Collection and Analysis of Code-switch Egyptian Arabic-English Speech Corpus

Injy Hamed, Mohamed Elmahdy, Slim Abdennadher
The German University in Cairo
El Tagamoa El Khames, New Cairo, Cairo, Egypt
{injy.hamed, mohamed.elmahdy, slim.abdennadher}@guc.edu.eg

Abstract

Speech corpora are key components needed by both: linguists (in language analyses, research and teaching languages) and Natural Language Processing (NLP) researchers (in training and evaluating several NLP tasks such as speech recognition, text-to-speech and speech-to-text synthesis). Despite of the great demand, there is still a huge shortage in available corpora, especially in the case of dialectal languages, and code-switched speech. In this paper, we present our efforts in collecting and analyzing a speech corpus for conversational Egyptian Arabic. As in other multilingual societies, it is common among Egyptians to use a mix of Arabic and English in daily conversations. The act of switching languages, at sentence boundaries or within the same sentence, is referred to as code-switching. The aim of this work is a three-fold: (1) gather conversational Egyptian Arabic spontaneous speech, (2) obtain manual transcriptions and (3) analyze the speech from the code-switching perspective. A subset of the transcriptions were manually annotated for part-of-speech (POS) tags. The POS distribution of the embedded words was analyzed as well as the POS distribution for the trigger words (Arabic words preceding a code-switching point). The speech corpus can be obtained by contacting the authors.

Keywords: Speech corpus, Dialectal Egyptian Arabic, Conversational Egyptian Arabic, Egyptian Arabic-English, code-switching, code-mixing

1. Introduction

The Arabic language is one of the most widely used languages in the world. It is the 6^{th} most used language based on number of of native speakers. Nearly 250 million persons use Arabic as their first language and it is the second language for around four times that number (Elmahdy et al., 2009). There are three types of the Arabic language: Classical Arabic, Modern Standard Arabic (MSA), and Colloquial (dialectal) Arabic. The classical Arabic is the most formal type of Arabic. It is used in the Quran and early Islamic literature. MSA is the official modern language used in the Arab world. It is derived from the Classical Arabic, with some simplifications such as removal of diacritic marks. MSA is the form of Arabic taught in schools and is used in writings, formal speeches, interviews, news broadcasts, movies' subtitling and education. However, MSA is not the language used in everyday life and is considered as a second language for all Arabic-speakers. Each country or region within a country has its own dialect. Colloquial (dialectal) Arabic is the language used in daily conversations and informal writings such as chats, blogs and comments on on-line social media such as Facebook and Twitter. However, they do not necessarily have a standard written form.

In addition to the three forms of Arabic, as in other multilingual environments, many Arabic-speakers use code-switching and code-mixing in their conversations. Such mixed speech is usually defined as a mixture of two distinct languages: primary language (also known as the matrix language); which is spoken in majority and secondary language (also known as the embedded language); by which words (in the case of code-mixing) or phrases (in the case of code-switching) are embedded into the conversation. There are two types of code-switching:

- Inter-sentential Code-switching: defined as switching languages from one sentence to another. For example: "كان عاجبني. It was very interesting."
  (I liked it. It was very interesting.)

- Intra-sentential Code-switching (also known as code-mixing): defined as using multiple languages within the same sentence. For example:
  "احنا كنا بن implement a semantic search engine."
  (We were implementing a semantic search engine.)

Another type of language alternation is referred to as ”borrowing“ or ”loanword“. This is defined as having the whole sentence in one language except for words which are borrowed from the secondary language. For example: "انا بحب الresearch عمتا." (I generally love research.)

Researchers use different definitions for code-switching. Some researchers define code-switching as incorporating sentences from different languages, each having its own grammatical rules. Therefore, following this definition, borrowing is not considered as a type of code-switching. On the other hand, some researchers consider borrowing as a type of intra-sentential Code-switching. In the scope of this paper, for the sake of simplicity, we will use the term "code-switching" to refer to all types of language alternation; inter-sentential code-switching, intra-sentential code-switching as well as borrowing.

This phenomenon of multilingualism comes as a result of several factors such as globalization, immigration, colonization, the rise of education levels, as well as international business and communication. Code-switching is seen in several Arab countries, where for example, English is commonly used in Egypt and French in Morocco, Tunisia, Lebanon and Jordon.

