruptions within text with a tilde ($\sim$). We also use ($\sim$) to mark the end of unfinished utterances due to the speaker being interrupted.

[*CR-nonSpeech*]^H For non-speech segments, we use the following tags: {laugh}, {cough}, {sneeze}, {breath}, {lipsmack}, {pause}, {gasp}, {shush}, and {hem} (clearing throat).

[*CR-interjections*]^H Interjections should be preceded with a percentage sign (%). We use the following closed set of Arabic and English interjections: %oh,أوه%, %aa, آآ%, %um, ام%, %mm, مم%, %hm, هم%, %ah, آه%, %aha, أها%, %ehm, إهم%, %nn,إن%, %ha, ها%, %tt, تت%, %er, ار%, and %oops. In case of extra long interjections (such as 'Oooooh'), a colon is added to the interjection, such as '%oh:'. The choice of script for the interjection is based on the main language of the utterance.

4.1.3. Code-switching Rules (CSR)

[*CSR-script*]^A Arabic words are written in Arabic script and English words are written in Latin script. For English words that are commonly used in Arabic (loanwords), the choice of script depends on the pronunciation. If the word is commonly used in Arabic and pronounced in Arabic accent, then Arabic script is used. If commonly used and pronounced in non-Arabic accent, Latin script is used.

[*CSR-MCS*]^A For morphologically code-switched words, where Arabic affixes and/or clitics are attached to English words, the following annotation is used: <Arabic Morphemes in Arabic script>+<English word in Latin>+<Arabic Morphemes in Arabic script>, e.g., وا+implement+ي *y+implement+wA* 'they implement'.

4.1.4. English Orthography Rules (ENR)

[*ENR-dictionary*]^C American English spelling is used throughout transcriptions, and grammatical errors are not corrected.

[*ENR-contractions*]^C Standard contractions are used when the contracted pronunciation is used, otherwise, the complete form is used. This also applies to non-standard contractions like 'gonna' and 'wanna'.

[*ENR-acronyms*]^C If an acronym is pronounced like a word, it is written in uppercase without spaces, such as AIDS. Acronyms pronounced as the individual letters are written in uppercase where letters are separated by underscores, such as U_A_E.

[*ENR-letters*]^C When the speaker utters single letters, including the case of spelling out a word, the letters are transcribed separately in upper case, for examples: 'We will go for plan B.'

4.1.5. Arabic Orthography Rules

As described in Section 3, the Arabic component of this corpus comprises MSA, dialectal Arabic, and a rich mix of both. As the official language of all Arab countries, MSA enjoys a well-defined official orthographic standard which we follow. For dialectal speech, we follow the Conventional Orthography for Dialectal Arabic (CODA), which is an on-going effort to specify orthographic conventions for a growing number of Arabic dialects. CODA has been used in a number of large-scale Arabic Dialect projects (Habash et al., 2018; Bouamor et al., 2018; Khalifa et al., 2018).

A core precept in CODA is to spell root radicals using an MSA cognate as a reference, according to a defined list of the most common sound to letter correspondences in Arabic.[3] This follows from the observation that when writing dialectal Arabic without conventional rules, spelling tends to reflect a tension between spelling according to the phonology of a given utterance on the one hand, and the spelling of a closely related MSA cognate on the other. For instance the Gulf word /w aa y i d/[4] meaning 'very' or 'a lot', may be spelled وايد *wAyd*, reflecting its phonology, or واجد *wAjd*, reflecting its MSA cognate.[5]

CODA regulates this tension by prioritizing the use of MSA cognates as references, more or less familiar to all Arabic speakers. At the same time CODA aims to "strike an optimal balance between maintaining a level of dialectal uniqueness and establishing conventions based on MSA-DA similarities" (Habash et al., 2018). One way this balance is struck is by allowing specific rules for certain morphemes, often highly marking, such as allowing the pronominal 2nd person feminine clitic when it is pronounced /tsh/ to be spelled with ج *j*, as most Gulf writers prefer, such as in the word /3 i n d i tsh/ عند+ج *nd+j*, 'with you [fs]'. Beyond root radicals, CODA also regularizes templatic pattern spelling according to an MSA reference modulo regular minor sound changes, e.g.: /(2 i) t. t. a w w i r/ meaning 'to develop' is spelled according to how it would be spelled in MSA, تطور *tTwr*, instead of how it is pronounced *اتطور *AtTwr*.

