Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4237–4246
Marseille, 11–16 May 2020

ArzEn: A Speech Corpus for Code-switched Egyptian Arabic-English

Injy Hamed[1,2], **Ngoc Thang Vu**[2], **Slim Abdennadher**[1]
[1]Computer Science Department, The German University in Cairo, Cairo, Egypt
[2]Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany,
{hamediy,thang.vu}@ims.uni-stuttgart.de,
slim.abdennadher@guc.edu.eg

Abstract

In this paper, we present our ArzEn corpus, an Egyptian Arabic-English code-switching (CS) spontaneous speech corpus. The corpus is collected through informal interviews with 38 Egyptian bilingual university students and employees held in a soundproof room. A total of 12 hours are recorded, transcribed, validated and sentence segmented. The corpus is mainly designed to be used in Automatic Speech Recognition (ASR) systems, however, it also provides a useful resource for analyzing the CS phenomenon from linguistic, sociological, and psychological perspectives. In this paper, we first discuss the CS phenomenon in Egypt and the factors that gave rise to the current language. We then provide a detailed description on how the corpus was collected, giving an overview on the participants involved. We also present statistics on the CS involved in the corpus, as well as a summary to the effort exerted in the corpus development, in terms of number of hours required for transcription, validation, segmentation and speaker annotation. Finally, we discuss some factors contributing to the complexity of the corpus, as well as Arabic-English CS behaviour that could pose potential challenges to ASR systems.

Keywords: Arabic-English, Dialectal Egyptian Arabic, code-switching, speech corpus, spontaneous speech

1. Introduction

The language in Egypt is rather complex and poses many challenges to Natural Language Processing (NLP) tasks. First of all, it is considered as a classic example of "Diglossia", that is, a situation in which one language is used in formal or written realms and a second language in informal or spoken realms (Ferguson, 1959). While Modern Standard Arabic (MSA) is taught in schools and used by people in formal contexts, the Egyptian Dialectal language is the lingua franca used by Egyptians in their daily lives. Beyond this diglossia of Standard and Dialectal Arabic, Egyptians tend to embed foreign languages into their conversations. It is mostly common to embed English, followed by French as a second language. The act of mixing more than one language in a conversation is referred to by linguists as "Code-switching" (CS). Given that MSA and Dialectal Egyptian are considered to be two different languages (Ferguson, 1959), code-switching in Egypt is considered to be rather challenging with the mixing of MSA and Dialectal Arabic as well as Arabic and English. In our work, we are more interested in the latter language pair.

According to Poplack (1980), there are three types of language alternations; extra-sentential CS (where a loan word is borrowed from the secondary language), intra-sentential CS (where the language, or code, switch is done within the same sentence) and inter-sentential CS (where the switch is done across sentences). In the scope of this paper, for the sake of simplicity, we will use the term "code-switching" to refer to all types of alternations.

Several factors, including globalization and immigration, have given rise of CS among many bilingual/multilingual societies. In the middle east, colonization and international businesses and education have played a major role in introducing English and French into everyday conversations. CS is prevalent in Arab countries such as Arabic-French in Morocco (Bentahila, 1983) and Algeria (Bentahila and Davies, 1983), Arabic-English in Egypt (Abu-Melhim, 1991), Saudi Arabia (Omar and Ilyas, 2018), Jordan (Mustafa and AL-KHATIB, 1994), Kuwait (Akbar, 2007), Oman (Al-Qaysi, 2016) and UAE (Khuwaileh, 2003) and a high level of multilingualism is found in Lebanon (Bacha and Bahous, 2011) and Tunisia (Baoueb, 2009) with the mixing of Arabic and both English and French.

Given that CS is language-dependent, a corpus for each language pair is needed. However, collecting CS corpora is a very challenging task, thus the collected, and available, corpora are very scarce and cover few language pairs. This lack of data is one of the main problems hindering the development of NLP applications that handle CS data (Çetinoğlu et al., 2016). Despite Arabic being one of the most widely used languages, there is a huge gap in the available CS speech corpora. For the Egyptian dialect, the only corpus available is the corpus gathered by Hamed et al. (2018). However, this corpus suffers from two main drawbacks: (1) it only covers 4 hours of transcribed speech and (2) the transcriptions are not segmented into sentences. In this paper, we extend our efforts in developing another Egyptian Arabic-English speech corpus.

We present ArzEn, a mixed Egyptian Arabic-English speech corpus consisting of a new set of 12 hours of speech. The corpus is obtained through informal interviews, where participants discuss broad topics, including education, hobbies, work, and life experiences. The recordings are segmented into sentences and transcribed. We also gather participants' meta-data, including their gender, age, education, occupation, perceptions about CS as well as their personality traits, gathered through the Big-5 Personality Test. Thus, this corpus serves as a useful resource in multiple fields, including NLP applications (mainly designed for ASR systems), linguistic analysis of the CS phenomenon, as well as sociolinguistic and psycholinguistic analyses.

2. Language Development in Egypt

Throughout history, Egypt has been occupied and conquered by many foreigners, including Greeks, Persians, French, Ottomans and British. The language has accordingly changed and evolved across the years. The contemporary language involves the use of the following three languages: Standard Arabic, Egyptian Arabic and English. In this Section, we give an overview on factors that have affected the language in Egypt. Several researchers provide surveys on the impact of history, politics and economy on the language in Egypt, with interesting discussions and reflections on language ideologies (Warschauer et al., 2002; Simpson and others, 2008; Stadlbauer, 2010; Bassiouney, 2015; Abouelhassan and Meyer, 2016) as well as the history of English introduction in the Egyptian education system (Schaub, 2000; Cochran, 2013).