Arabic varieties are characterized by Diglossia where dialectal forms usually differ considerably from their formal language, and are thus considered by researchers to be a separate language (Ferguson, 1959). For instance, the huge variation between MSA and the Egyptian Colloquial Arabic was shown in the study conducted by Kirchhoff and Vergyri (Kirchhoff and Vergyri, 2005). The authors studied the data sharing between both languages, where the author calculated the percentage of shared unigrams, bigrams and trigrams in the LDC CallHome corpus (telephone conversations in Egyptian Colloquial Arabic) and the FBIS corpus (broadcast news in MSA). It was found that the corpora only overlap by 10.3% in unigrams, 1% in bigrams and $< 1\%$ in trigrams. The same overlap computation was done for the CHRISTINE corpus (conversational British English) and the American English Broadcast News data from the NIST 2004 Rich Transcription evaluations. The overlap was found to be 44.5% in unigrams, 19.2% in bigrams and 5.3% in trigrams. Consequently, existing speech corpora in Modern Standard Arabic will be unsuitable for recognizing dialectal Arabic. Moreover, it is also unsuitable to use monolingual speech corpora to recognize code-switch speech.

In literature, there are two main approaches used to recognize code-switched speech: language-dependent and language-independent. In the language-dependent approach, monolingual ASR systems are used. This is done by first detecting the boundaries at which language switching occurs using language boundary detection (LBD) algorithms. For each language-homogeneous segment, the language is identified using language identity detection (LID) algorithms. Finally, each segment is recognized by its respective monolingual ASR system. This approach is suitable for building multilingual ASR systems that handle multiple languages, where the input speech is monolingual, as well as those that handle inter-sentential code-switching. However, in the case of intra-sentential code-switching, the language switching occurs within the same sentence and the language-homogeneous segments can be short. The accuracy of the LBD and LID algorithms can then become a limitation to the overall ASR system performance. In this case, the language-independent approach is more suitable for the task. It involves building a truly multilingual language model, acoustic model and pronunciation dictionary that encompass both- or all- languages involved. Recognition is then done in a one-pass approach. Building multilingual language and acoustic models is considered to be a holistic task. One of the main challenges is the need for code-switching corpora for training and testing purposes. Despite the great efforts done in collecting speech and text corpora, there is still shortage in the available corpora for code-switching languages. The lack of existing speech corpora for code-switched Egyptian Arabic-English is a bottleneck in the creation of ASR systems for conversational Egyptian Arabic. Few speech corpora are available for Dialectal Egyptian Arabic, such as CALLHOME (Canavan et al., 1997). Dialectal Egyptian Arabic speech corpora were also gathered by and (Elmahdy et al., 2009) and (El-Sakhawy et al., 2014). These corpora may contain a small percentage of code-switching, however, they are mainly designed for Dialectal Egyptian Arabic. Up to our knowledge, there are no speech corpora dedicated to code-switching occurring in dialectal Egyptian Arabic. In this paper, we present our first efforts in filling this gap. The collected corpus can be used in training and testing data for building a multilingual Egyptian Arabic-English ASR system. It can also serve as a data set for linguistic analysis on the Egyptian Arabic-English code-switching behaviour.

The remainder of the paper is structured as follows: Related work will be discussed in Section 2 . Section 3 describes the process of corpus creation; speech collection and transcription. In Section 4 , the corpus is analyzed to examine/provide insights on its code-switching features. Finally, Section 5 concludes and provides future work.

2. Related Work

There are two types of speech corpora: read speech and spontaneous speech. Spontaneous speech corpora are usually preferred in building ASR systems as they are closer to natural speech. However, they require manual transcription, which is a labour-intensive and time-consuming task. Both approaches have been recorded for collecting code-switched speech corpora. (Chan et al., 2005) collected a Cantonese-English speech corpus through read newspaper content. (Li et al., 2012) gathered a Mandarin-English speech from four different sources: (1) conversational meetings ; (2) group project meetings; (3) student interviews speech. The speech was then transcribed by an annotator and manually verified by Mandarin-English bilingual speakers. (Lyu et al., 2015) collected the SEAME corpus, where Mandarin-English audio recordings were collected from interviews and conversational speech and were manually transcribed. (Chong et al., 2012) developed a Malay conversational speech corpus, containing Malay-English code-switching through the recording and transcription of conversational speech.

3. Approach

The purpose of this work is to develop a corpus for spontaneous Egyptian Arabic-English code-switching speech. This was done by recording informal interviews rather than collecting read speech to gather casual speech of spontaneous nature. This decision was made under the assumption that code-switching occurs most frequently in spontaneous speech.