Efforts employing CODA have thus far dealt with text based data, but this project marks the first time annotators have used it for representing a speech corpus. This has called for additional specifications on two CODA rules. The first rule involves hamzas (glottal stops) that appear at the beginning

[3] coda.camel-lab.com/#4321-root-radical-spelling

[4] CAPHI phonetic scheme (Habash et al., 2018).

[5] HSB transliteration scheme (Habash et al., 2007).

of basewords as part of Alif-Hamza letters (i.e., أإآ *ÂĂ*). Whereas previous CODA annotations have stripped baseword initial hamzas (to ا *A*), our annotations spell out these hamzas when they are *audibly pronounced* in the recording. The second rule involves the spelling of particles which have alternative proclitic forms with *shortened vowel*, e.g., شو *šw* /sh uu/ and +ش *š+* /sh u/ 'what'. In CODA both forms are valid and depend on vowel length, which may not always be evident. As such we opted to always spell their non-clitic elongated form.

4.1.6. Redacting Participants' Names

In the public release, we redact the mentions of participants' names during the meetings, to protect the participants' identities. This is done on the word-level, where the name mentions are bleeped in the audio files and the names in the transcriptions are replaced with the following references: <Moderator>, <Student1>, <Student2>, and <Interlocutor>.

4.2. Dialectness Level Annotation

As previously mentioned, the corpus contains code-switching between Arabic and English mainly occurring in *Phase 2* as well as between MSA and dialects occurring in *Phase 3*. While the former type of code-switching can be automatically detected as Arabic and English use different scripts which is maintained by our transcription guidelines, the distinction between MSA and dialectal words in the latter code-switching type is a challenging task. This is not only due to the shared script, but also due to the shared vocabulary. There have been a number of efforts on defining and measuring the degree of dialectness in Arabic (Habash et al., 2008; Zaidan and Callison-Burch, 2011; Keleg et al., 2023). In this work, we follow the guidelines introduced in Habash et al. (2008) where five levels are defined for annotating Arabic dialectness:

L0 denotes perfect MSA.

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

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

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

L4 denotes pure dialect.

Category	Value
# Speakers	27
# Moderator	1
# Students	16+2(pilot)
# Interlocutors	8
# Meetings	15
Total Duration (h)	11.9
Speech Duration (h)	10.5
Average Meeting Duration (h)	0.8
# Utterances	6,033
# Tokens	94,101

Table 1: Corpus Size Overview

We perform manual dialectness level annotation on the utterances in *Phase 3* for the seven recordings having Arabic-speaking Interlocutors. We only annotate *Phase 3*, as it is the phase containing MSA-dialectal Arabic code-switching. The annotators are asked to listen to each utterance and annotate it according to the guidelines mentioned above. The annotation is placed at the start of the utterance transcription followed by '||'. For example, for an utterance identified as *L2*, the annotation is: "L2|| *transcription*". The annotation is performed by annotator *A1* and afterwards revised by annotator *A2*. Cases of disagreements, comprising 22.5% of utterances, were annotated by annotator *A3*, providing the final decisions for the annotations.

5. Corpus Statistics

Table 1 presents general corpus statistics, including the number of speakers, the corpus size in hours, and the total number of tokens and utterances. Among utterances containing language-specific words (not solely annotations), the average utterance duration is 7.2 seconds containing on average 17.7 tokens. Table 2 presents token-level and utterance-level statistics across recordings with Arabic- and English-speaking Interlocutors. On the token-level, we report the number of Arabic, English, and morphologically code-switched (MCS) words and partial words. We also report counts for punctuation, digits, and several annotations. On the utterance-level, we report the counts for monolingual Arabic, monolingual English, and code-switched Arabic-English utterances as well as those composed only of annotations. For the code-switched utterances, we also present a breakdown of their counts according to the extent of code-switching, measured as the percentage of English words in the utterance. With regards to *GR-unclear* annotations due to corruption in the audio file, it is found in 152 utterances (2.5% of the corpus), with 174 instances of the annotation.