The first European language was introduced in Egypt by the French conquerors (1798-1801), however with no major impact on the Arabic language. Under the rule of Muhammad Ali during the Ottoman occupation (1805-1853), other foreign languages were introduced, including Turkish, Persian and Italian. Throughout the British rule (1882-1922), while French was perceived to be the "elite" language dominating the private schools, the English language started gaining more popularity.

According to Warschauer et al. (2002), the British administration in Egypt, aiming to weaken the influence of Arabic, adopted two courses of action: (1) introduced English and French as required languages in the education system and (2) elevated the status of dialectal Arabic. It is claimed that dialectal Arabic was used as a tool to heighten the distinctiveness and distance of Egypt from the rest of the Arab world. Thus, MSA was facing a threat posed by both, the foreign and dialectal languages. On the other hand, Abouelhassan and Meyer (2016) believe that the spread of English was a result of Egyptians' willingness to use the language for their own benefit, and that there were no explicit educational policies aiming at spreading the English language. The demand on learning English increased as a by-product of economic needs, modernization pressures, and people seeking better jobs.

After Egypt gained its independence (1953), two strong forces affected language in opposite directions. On one hand, English was still a symbol for modernization. On the other hand, the change in language was not welcomed by many pan-Arab nationalists and religious conservatives, who longed for empowering the use of Arabic. With Gamal Abdel Nasser being assigned the president of Egypt (1956-1970) and with the rise of pan-Arabism, the use of MSA was revitalized as a common language and a unifying force to all Arabic speaking countries (Stadlbauer, 2010). As quoted from one of Nasser's speeches, he declared that "We announce that we believe in a single Arab nation. The Arab nation was always united linguistically. And linguistic unity is unity of thought". Although Egypt kept its own dialect in everyday interactions, MSA was introduced as an obligatory language in all schools.

According to Abouelhassan and Meyer (2016), the real invasion of English in the education system did not happen under the British rule, but rather in the years of Anwar El Sadat (1970-1981) and Hosni Mubarak (1981-2011) when the middle class grew in size and wealth. This gave a rising demand on private schools where English was the primary instruction language, as more people afforded better education and sought higher language proficiency.

Nowadays, English is taught in public and private education sectors. In the public sector, English is a compulsory subject at preparatory level (grades 7-9) (Schaub, 2000). It is the main instruction language in some universities, including medicine, dentistry and engineering, as well as in the special English-medium sections in other universities, such as commerce and law (Warschauer et al., 2002). In the private sector, English is the main instruction language in schools and universities, where English language instruction begins in kindergarten. There are also private French schools, however, much less in number.

The attitude of Egyptians towards the use of English has shifted from post-colonization state, where English shop signs were criticized as "Arabic in the Valley of Neglect", "Winds of Foreignization Sweep the Egyptian Street" and "Before Arabenglish Spreads" (Simpson and others, 2008), to being accepted by the majority as a symbol of modernization. As (Imhoof, 1977) states, people's perception towards the use of English has shifted from a "necessary evil" during the British occupation to a "practical vehicle for educational, economic and social mobility".

Not only do Egyptians use the three languages, but they also alternate between them. Code-switching can be seen mainly between the two language pairs: MSA + Egyptian Arabic and Egyptian Arabic + English. The former language pair has been examined more thoroughly by linguists (Eid, 1988) and computational linguists (Elfardy and Diab, 2012; AlGhamdi et al., 2019). According to Bassiouney (2006), MSA and Egyptian Arabic are commonly mixed in relatively formal contexts, such as political speeches, sermons in mosques and university lectures. It has also become common practice for well-educated young Egyptians to embed English in everyday conversations (Schaub, 2000). However, this language pair has been less studied by researchers. In Hamed et al. (2018), high usage of CS was found among university teaching assistants during informal interviews; where 37.4% of the words were English and 79.8% of the sentences were code-mixed.

In summary, the contemporary language is a product of the past and the present, where each language stands for a certain ideology. The standard Arabic identifies Egyptians as part of the Arab world, and the Islamic nation. The Egyptian Arabic, used in everyday communication, jokes, songs, and cinema is central to their identity as Egyptians (Warschauer et al., 2002). The English language represents the link to the modern world and a tool for better job opportunities.

3. Related Work

According to Ethnologue (Eberhard et al., 2019), the Arabic language has 273.9 million total number of speakers (ranking $6th$ among all languages). Despite Arabic being one of the most widely used languages, there is huge scarcity in the available resources. Most of the collected speech corpora are for MSA, mainly covering news data, politics and economics, such as GALE data (Walker, 2017; Walker, 2018), GlobalPhone (Schultz, 2002), United Nations Proceedings Speech (Chay et al., 2014), NetDC Arabic BNSC (Choukri et al., 2004), and Arabic Broadcast News Speech and Transcripts (Maamouri et al., 2001; Maamouri et al., 2006). Fewer resources are available for dialectal Arabic, such as CALLHOME Egyptian Arabic corpus (Gadalla et al., 1997) and JANA: A Human-Human Dialogues Corpus for Egyptian Dialect (Elmadany et al., 2016). Very few speech corpora have been collected for mixed Arabic. Up to our knowledge, Arabic CS speech corpora have only been collected for the following dialects:

- Egyptian Arabic-English (Hamed et al., 2018): 6 hours of informal interviews with 12 speakers in technical domain. Transcriptions were obtained for 4 hours.
- Saudi Arabic-English (Ismail, 2015): 89 minutes gathered from informal dinner gatherings involving 6 participants, where the speech is transcribed.
- Algerian Arabic-French (Amazouz et al., 2018): 7.5 hours of read speech from books and movie transcripts, as well as informal conversations, gathered from 20 speakers. The corpus contains transcriptions, sentence segmentation, language boundary and phone-level time codes information.
- Maghrebian Arabic-French (MOHDEB-AMAZOUZ et al., 2016): 53 hours of spontaneous speech gathered from TV entertainment and talk shows, involving speakers from Algeria, Morocco and Tunisia. The corpus contains sentence segmentation and language annotation.