For the interview setups, each interview involved two interviewers and one or two interviewee(s). Only the interviewees' speech were transcribed. The participants were asked to discuss technical topics such as the courses they teach, work experiences, as well as their B.Sc., M.Sc. and Ph.D. projects. The participants were Egyptian teaching assistants in the German University in Cairo. The participants were of both genders, with their ages ranging between 23 and 28. The mother tongue of all participants is Arabic, and they are all fluent in English.

The interviews were recorded in a quiet closed room at the German University in Cairo. The recordings were collected using two table-mounted microphones. All audio recordings were stored in mono channel pulse-code modulation (PCM) at 16kHz sampling rate. In order to minimize the interference with power lines, the laptop charger was disconnected. The transcriptions were done manually by two Egyptian annotators who master both languages. The transcriptions were done using TranscriberAG. All transcriptions were recorded in UTF-8. Arabic transcriptions were written without diactritic marks. All non-speech sounds, pauses, and unintelligible speech segments were transcribed with some predefined filler tags.

4. Corpus Evaluation

The interviews were held with 12 participants; 6 males and 6 females. A total of 5.3 hours were recorded, with an average duration of a participant's recording of 26.5 minutes. A total of 4.5 hours were transcribed. The transcriptions made up a total of 1,234 sentences and 17,769 words, with an average of 14 words per sentence. The transcribed text is analyzed from two perspectives: (1) code-switching and code-mixing distribution, (2) part-of-Speech (POS) distribution of embedded words and POS trigger tags. The most frequently used Arabic trigger words (preceding a code-switching point) are identified. The most frequently used unigrams, bigrams and trigrams are also presented as well as examples from the transcriptions.

4.1. Code-switching and code-mixing analysis

The transcriptions contain a total of 17,769 words; 11,045 (62.1%) Arabic and 6,641 (37.4%) English words. This shows a high usage of the embedded English language in the conversations. The transcriptions show a high rate of code-switching and code-mixing. The 1,234 sentences are divided by language as follows: 124 monolingual Arabic, 125 monolingual English and 985 mixed. This shows a percentage of 79.8% of code-mixing in speech. It also shows a percentage of 10.1% of code-switching, where the whole sentence was uttered in the embedded language. Only 10% of the sentences are uttered purely in the matrix (Arabic) language.

The transcriptions were further analyzed from the code-mixing perspective. In the scope of the code-mixed sentences, 34.4% of the words are in the embedded language. The average number of code-switch points (Arabic-English and English-Arabic) in a sentence is 4. On average, there are 2 embedded English parts. The average number of English words in an embedded part is 2.

4.2. Trigger Words

Trigger words are defined as the Arabic words preceding a code-switching point. There are in total 535 unique Arabic words preceding a code-switching point. Table 4.2. shows the most frequent trigger words.

Trigger word	Percentage
ال	31.0%
في	4.8%
و	3.4%
يعني	1.5%
هو	1.3%

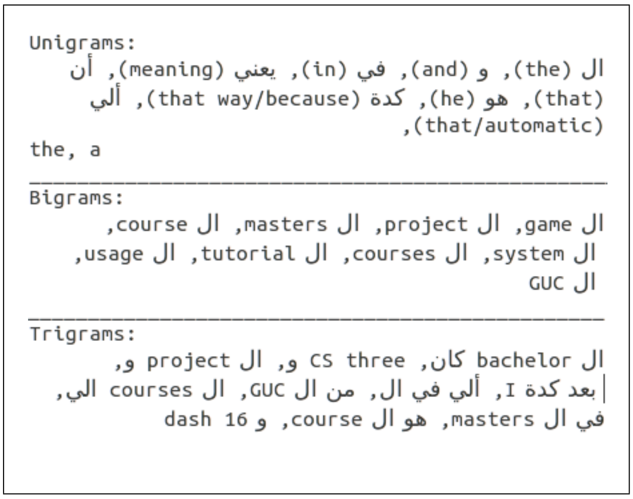
Table 1: The most frequent trigger words.

4.3. Top unigrams, bigrams and trigrams

The number of unigrams, bigrams and trigrams gathered from the transcriptions are shown in Table 4.3.. Figure 4.3. shows the most frequently used mixed n-grams in the conversations.

N-grams order	Count
unigrams	4,330
bigrams	14,205
trigrams	15,855

Table 2: The number of unigrams, bigrams and trigrams gathered from the transcriptions.



```
Unigrams:
    ال (the), و (and), في (in), يعني (meaning), أن
    (that), هو (he), كدة (that way/because), ألي
    (that/automatic),
the, a
______________________________________________
Bigrams:
    ال game, ال project, ال masters, ال course,
    ال system, ال courses, ال tutorial, ال usage,
    ال GUC
______________________________________________
Trigrams:
    ال bachelor كان, CS three و, ال project و,
    بعد كدة I, ألي في ال, من ال GUC, ال courses الي,
    في ال masters, هو ال course, و 16 dash
```

Figure 1: The top most frequently used unigrams, bigrams and trigrams.