Type	Rec-Ar	Rec-En	Total
Token-level Analysis			
Arabic words	24,301	3,571	28,515
English words	9,692	33,059	43,767
MCS words	233	76	346
Arabic partial words	932	81	1,015
English partial words	71	241	317
MCS partial words	4	3	7
Punctuation	5,078	5,741	11,112
Digits	0	0	0
GR-background	96	126	238
GR-unclear	708	826	1,610
GR-mispronounciation	118	80	198
GR-newTerm	0	0	0
CR-flow	598	730	1,366
CR-interruption	137	160	318
CR-nonSpeech	291	187	500
CR-interjections	2,571	2,155	4,792
Total words (full+partial)	35,233	37,031	73,967
Total tokens	44,830	47,036	94,101
Utterance-level Analysis			
Monolingual Arabic	1,577	200	1,840
Monolingual English	584	1,808	2,424
Code-switched Ar-En	549	344	999
English: 1-20%	274	80	359
English: 21-40%	102	54	180
English: 41-60%	57	40	121
English: 61-80%	38	49	114
English: 81-99%	78	121	225
Annotations only	504	242	770
Total Utterances	3,214	2,594	6,033

Table 2: Corpus statistics showing the counts of token and utterance types across recordings with Arabic-speaking (Rec-Ar) and English-speaking (Rec-En) Interlocutors. The reported total is the summation of both in addition to the pilot recording.

Metric	Average	SD
Code-Mixing Index (CMI)	0.20	0.13
Switch Point Fraction (SPF)	0.20	0.14
Percentage of English words	44.0%	32.6%

Table 3: Arabic-English code-switching statistics, reporting CMI, SPF and percentage of English words, calculated as the mean and standard deviation over code-switched utterances.

6. Code-switching Analysis

In this section, we present analyses for Arabic-English and MSA-dialectal Arabic code-switching. Statistics on the former code-switching type is presented in Table 3. Table 4 demonstrates examples, covering different dialectness level annotations and showing both types of code-switching.

6.1. Arabic-English Code-switching

As reported in the utterance-level analysis presented in Table 2, Arabic-English code-switched utterances constitute 19.0% of all non-annotation-only utterances. In order to measure the amount of Arabic-English code-switching, we use three metrics: Code-mixing Index (CMI) (Gambäck and Das, 2016), Switch Point Fraction (SPF) (Pratapa et al., 2018), and the percentage of English words. Afterwards, we analyze the morphological code-switching constructs.

Code-Mixing Index CMI measures the level of mixing between languages, and is defined on the utterance-level as follows:

$$C(x) = \frac{\frac{1}{2} * (N(x) - max_{L_i \in \mathbf{L}}\{t_{L_i}\}(x)) + \frac{1}{2}P(x)}{N(x)}$$

where N is the number of language-dependent tokens in utterance x; $L_i \in \mathbf{L}$ the set of languages in the corpus; $max\{t_{L_i}\}$ represents the number of tokens in the dominating language in x, with $1 \leq max\{t_{L_i}\} \leq N$; and P is the number of code alternation points in x; $0 \leq P < N$. The corpus-level CMI is calculated as the average of utterance-level CMI values. Among the code-switched utterances, the value is 0.20 with standard deviation of 0.13.

Switch Point Fraction SPF is calculated as the number of switch points over the number of word boundaries in an utterance. The corpus-level SPF is calculated as the average SPF values over utterances. Among the code-switched utterances, the SPF value is 0.20 with standard deviation of 0.14.

Percentage of English Given that the CMI metric does not distinguish between the primary and secondary languages, we report the percentage of English words to get a better understanding on the amount of English usage. Among the code-switched utterances, the average percentage of English words over utterances is 44.0% with standard deviation of 32.6%.

Morphological Code-switching Among the 353 MCS constructs in the corpus, we report the attachment of the Arabic definite article ال *Al* 'the' as the dominating construct, where 78.5% of the MCS constructs are *Al*+English word. This is in-line with the figures reported for Egyptian Arabic-English code-switching in Hamed et al. (2021), where this construct had a share of around 74%. Other MCS constructs include the attachement of Arabic conjunction proclitics (7.9%), prepositional proclitics (2.0%), and feminine plural suffix (1.4%). The remaining 10.2% include the use of the definite article