The vast majority of the available CS speech corpora have covered Chinese-English (Chan et al., 2009; Shen et al., 2011; Li et al., 2012; Lyu et al., 2015; Ahmed and Tan, 2012), Hindi-English (Dey and Fung, 2014; Ramanarayanan and Suendermann-Oeft, 2017; Sivasankaran et al., 2018; Sreeram et al., 2018; Pandey et al., 2017) and Spanish-English (Solorio and Liu, 2008; Deuchar et al., 2014; Ramanarayanan and Suendermann-Oeft, 2017) language pairs. Less work has covered Arabic-English (Hamed et al., 2018; Ismail, 2015), Arabic-French (Amazouz et al., 2018; MOHDEB-AMAZOUZ et al., 2016), Frisian-Dutch (Yilmaz et al., 2016), Mandarin-Taiwanese (Lyu et al., 2006; Lyu and Lyu, 2008), Turkish-German (Çetinoğlu, 2017), English-Malay (Ahmed and Tan, 2012), English-isiZulu (van der Westhuizen and Niesler, 2016) and Sepedi-English (Modipa et al., 2013).

Although CS has received more attention from the linguistic and NLP communities and researchers have made significant progress in the available CS speech corpora, the available corpora are yet limited; as they are mostly relatively-small and covering few language pairs. There is still a huge need to further collect corpora for the other language pairs, as well as extend the corpora for the previously mentioned languages, in order to give foundation for multilingual NLP applications to spur in that direction.

4. The ArzEn Corpus

In this section we present the ArzEn[1] corpus. We discuss how it is collected, provide an overview on participants' profiles, as well as statistics on the CS behaviour in the corpus. We also summarize the efforts needed in terms of working hours for collecting the corpus. Table 1 gives a summary on the main properties of the corpus.

Language	Egyptian Arabic (Arz) & English (En) (+ few French words)
Size	12 hours (38 interviews)
Speakers	40 speakers (2 interviewers + 38 interviewees)
Type	Spontaneous Speech
Domain	Informal interviews discussing general topics such as education, career, work and traveling experiences.
Environment	Soundproof room
Recordings Quality	48kHz sampling rate

Table 1: Summary on ArzEn corpus

4.1. Data Collection

The interviews were held at The German University in Cairo (GUC), a private university where English is the instruction language. To ensure good audio quality, all recordings were carried out in a soundproof room. The data was recorded with mono channel at 48kHz sampling rate. Participants included students and employees from the GUC. A total of 38 interviews were recorded, with an average duration of 19.1 minutes each.

Interviews included one interviewee and two interviewers. The interviewees were asked questions covering several topics, including education, personality, personal life, career, hobbies, travelling, work and life experiences, role model and technology. In order to allow for the corpus to be used for further linguistic investigations, some considerations were taken in the setup. The interviewers were composed of a male and a female. Also, the set of questions as well as the way the questions were asked were fixed throughout the interviews, so as to avoid having different effects on the CS behaviour across participants' answers.

After the interview, the participant was asked to fill in three forms:

[1] Arz and En are the codes for Egyptian Arabic and English in Ethnologue.

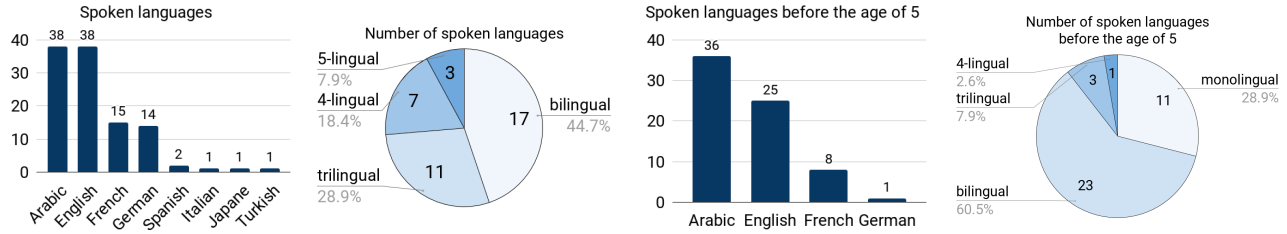

Figure 1: Participants' languages knowledge.

- Questionnaire: gathering information about the participants including their demographics, education, work, travelling experiences, and their perceptions and opinions regarding code-switching.

- Big Five Personality Test (Goldberg, 1992): which assesses five major dimensions of personality: Openness, Conscientiousness, Agreeableness, Extraversion, and Neuroticism. The Big Five Personality test was chosen as it is the most widely used and extensively researched model of personality Gosling et al. (2003) and because it consists of 50 questions, thus only requires around 10 minutes from the participants.

- Self-Assessment Manikin Test (Bynion and Feldner, 2017): which is a picture-oriented questionnaire developed to measure emotional response.

4.2. Overview on Participants

All participants are Egyptians in the age range of 18-35. 61.5% of the participants are males, while 38.5% are females. All participants are students or employees of the GUC. Out of the 38 participants, 21 are students, 15 are teaching assistants and 1 is a GUC employee. The majority of the participants graduated from private English schools (63.2%), followed by private French schools (21.1%) and public national schools (15.8%), where they are all fluent in English.

All participants were asked to report the languages they currently speak, as well as the languages learnt before the age of 5. Figure 1 shows the level of participants' multilingualism. While all participants are Arabic-English bilinguals, participants also show high multilingualism, where more than half (55%) of the participants can speak more than two languages. The reported languages are Arabic (38), English (38), French (15), German (14), Spanish (2), Turkish (1), Italian (1), and Japanese (1). Also, 71% of the participants reported that they spoke at least two languages before the age of 5. Moreover, 86.8% of the participants reported that their family members speak more than one language, which is expected as the family members of the participants involved in this experiment are the generations who received their education in the post-colonial era, in which English has been a mandatory subject in all schools.