4.4. Part-of-Speech analysis

The transcriptions were analyzed to determine the Part-of-speech (POS) of the embedded English words. A total of 388 sentences were manually annotated for POS tags. Figure 2 shows the POS distribution of the embedded English words. It can be seen from the data that the participants used the English language mostly in nouns.

The POS of the Arabic words preceding a code-switching point was also examined. These are considered to be trigger POS tags. Figure 3 shows the POS distribution of the trigger POS tags. It can be seen that code-switching mostly occurs after articles, followed by verbs and prepositions.

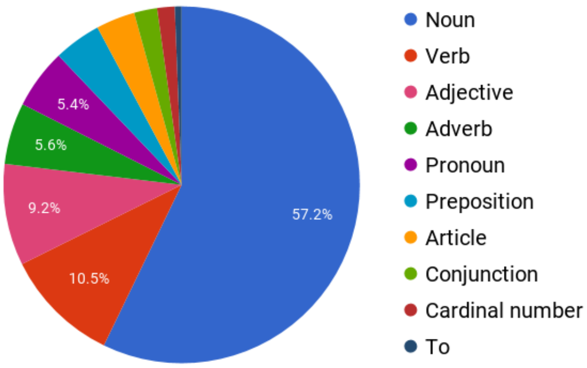


Figure 2: The POS distribution of the embedded English words in the set of manually annotated sentences.

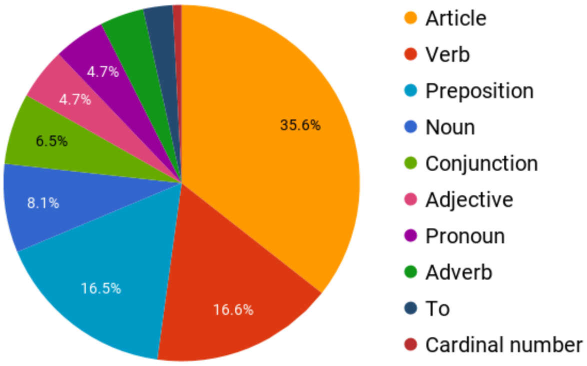


Figure 3: The POS distribution of the trigger POS tags in the set of manually annotated sentences.

4.5. Transcriptions examples

Figure 4 presents a sample of the transcribed sentences.

5. Conclusion

Code-switching has become a prevalent phenomenon in everyday conversations in multilingual societies. In Egypt, it has become common, especially among youth, to mix Arabic and English in conversations. This created a need for ASR systems to be able to recognize mixed Egyptian Arabic-English speech. In order to build such a multilingual ASR system, an Egyptian Arabic-English speech corpus is crucial. In this work, we collected spontaneous speech through informal interviews. The topics of the interviews were chosen to be technical, thus more probable to contain code-switching. The collected speech was manually annotated. The transcriptions were analyzed in terms of the code-switching and code-mixing behaviour. It was seen that both phenomena were used extensively. A subset of the transcribed data was manually annotated for POS tags. Analyses were performed on the POS tags of the English embedded words as well as the Arabic words preceding a code-switching point. It was found that English embedded words were mostly used for nouns, and that code-switching (from Arabic to English) occurs most frequently after articles. It was found that code-switching occurs at a rate of 30% after the artice ال. It is to be noted that in technical contexts, many technical terms are embedded as loanwords, thus contributing to this high code-switching rate. Accordingly, further research needs to be done to investigate the code-switching rate in other domains. We also intend to continue working on the corpus and collect more recordings and transcriptions.

6. Bibliographical References

Canavan, A., Zipperlen, G., and Graff, D. (1997). Callhome egyptian arabic speech. *Linguistic Data Consortium*.

Chan, J. Y., Ching, P., and Lee, T. (2005). Development of a cantonese-english code-mixing speech corpus. In *Ninth European Conference on Speech Communication and Technology*.

Chong, T. Y., Xiao, X., Tan, T.-P., Chng, E. S., and Li, H. (2012). Collection and annotation of malay conversational speech corpus. In *Speech Database and Assessments (Oriental COCOSDA), 2012 International Conference on*, pages 30–35. IEEE.