Level	Example
L0	كيف يمكن أن نشجع عملهم كفريق ولكن في نفس الوقت نحافظ على التخصص؟ How can we encourage their work as a team, but at the same time maintain specialization?
L1	%أم أي مخاطر أخرى من وجهة نظركم ت-- **حددتوها** ؟ %mm any other risks from your point of view **y– you specified them**?
L2	شكرا، شكرا لك على الفرصة الحلوة **هذي** . Thanks, thank you for **this** nice opportunity.
L3	**طيب** أنا **باتكلم** عن مادة آخذها أنا في هذا ال+*semester* حاليا. **Ok, I am talking** about a course that I am in this *semester* currently.
L4	**ونقدر نفتح** حسابات في ال+*social media* مثلا. **صح ما بنبيع حق العالم بس** يعني نعرض منتجاتنا. **We can open** accounts on *social media*, for example. **It's true we will not sell to** the world, **but** I mean, we offer our products.

Table 4: Examples of utterances receiving *L0-L4* dialectness level annotations. Dialectal words are bolded.

in combination with the other mentioned proclitics, plural suffix, as well as cliticized demonstrative pronoun. While MCS can also involve verbal inflection, we do not observe such constructs in our corpus.

6.2. MSA-Dialectal Code-switching

With regards to the dialectness level annotation task outlined in Section 4.2, a total of 1,158 utterances were annotated. We report that 40.7% of the utterances are annotated as *L0*, 17.4% as *L1*, 11.5% as *L2*, 14.2% as *L3*, and 16.2% as *L4*.

7. Morphological Annotation

We automatically annotate the corpus for tokenization, POS tags, and lemmas following Habash and Palfreyman (2022). We discuss the annotation process as well as analyses and observations.

7.1. Annotation Guidelines

For **English words**, we use Stanza (Qi et al., 2020) to obtain lemmas and Universal Dependency (UD) POS tags (Nivre et al., 2017). English tokenization is minimally intrusive; one exception is contractions, which Stanza tokenizes, e.g., *I'll* is separated to *I+'ll*. For **Arabic words**, we first morphologically disambiguate using CAMeL Tools BERT-based model (Obeid et al., 2020; Inoue et al., 2021), which produces lemmas and a wide range of features. We utilize the MSA model for Arabic-speaking Interlocutors, as that is their intended language. For Students' utterances, we use the Gulf Arabic model, and back off to the MSA model in case of missing analyses. For tokenization, we follow the ATB tokenization, where all clitics are tokenized except for the definite article (Maamouri et al., 2004; Habash, 2010). For POS tags, we use UD (Taji et al., 2017). In the case of MSA, these features are readily provided by CAMeL Tools. But for Gulf Arabic, we convert the Buckwalter tag features to get tokenization and UD POS tags using the mapping provided by Taji et al. (2017). Finally, we use the diacritized Arabic lemmas produced by CAMeL Tools.

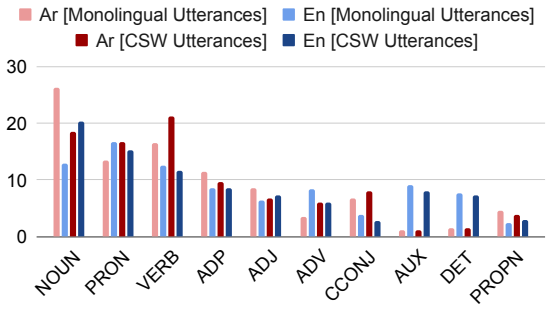

Figure 2: POS distribution for Arabic and English words in monolingual and code-switched (CSW) utterances.

As part of the pre-processing step for this annotation task, the transcriptions were automatically white-space-and-punctuation tokenized, except for the following annotation tokens, which were left untokenized: *GR-background*, *CR-nonSpeech*, *CR-interjections*, *CR-partial*, *CR-flow*, *CSR-MCS*, and *ENR-acronyms*.

7.2. Statistics and Observations

Part-of-Speech In Figure 2, we present the distribution of top-occurring POS tags for Arabic and English words across monolingual and code-switched utterances.[6] In the context of monolingual utterances, we report that Arabic has a higher usage of NOUN and CCONJ over English, while DET and AUX are more common in English than Arabic. These observations were previously reported and

[6]We report the distribution for Arabic and English words only, excluding partial words, punctuation, annotations, and morphologically code-switched words.