Participants' perceptions on CS On a scale of 1-5, participants were asked to rate their frequency of CS as well as how aware they are of their CS usage (which would reflect how accurate their frequency rating would be). Participants report an average of 3.8 rating for CS frequency (with a least rating of 2, meaning that they all code-switch) and 4.1 of usage awareness. On a survey on their mother tongue, 59% of the participants identified their mother tongue to be "Arabic", 41% as "Code-switched, Arabic-English", and 0% as "Pure English". Not only does this observation show the high usage of CS among the participants, it also reflects the participants' acceptance towards the CS phenomenon and identifying it as their mother tongue. When asked if they think code-switching pollutes their mother tongue, 30% reported "No", while 46% and 24% reported "Yes" and "Maybe", respectively.

Around 73.6% of the participants state that code-switching "says something about who they are". This could be inline with Nerghes (2011), stating that code-switching can be used to reflect a certain socioeconomic identity. While some participants believe that code-switching is done due to weakness in one of the languages, the majority of the participants believe it is done due to strength in both languages more than weakness in one them. When asked if they believe that people code-switch to "show-off", 22% reported "No", while 32% affirmed it, and 46% answered "Maybe".

4.3. Corpus Transcription and Annotation

The interviews were manually transcribed by professional transcribers. In order to address the unstanderdized issue of Dialectal Arabic orthography, we based our transcription guidelines on the conventions developed and used by the Egyptian Arabic Wikipedia community [2]. For ambiguous cases and when several orthographic varieties are permitted, we made decisions to restrict the number of possibilities to usually one variant only, to reduce ambiguity. These guidelines were provided to the transcribers and the transcriptions were revised by at least one of the authors. The corpus was then segmented into segments of maximum 25 seconds each. Each speech segment was annotated with the speaker ID. The following tags were used for non-speech parts: hesitation, humming, cough, laugh, noise, and silence. The main source of the produced noise was due to the moving of the microphone between interviewee and interviewer. Although this type of noise sometimes occurred within participants' sentences, it mainly occurred at the beginning of the interviewer/interviewee's turn, in which the noisy part was segmented as a stand-alone segment with the

[2] https://arz.wikipedia.org/wiki/ويكيبيديا:Introduction_in_English#Rules_of_writing

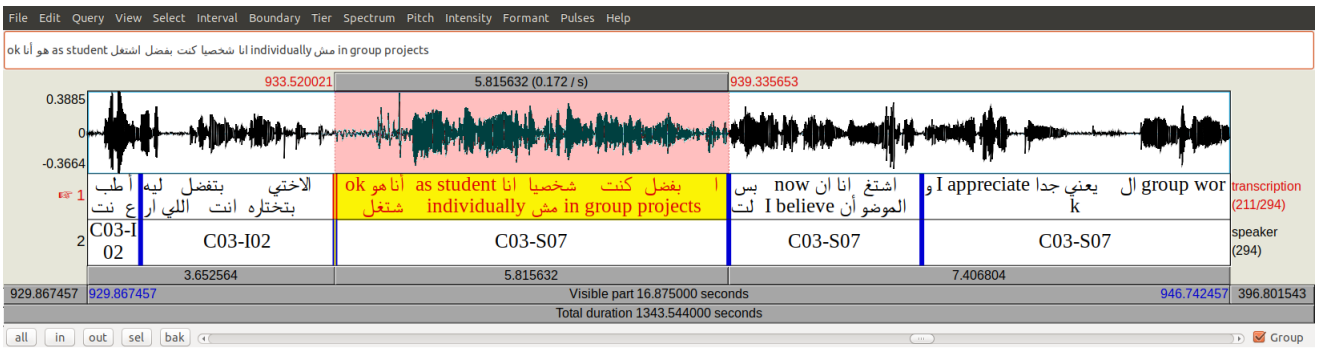

Figure 2: Using praat for the annotation process.

```
intervals [178]:
   xmin = 767.3698816683088
   xmax = 768.6999959016489
   text = "عملت صحاب بس مش في الشغل"←
intervals [179]:
   xmin = 768.6999959016489
   xmax = 772.5488124999999
   text = →"It's very difficult making friends at work"
intervals [180]:
   xmin = 772.5488124999999
   xmax = 778.5911041666667
   text = →"I actually made more friends من ال course بتاع ال Deutsch اللي انا كنت باخده"
```

Figure 3: Example of the TextGrid output from Praat for the transcription tier. The arrows beside the sentences show the sentence starting direction, while the arrows on top of the words show the direction of the language-homogeneous blocks.

noise tag. We adopt the same measures presented in Lyu et al. (2015) to report on the effort needed for the corpus development, and present them in Table 2.

Task	Effort
Collecting Recordings	12 h
Transcriptions	8 RT
Revision and Segmentation	16 RT
Speaker Annotation	0.4 RT

Table 2: Effort on corpus development

Praat[3] was used for annotation. We used two tiers: transcription and speaker. Figure 2 shows an example of the annotation process. Figure 3 shows the TextGrid file produced by Praat.

Reading code-mixed Arabic-English sentences could be confusing as Arabic is written from right to left, thus each language is read in a different direction. In order to read a code-mixed sentence, the reader must first determine the starting direction of the sentence, and then switch direction whenever there is a language switch. Sentences beginning with an Arabic character start from the right side, while sentences beginning with a Latin character start from the left side. In order to make it easier throughout the paper, we will use two notations. A small arrow will be placed at the start of sentences to mark the starting point. Arrows will be placed on top of words to guide you through the word sequences. This annotation is used in Figure 3.