El-Sakhawy, D., Abdennadher, S., and Hamed, I. (2014). Collecting data for automatic speech recognition systems in dialectal arabic using games with a purpose. In *Multimodal Analyses enabling Artificial Agents in Human-Machine Interaction*. Springer LNAI.

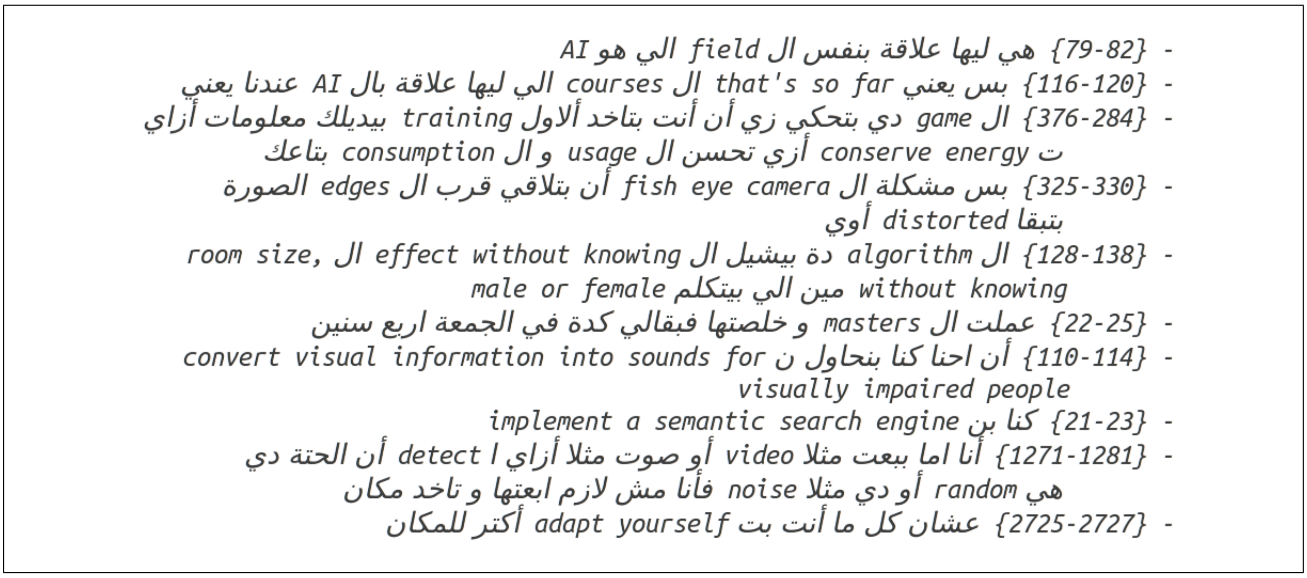
Elmahdy, M., Gruhn, R., Minker, W., and Abdennadher, S. (2009). Modern standard arabic based multilingual approach for dialectal arabic speech recognition. In *Eighth International Symposium on Natural Language Processing (SNLP'09)*, pages 169–174. IEEE.

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

Kirchhoff, K. and Vergyri, D. (2005). Cross-dialectal data sharing for acoustic modeling in arabic speech recognition. *Speech Communication*, 46(1):37–51.

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

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.



- {79-82} هي ليها علاقة بنفس ال field الي هو AI
- {116-120} بس يعني that's so far ال courses الي ليها علاقة بال AI عندنا يعني
- {376-284} ال game دي بتحكي زي أن أنت بتاخد ألاول training بيديلك معلومات أزاي
  ت conserve energy أزي تحسن ال usage و ال consumption بتاعك
- {325-330} بس مشكلة ال fish eye camera أن بتلاقي قرب ال edges الصورة
  بتبقا distorted أوي
- {128-138} ال algorithm دة بيشيل ال effect without knowing ال room size,
  without knowing مين الي بيتكلم male or female
- {22-25} عملت ال masters و خلصتها فبقالي كدة في الجمعة اربع سنين
- {110-114} أن احنا كنا بنحاول ن convert visual information into sounds for
  visually impaired people
- {21-23} كنا بن implement a semantic search engine
- {1271-1281} أنا اما ببعت مثلا video أو صوت مثلا أزاي ا detect أن الحتة دي
  هي random أو دي مثلا noise فأنا مش لازم ابعتها و تاخد مكان
- {2725-2727} عشان كل ما أنت بت adapt yourself أكتر للمكان

Figure 4: Some of the transcribed sentences. The time interval in seconds is shown for each sentence between braces.