Figure 3: Lemma cloud for the top-occurring 100 noun, verb, adjective, or adverb lemmas.

justified in Habash and Palfreyman (2022). However, unlike them, we report Arabic having a higher frequency of VERBs over English. In the context of code-switching, we see a significant increase in frequency for English NOUNs over their occurrence in monolingual utterances, showing +56.9% relative increase. The prevalence of NOUNs in embedded code-switched words is expected as NOUNs are widely-used in borrowing. However, we note that in our corpus the percentage of NOUNs in code-switched utterances (20%) is lower than that previously reported in Bali et al. (2014) and Hamed et al. (2018) (57-66%). This could reflect that the code-switching in the corpus involves more complex intra-sentential code-switching than borrowing of NOUNs. It can also be due to a significant amount of code-switched utterances being dominated by English, thus the POS distribution of the English words being closer to that of monolingual English utterances. This is shown in Table 2, where we report that code-switched utterances having 1-20% English words only constitute 35.9% of the code-switched utterances, while the rest are well distributed across the other ranges, with 33.9% of the code-switched utterances being dominated (61-99%) by English.

Tokenization Under the utilized tokenization schemes, we find that on average, Arabic and English words have 1.23 and 1.03 morphemes, respectively. For English, 96.6% of the words do not get tokenized, 3.4% have 2 morphemes, and 0.03% have 3 morphemes. The figures for Arabic words are higher, where the percentages of words consisting of 1, 2, 3, and 4 morphemes are 78.7%, 19.8%, 1.5%, and 0.02%. It is to be noted that the figures for Arabic do not fully reflect its morphological richness, as affixes are not tokenized.

Lemmatization In Figure 3, we show the lemma cloud for the 100 top-occurring noun, verb, adjective, or adverb lemmas in the corpus. The top-occurring lemmas constitute words that are normally frequently present in speech in addition to words that are specific to the discussed topics. The former case includes lemmas such as عنى *ʿný* 'mean', مثل *mθl* 'like', هيه *hyh* 'yes', كان *kAn* 'be', *so*, *have*, *think*, and *know*. The latter includes lemmas such as شركة *šrkħ* 'company', محل *mHl* 'shop', ناس *nAs* 'people', عمل *ʿml* 'work', *student*, *study*, *university*, and *business*.

8. Conclusion and Future Work

In this work, we extend the currently available speech corpora with a bilingual, multidialectal corpus, containing (accented) English and Modern Standard Arabic, as well as Gulf and Egyptian Arabic dialects. The recordings are collected through Zoom meetings and are manually transcribed. We develop our transcription guidelines to handle challenges introduced by conversational speech, code-switching, and unstandardized orthography. We provide an analysis on the code-switching involved in the corpus. We also automatically annotate the corpus for POS tags, tokenization and lemmatization, and plan on extending the corpus with manual revisions. Also, for the subset of the corpus containing code-switching between Arabic variants, we provide dialectness level annotations.

Looking ahead, we plan to manually annotate the corpus for morphological features and syntactic representations. We also plan to use it to benchmark a range of NLP tasks from ASR to morphological disambiguation and syntactic parsing.

Ethics Statement

We conducted this research with full approval from Zayed University Institutional Review Board. All annotators and participants were compensated fairly according to campus payment policies. Participants were informed that the meetings are being recorded, and that the purpose of the study is collection of natural speech. Participants gave explicit consent for the public release of audio files and transcriptions. We've taken rigorous steps to anonymize any sensitive or identifying information.

Acknowledgments

ZAEBUC was funded by a Zayed University Research Incentive Fund (RIF #R19068). We thank Dr. Michael Bowles for his project management help.

9. Bibliographical References

Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and San'ani Yemeni Arabic. In *Proceedings of LREC*, pages 1300–1306.

Ahmed Ali, Peter Bell, James Glass, Yacine Messaoui, Hamdy Mubarak, Steve Renals, and Yifan Zhang. 2016. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. In *SLT*, pages 279–284.

Ahmed Ali, Suwon Shon, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, James Glass, Steve Renals, and Khalid Choukri. 2019. The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech. In *Proceedings of ASRU*, pages 1026–1033.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, pages 316–322.

Khalid Almeman, Mark Lee, and Ali Abdulrahman Almiman. 2013. Multi dialect Arabic speech parallel corpora. In *Proceedings of the International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2018. The French-Algerian code-switching triggered audio corpus (FACST). In *Proceedings of LREC*, pages 1468–1473.