[3] http://www.fon.hum.uva.nl/praat/

For each interview, the following data is gathered: (1) recording, (2) transcriptions, segmentation and speaker annotation, (3) speaker information gathered from the questionnaire, (4) personality traits gathered through the Big Five Personality test and (5) the Self-Assessment Manikin test information.

4.4. Corpus Overview

4.4.1. Overall Statistics

In Table 3, we provide an overview on the corpus. We report the number of tokens, rather than words, as it is common for speakers to use Arabic prefixes and suffixes in combination with English words, as will be discussed in Section 4 4 7 . In this case, we separate the English words from the Arabic prefixes and suffixes using spaces, such as: "ات course ال"←.

4.4.2. CS Analysis

Percentage of code-switching types: The most prevalent CS type in the corpus is code-mixing. A total of 63.2% of the sentences are code-mixed, where 88.9% of the CS sentences are mainly in Arabic with English embeddings, while 7.4% have more English than Arabic words. Intersentential CS is also seen in the corpus, however, with very low frequency. Pure monolingual English sentences only constitute 3.7% of the sentences. However, it is seen that some sentences are mainly in English but contain few Arabic conjugations. For example:

"I would say probably martial arts أو tennis."

(I would say probably martial arts or tennis.)

Category	Value
# Speakers	40
# Interviewers	38
# Interviewees	2
Average Interview Duration	0.32 hours
Total Duration	12 hours
Speech Duration	11.4 hours
Non-speech Duration	0.7 hours
# Sentences	6,290
% Monolingual Arabic Sentences	33.1%
% Monolingual English Sentences	3.7%
% CS Sentences	63.2%
# Tokens	102,332
% Arabic Tokens	84.9%
% English Tokens	15.1%
# Arabic Tokens	86,851
# Unique Arabic Tokens	7,406
# English Tokens	15,481
# Unique English Tokens	2,594
Sentence Duration Range (s)	0.3-24.9
Av. Sentence Duration (s)	6.6
Sentence Length Range (words)	1-95
Av. Sentence Length (words)	16.3
Average speaking rate (words per minute)	149.1

Table 3: Corpus overview

Percentage of embedded language: Throughout the corpus, there are 15,481 English words, which are 15.1% of the total words. Among the CS sentences, 18.8% of the words are in English.

Switches per sentence: On average, in each CS sentence, there are 1.98 switches from Arabic to English and 1.91 switches from English to Arabic. Among the CS sentences, 80.5% of the sentences start in Arabic and 19.5% start in English.

Code-mixing Index (CMI): We use the CMI introduced by Das and Gambäck (2014). It is defined as:

$$CMI = \frac{\sum_{i=1}^{N}(w_i) - max\{w_i\}}{n - u}$$

where $\sum_{i=1}^{N}(w_i)$ is the total number of words over all languages, $max(w_i)$ is the highest number of words across the languages, n is the total number of words, and u is the total number of language-independant words. Monolingual sentences would have a CMI of 0 and sentences with equal word distributions across languages would have a CMI of n/N, which is 0.5 in the case of bilingual utterances.

We calculate CMI over each utterance and average over all sentences. The CMI over the whole corpus is 0.12, and over the CS sentences only is 0.17.

4.4.3. Word Distributions

Across the CS sentences, we analyze the language-homogeneous blocks. In total, there are 10,788 and 8,632 Arabic and English blocks, respectively. We also analyze the number of words in each block. On average, an Arabic block spans 6.0 words (with a range of 1-62) and an English block spans 1.7 words (with a range of 1-32). The words' distribution in terms of length is shown in Figure 4.

From Figure 4, we can interpret the types of CS available in the code-mixed sentences in the corpus. In the case of extra-sentential CS, people borrow loan words or compound words from the secondary language, where the embedded length is usually between 1 to 2 words. More than 2 words would probably indicate intra-sentential CS, where segments of sentences are used in the secondary language. Using the word distribution analysis, we can see that 66.3% of the English blocks are of length 1, 20.2% of the blocks are of length 2, and 13.5% are longer than 2 words. This can give a rough limit for extra-sentential CS and a lower limit for intra-sentential CS.

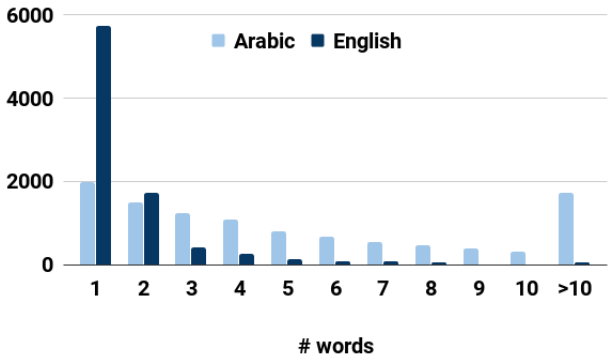

Figure 4: Word distributions showing the number of words in Arabic and English blocks in CS sentences.

4.4.4. Word Frequencies

In Table 4, we show the top frequent unigrams and bigrams. The word frequencies are calculated only over the interviewees' speech, as the interviewers use the same set of questions across interviews, thus the repetition would affect the word frequencies. It can be seen that the most common Arabic-English words are noun+ال← (the+noun), while the most common English-English words are pronoun+verb and compound words.

4.4.5. Trigger Words

A trigger word is defined as the words preceding a code-switching point. There are in total 535 unique Arabic words preceding a code-switching point. Table 5 shows the top frequent trigger words. The trigger words are also calculated only over the interviewees' speech.