Mohamed Balabel, Injy Hamed, Slim Abdennadher, Ngoc Thang Vu, and Özlem Çetinoğlu. 2020. Cairo student code-switch (CSCS) corpus: An annotated Egyptian Arabic-English corpus. In *Proceedings of LREC*, pages 3973–3977.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. I am borrowing ya mixing?" an analysis of English-Hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching*, pages 116–126.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic Dialect Corpus and Lexicon. In *Proceedings of LREC*, pages 3387–3396.

Hennie Brugman, Albert Russel, and Xd Nijmegen. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of LREC*, pages 2065–2068.

Shammur A Chowdhury, Younes Samih, Mohamed Eldesouki, and Ahmed Ali. 2020. Effects of dialectal code-switching on speech modules: A study using Egyptian Arabic broadcast speech. In *Proceedings of Interspeech*, pages 2382–2386.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching Arabic ASR. In *Proceedings of Interspeech*, pages 2466–2470.

A Seza Doğruöz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of ACL-IJCNLP*, pages 1654–1666.

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a TV broadcasts speech recognition system for Qatari Arabic. In *Proceedings of LREC*, pages 3057–3061.

Charles A Ferguson. 1959. Diglossia. *word*, 15(2):325–340.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of LREC*, pages 1850–1855.

Marwa Gaser, Manuel Mager, Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. Exploring segmentation approaches for neural machine translation of code-switched Egyptian Arabic-English text. In *Proceedings of EACL*, pages 86–100.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh, and Hind Saddiki. 2018. Unified guidelines and resources for Arabic dialect orthography. In *Proceedings of LREC*, pages 3628–3637.

Nizar Habash and David Palfreyman. 2022. ZAEBUC: An annotated Arabic-English bilingual writer corpus. In *Proceedings of LREC*, pages 79–88.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of Arabic dialectness. In *Proceedings of*

the LREC Workshop on HLT & NLP within the Arabic world, pages 49–53.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*, pages 15–22. Springer, Netherlands.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Injy Hamed, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu. 2021. Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech. *Computer Speech & Language*, 72:101278.

Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2018. Collection and analysis of code-switch Egyptian Arabic-English speech corpus. In *Proceedings of LREC*, pages 3805–3809.

Injy Hamed, Nizar Habash, Slim Abdennadher, and Ngoc Thang Vu. 2022. ArzEn-ST: A three-way speech translation corpus for code-switched Egyptian Arabic-English. In *Proceedings of the Arabic Natural Language Processing Workshop (WANLP)*, pages 119–130.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. ArzEn: A speech corpus for code-switched Egyptian Arabic-English. In *Proceedings of LREC*, pages 4237–4246.

Go Inoue, Salam Khalifa, and Nizar Habash. 2021. Morphosyntactic tagging with pre-trained language models for Arabic and its dialects. In *Findings of ACL*, pages 1708—-1719.

Manal A Ismail. 2015. The sociolinguistic dimensions of code-switching between Arabic and English by Saudis. *International Journal of English Linguistics*, 5(5):99.

Mustafa Jarrar, Nizar Habash, Faeq Alrimawi, Diyam Akra, and Nasser Zalmout. 2016. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation*, pages 1–31.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of EMNLP*.

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A Large Scale Corpus of Gulf Arabic. In *Proceedings of LREC*, pages 4282–4289.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of Emirati Arabic. In *Proceedings of LREC*, pages 3839–3846.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Mona Diab, Nizar Habash, Owen Rambow, and Dalila Tabessi. 2006. Developing and Using a Pilot Dialectal Arabic Treebank. In *Proceedings of LREC*, pages 443–448.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, pages 102–109.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of LREC*, pages 2348–2354.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. QASR: QCRI Aljazeera speech resource–a large scale annotated Arabic speech corpus. In *Proceedings of ACL*, pages 2274–2285.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Eckhard Bick, Cristina Bosco, Gosse Bouma, Sam Bowman, Marie Candito, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Fabricio Chalub, Jinho Choi, Çağrı Çöltekin, Miriam Connor, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Nizar Habash, Jan Hajič, Linh Hà Mỹ, Dag Haug, Barbora Hladká, Petter Hohle, Radu Ion, Elena Irimia, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Hiroshi Kanayama, Jenna Kanerva, Natalia Kotsyba, Simon Krek, Veronika Laippala, Phuong Lê Hồng, Alessandro Lenci, Nikola Ljubešić, Olga Lyashevskaya, Teresa Lynn, Aibek Makazhanov, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, André Martins, Jan

Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Shunsuke Mori, Bohdan Moskalevskyi, Kadri Muischnek, Nina Mustafina, Kaili Müürisep, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Hanna Nurmi, Stina Ojala, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Lauma Pretkalniņa, Prokopis Prokopidis, Tiina Puolakainen, Sampo Pyysalo, Alexandre Rademaker, Loganathan Ramasamy, Livy Real, Laura Rituma, Rudolf Rosa, Shadi Saleh, Manuela Sanguinetti, Baiba Saulīte, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Lena Shakurova, Mo Shen, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Takaaki Tanaka, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Larraitz Uria, Gertjan van Noord, Viktor Varga, Veronika Vincze, Jonathan North Washington, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, and Hanzhi Zhu. 2017. Universal dependencies 2.0.

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 LREC*, pages 7022–7032.

David Palfreyman and Nizar Y Habash. 2022. *Bilingual writers and corpus analysis*. Taylor & Francis Group.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý. 2011. The Kaldi speech recognition toolkit. In *Proceedings of ASRU*, pages 1–4.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of ACL*, pages 1543–1553.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of ACL: System Demonstration*, pages 101–108.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 823–848.

Colleen Richey, Cynthia D'Angelo, Nonye Alozie, Harry Bratt, and Elizabeth Shriberg. 2016. The SRI speech-based collaborative learning corpus. In *Proceedings of Interspeech*, pages 1550–1554.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal dependencies for Arabic. In *Proceedings of the Workshop for Arabic Natural Language Processing (WANLP)*, pages 166–176.

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

10. Language Resource References

Pty Ltd, Sydney, and Australia Appen. 2006a. Gulf Arabic conversational telephone speech LDC2006S43. Web Download. Philadelphia: Linguistic Data Consortium.

Pty Ltd, Sydney, and Australia Appen. 2006b. Gulf Arabic conversational telephone speech, transcripts LDC2006T15. Web Download. Philadelphia: Linguistic Data Consortium.

Pty Ltd, Sydney, and Australia Appen. 2007. Levantine Arabic conversational telephone speech, transcripts LDC2007T01. Web Download. Philadelphia: Linguistic Data Consortium.

BBN Technologies, John Makhoul, Bushra Zawaydeh, Frederick Choi, and David Stallard. 2005. BBN/AUB DARPA Babylon Levantine Arabic speech and transcripts LDC2005S08. Web Download. Philadelphia: Linguistic Data Consortium.

Hassan Gadalla, Hanaa Kilany, Howaida Arram, Ashraf Yacoub, Alaa El-Habashi, Amr Shalaby, Krisjanis Karins, Everett Rowson, Robert MacIntyre, Paul Kingsbury, David Graff, and Cynthia McLemore. 1997. CALLHOME Egyptian Arabic transcripts. *Linguistic Data Consortium, Philadelphia*.

H. Kilany, H. Gadalla, H. Arram, A. Yacoub, A. El-Habashi, and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22.

Mohamed Maamouri, Ann Bies, Seth Kulick, Dalila Tabessi, and Sondos Krouna. 2012. Egyptian Arabic Treebank DF Parts 1-8 V2.0 - LDC catalog numbers LDC2012E93, LDC2012E98, LDC2012E89, LDC2012E99, LDC2012E107, LDC2012E125, LDC2013E12, LDC2013E21.

Mohamed Maamouri, Tim Buckwalter, Dave Graff, and Hubert Jin. 2007. Fisher Levantine Arabic conversational telephone speech LDC2007S02. Web Download. Philadelphia: Linguistic Data Consortium.

Paul Boersma and David Weenink. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.14, retrieved from https://www.praat.org.

Colleen Richey, Nonye Alozie Cynthia D'Angelo, Harry Bratt, and Elizabeth Shriberg. 2019. SRI speech-based collaborative learning corpus LDC2019S01. Web Download. Philadelphia: Linguistic Data Consortium.