It can be seen from Tables 4 and 5 that the most common Arabic-English switches occur after the ال (the) token. This could be because, as seen in Figure 4, it is most common in the corpus for users to embed English segments made up of only one word, which is most commonly the case

Arabic		English	Arabic-English		English-Arabic		English-English
Word	**Translation**	**Word**	**Words**	**Translation**	**Words**	**Translation**	**English**
و	and	I	mobile ال ←	the mobile	→ English و	and English	I think
يعني	meaning/so	a	project ال ←	the project	→ okay أنا	okay I	turning point
ال	the	to	bachelor ال ←	the bachelor	→ okay هو	okay he	computer science
أنا	me	English	I يعني←	so I	→ project بتاعي	my project	role model
بس	but	of	major ال ←	the major	→ masters و	masters and	to be
مش	not	in	I ف ←	so I	→ French و	French and	I would
هو	he	it	mobile غير←	without mobile	→ English بس	English but	at least
كده	that way	the	masters ال ←	the masters	→ routine بتاعي	my routine	dream job
حاجة	something	okay	English ال ←	the English	→ major ده	this major	it was
اللي	that	and	working ال ←	the working	→ national و	national and	I don't

Table 4: Frequent unigrams and bigrams.

Word/Token	Translation	Frequency
ال	the	24.7%
يعني	meaning/so	3.6%
في	in	3.2%
و	and	2.8%
بس	but	2.1%

Table 5: The most frequent trigger words.

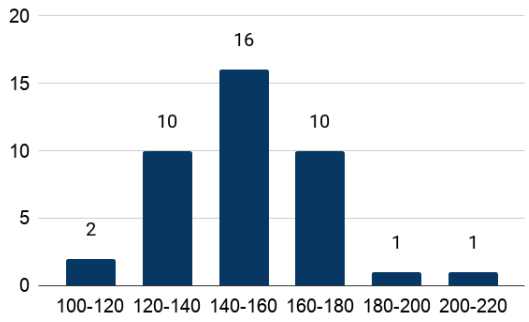

Figure 5: wpm distribution

in extra-sentential code-switching or borrowing. In this case, loan words are used from the secondary language, and these loan words are most commonly nouns, therefore, frequently preceded by the word "the" or "ال". This is also aligned with our observations in the previously collected corpus (Hamed et al., 2018), where the most frequent trigger word was ال (30%), followed by في, و, يعني, and هو.

Even though it is reasonable that most trigger words are conjugations, as they are used to join two parts of sentences, it should be noted that the frequent trigger words are affected by words frequency in general.

4.4.6. Corpus Complexity

The speaking rate affects the overall corpus complexity in terms of transcription difficulty and potential challenge for ASR systems. In order for the speech to be clear for the listener, the words per minute (wpm) rate should be between 140-160 (Li and Vu, 2019). In Figure 5, we show the distributions for the speakers' wpm. It can be seen that 40% of the speakers are within the normal 140-160 range, 30% are below the rate and 30% are above. Therefore, for 30% of the interviews, accurate speech recognition for humans and ASR systems would be more challenging

Another factor adding to the complexity of the corpus is hesitation. Given the spontaneous nature of the gathered corpus, it is common that the utterances contain hesitation. In our corpus, we mark hesitations with "..". A total of 2,732 hesitations are seen in the corpus in more than 26% of the sentences. Three main types of hesitations are observed:

- Repetition:
  "طيب ده .. ده علشان تب .. تبهري اللي قدامك."←
  (Well, that .. that is to im .. impress the present person.)

- Correction:
  "لو هو بيلعب .. لعب في ماتش كورة"←
  (If he plays .. played in a football match)

- Changing course/structure mid-sentence:
  "ساكن .. أقول أنا ساكن فين؟"←
  (I live in .. Do I say where I live?)

- Combination of the three types:
  "امتى .. امتى حددت .. أنت قلت اه في .. بعد أول سنة."←
  (When .. when did you decide .. oh, you already said in.. after freshman year)

Set	Interviews	Females	Males	Duration(h)	wpm	CMI	%CS Sentences	%English Sentences	% English words
Train	12	4+1	8+1	5.6	156.9	0.14	68.8%	3.9%	16.0%
Dev	13	5	8	2.9	143.9	0.14	70.0%	2.6%	14.8%
Test	13	5	8	2.9	148.7	0.14	68.7%	2.1%	17.0%
ArzEn	38	14	24	11.4	151.5	0.14	69.0%	3.17%	16.0%

Table 6: Overview on training, development and testing sets.

4.4.7. CS Complexity

Arabic is a highly morphological language, where words contain prefixes and suffixes. For example, the word "فهنعلمهم" (ف+ه+ن+علم +هم) in Egyptian Arabic means "so we will educate them", where the stem is "علم" meaning "to educate". When code-switching, Egyptians inject the same morphology into English words, where they use the English equivalent of the Arabic stem word (which is the least form of the word without any prefixes and suffixes) and add to it Arabic prefixes and/or suffixes. Examples of such combinations:

- Prefix: "job offer ك" ← (as a job offer), "attitude ب" ← (with attitude), "build+ه"← (will build).
- Suffix: →"course+ات" (courses), →"mobile+ي" (my mobile), →"career+ه" (his career).
- Both: "ها+propagate+ي"← (to propagate it (feminine)).

Moreover, code-mixing is done within English phrases. For example, participants used the phrase "to[1] break the[2] ice" as "ice+ال[2] break+ي[1]"←. This characteristic of the Arabic language that is embedded to Arabic-English CS poses potential challenges to NLP applications.

Another challenge is that the pronunciation of English words may differ in the context of Arabic-English speech. Firstly, because of accents, but secondly, and more interestingly, because, even among unaccented English speakers, the pronunciation of English characters can change when the English word is combined with Arabic prefixes/suffixes. There are similar characters in both languages that are considered to be equivalents but are pronounced slightly different, such as *t* and ت and r and ر. When Arabic prefixes/suffixes get attached to English words, some English phonemes can be replaced by the Arabic close phonemes. For example, the pronunciation of the *t* and *r* differ in both contexts: "the computer" and "computer ال". Also, although the word "target" is pronounced the same as in English in the case of "target+بي"← (he targets), it is pronounced differently in "وا+target+بي"← (they target), where the *r* is pronounced as ر. Also, a speaker might pronounce "to skip it" correctly, however, when saying "ها+skip+ي"←, an emphasis is placed on the *p*. We have only shed light on this problem, however, further investigation is needed to identify when people change the pronunciation of English characters within Arabic-English CS contexts.

4.5. Adaptation, Development, and Test Sets

We have divided the corpus into three sets: train, dev and test. The split has been done taking into consideration having balanced dev and test sets in terms of gender, number of interviews, duration, wpm and CS metrics, as shown in Table 6. Although the corpus is gender-biased, the split is done such that the number of males and females are equal across the dev and test sets. In order to avoid having the interviewers as common speakers across all sets, their utterances have been placed in the train set. Therefore, the train set contains utterances from 4 female and 8 male interviewees, in addition to 1 female and 1 male interviewer.

5. Conclusion

With the widespread of the code-switching phenomenon, a demand has been placed on ASR systems to be able to handle such mixed speech. One of the main challenges hindering the advancements in this direction is the lack of speech corpora. Researchers have collected speech corpora that only cover a few language pairs and there is still a huge gap in the case of Arabic-English. In order to fill the gap, we present our ArzEn corpus. The corpus contains 12 hours of transcribed and segmented recordings gathered from 40 Egyptian Arabic-English bilingual speakers through informal interviews. Information about participants is also collected, including gender, age, educational background, perceptions about CS and personality traits. Thus, the corpus serves as a useful resource for multiple research directions. Firstly, it provides a large enough speech corpus that could be used as an evaluation benchmark for ASR systems. Secondly, it provides linguistic insight into Arabic-English CS. Thirdly, the meta-data collected for the participants can be used in further sociolinguistic and psycho-linguistic analyses. We plan on expanding our corpus with further recordings that would include a wider diversity of participants in terms of age, occupations, and socio-economic backgrounds as well as further annotations including topic domain and language boundary. Also, we intend to delve deeper into the reasons of code-switching and further investigate the factors that affect people's CS behaviour.

6. Acknowledgements

We would like to thank Dahlia Sabet and Nader Rizk for helping us in collecting the corpus. Special thanks also goes to all the participants who volunteered to help us with our project.

7. Bibliographical References

Abouelhassan, R. S. M. and Meyer, L. M. (2016). Economy, modernity, islam, and english in egypt. *World Englishes*, 35(1):147–159.

Abu-Melhim, A.-R. (1991). Code-switching and linguistic accommodation in arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250. John Benjamins Publishing.

Ahmed, B. H. and Tan, T.-P. (2012). Automatic speech recognition of code switching speech using 1-best rescoring. In *2012 International Conference on Asian Language Processing*, pages 137–140. IEEE.

Akbar, R. (2007). *Students' and teachers' attitudes towards Kuwaiti English code-switching.* Cardiff University.

Al-Qaysi, N. J. M. (2016). *Examining Students' and Educators' Attitudes Towards the use of Code-Switching within Higher Educational Environments in Oman.* Ph.D. thesis, The British University in Dubai (BUiD).

AlGhamdi, F., Molina, G., Diab, M., Solorio, T., Hawwari, A., Soto, V., and Hirschberg, J. (2019). Part of speech tagging for code switched data. *arXiv preprint arXiv:1909.13006*.

Amazouz, D., Adda-Decker, M., and Lamel, L. (2018). The french-algerian code-switching triggered audio corpus (facst).

Bacha, N. N. and Bahous, R. (2011). Foreign language education in lebanon: A context of cultural and curricular complexities. *Journal of Language Teaching and Research*, 2(6):1320.

Baoueb, L. B. (2009). Social factors for code-switching in tunisian business companies: A case study.

Bassiouney, R. (2006). *Functions of code switching in Egypt: Evidence from monologues*, volume 46. Brill.

Bassiouney, R. (2015). *Language and identity in modern Egypt*. Edinburgh University Press.

Bentahila, A. and Davies, E. E. (1983). The syntax of arabic-french code-switching. *Lingua*, 59(4):301–330.

Bentahila, A. (1983). Motivations for code-switching among arabic-french bilinguals in morocco. *Language & communication*.

Bynion, T.-M. and Feldner, M. T. (2017). Self-assessment manikin. *Encyclopedia of personality and individual differences*, pages 1–3.

Çetinoğlu, Ö., Schulz, S., and Vu, N. T. (2016). Challenges of computational processing of code-switching. *arXiv preprint arXiv:1610.02213*.

Çetinoğlu, Ö. (2017). A code-switching corpus of turkish-german conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 34–40.

Chan, J. Y., Cao, H., Ching, P., and Lee, T. (2009). Automatic recognition of cantonese-english code-mixing speech. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009*.

Chay, K., Elizalde, C., and Ziemski, M. (2014). United nations proceedings speech (ldc2014s08). *Linguistic Data Consortium (LDC)*.

Choukri, K., Nikkhou, M., and Paulsson, N. (2004). Network of data centres (netdc): Bnsc-an arabic broadcast news speech corpus. In *Linguistic Data Consortium (LDC)*.

Cochran, J. (2013). *Education in Egypt (RLE Egypt)*. Routledge.

Das, A. and Gambäck, B. (2014). Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387.

Deuchar, M., Davies, P., Herring, J., Couto, M. C. P., and Carter, D. (2014). Building bilingual corpora. *Advances in the Study of Bilingualism*, pages 93–111.

Dey, A. and Fung, P. (2014). A hindi-english code-switching corpus. In *LREC*, pages 2410–2413.

Eberhard, D., Simons, G. F., and Fennig, C. D. (2019). Ethnologue: Languages of the world (22nd edn.) dallas: Sil international. *Online at¡ http://www. ethnologue. com¿(Accessed March 22, 2019)*.

Eid, M. (1988). Principles for code-switching between standard and egyptian arabic. *al-'Arabiyya*, pages 51–79.

Elfardy, H. and Diab, M. (2012). Token level identification of linguistic code switching. In *Proceedings of COLING 2012: Posters*, pages 287–296.

Elmadany, A., Abdou, S., and Gheith, M. (2016). Jana: A human-human dialogues corpus for egyptian dialect (ldc2016t24). In *Linguistic Data Consortium (LDC)*.

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

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., Karins, K., Rowson, E., MacIntyre, R., Kingsbury, P., Graff, D., and McLemore, C. (1997). Callhome egyptian arabic transcripts. *Linguistic Data Consortium, Philadelphia*.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Gosling, S. D., Rentfrow, P. J., and Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528.

Hamed, I., Elmahdy, M., and Abdennadher, S. (2018). Collection and analysis of code-switch egyptian arabic-english speech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Imhoof, M. (1977). The english language in egypt. *English around the World*, 17:3.

Ismail, M. A. (2015). The sociolinguistic dimensions of code-switching between arabic and english by saudis. *International Journal of English Linguistics*, 5(5):99.

Khuwaileh, A. A. (2003). Code switching and multilingualism in a small multi-ethnic group society (uae). *JOURNAL OF LANGUAGE FOR INTERNATIONAL BUSINESS*, 14(2):32–49.

Li, C.-Y. and Vu, N. T. (2019). Integrating knowledge in end-to-end automatic speech recognition for mandarin-english code-switching. *International Conference on Asian Language Processing*.

Li, Y., Yu, Y., and Fung, P. (2012). A mandarin-english code-switching corpus. In *LREC*, pages 2515–2519.

Lyu, D.-C. and Lyu, R.-Y. (2008). Language identification on code-switching utterances using multiple cues. In *Ninth Annual Conference of the International Speech Communication Association*.

Lyu, D.-C., Lyu, R.-Y., Chiang, Y.-c., and Hsu, C.-N. (2006). Speech recognition on code-switching among the chinese dialects. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE.

Lyu, D.-C., Tan, T.-P., Chng, E.-S., and Li, H. (2015). Mandarin–english code-switching speech corpus in south-east asia: Seame. *Language Resources and Evaluation*, 49(3):581–600.

Maamouri, M., Graff, D., and Cieri, C. (2001). Arabic broadcast news speech (ldc2006s46). *Linguistic Data Consortium (LDC)*.

Maamouri, M., Graff, D., and Cieri, C. (2006). Arabic broadcast news transcripts (ldc2006t20). *Linguistic Data Consortium (LDC)*.

Modipa, T. I., Davel, M. H., and De Wet, F. (2013). Implications of sepedi/english code switching for asr systems.

MOHDEB-AMAZOUZ, D., Martine, A.-D., and LAMEL, L. (2016). Arabic-french code-switching across maghreb arabic dialects: a quantitative analysis.

Mustafa, Z. and AL-KHATIB, M. (1994). Code-mixing of arabic and english in teaching science. *World Englishes*, 13(2):215–224.

Nerghes, A. (2011). The impact of code-switching on persuasion: An elaboration likelihood perspective. *Wageningen University*.

Omar, A. and Ilyas, M. (2018). The sociolinguistic significance of the attitudes towards code-switching in saudi arabia academia. *International Journal of English Linguistics*, 8(3).

Pandey, A., Srivastava, B. M. L., and Gangashetty, S. V. (2017). Adapting monolingual resources for code-mixed hindi-english speech recognition. In *2017 International Conference on Asian Language Processing (IALP)*, pages 218–221. IEEE.

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Ramanarayanan, V. and Suendermann-Oeft, D. (2017). Jee haan, i'd like both, por favor: Elicitation of a code-switched corpus of hindi-english and spanish-english human-machine dialog. In *INTERSPEECH*, pages 47–51.

Schaub, M. (2000). English in the arab republic of egypt. *World Englishes*, 19(2):225–238.

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh International Conference on Spoken Language Processing*.

Shen, H.-P., Wu, C.-H., Yang, Y.-T., and Hsu, C.-S. (2011). Cecos: A chinese-english code-switching speech database. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 120–123. IEEE.

Simpson, A. et al. (2008). *Language and national identity in Africa*. Oxford University Press.

Sivasankaran, S., Srivastava, B. M. L., Sitaram, S., Bali, K., and Choudhury, M. (2018). Phone merging for code-switched speech recognition.

Solorio, T. and Liu, Y. (2008). Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

Sreeram, G., Dhawan, K., and Sinha, R. (2018). Hindi-english code-switching speech corpus. *arXiv preprint arXiv:1810.00662*.

Stadlbauer, S. (2010). Language ideologies in the arabic diglossia of egypt. *Colorado Research in Linguistics*, 22(1):4.

van der Westhuizen, E. and Niesler, T. (2016). Automatic speech recognition of english-isizulu code-switched speech from south african soap operas. *Procedia Computer Science*, 81:121–127.

Walker, Kevin, e. a. (2017). Gale phase 4 arabic broadcast conversation speech (ldc2017s15). *Linguistic Data Consortium (LDC)*.

Walker, Kevin, e. a. (2018). Gale phase 4 arabic broadcast news speech (ldc2018s05). *Linguistic Data Consortium (LDC)*.

Warschauer, M., Said, G. R. E., and Zohry, A. G. (2002). Language choice online: Globalization and identity in egypt. *Journal of Computer-Mediated Communication*, 7(4):JCMC744.

Yilmaz, E., Andringa, M., Kingma, S., Dijkstra, J., Kuip, F., Velde, H., Kampstra, F., Algra, J., Heuvel, H., and van Leeuwen, D. A. (2016). A longitudinal bilingual frisian-dutch radio broadcast database designed for code-switching research.