Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3973–3977
Marseille, 11–16 May 2020

Cairo Student Code-Switch (CSCS) Corpus: An Annotated Egyptian Arabic-English Corpus

Mohamed Balabel[1,2], Injy Hamed[2,3], Slim Abdennadher[3], Ngoc Thang Vu[2], Özlem Çetinoğlu[2]
[1]IBM Germany Research and Development, Böblingen, Germany
[2]Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany,
[3]Computer Science Department, The German University in Cairo, Cairo, Egypt
mohamed.balabel1@ibm.com,
{hamediy,thang.vu,ozlem.cetinoglu}@ims.uni-stuttgart.de,
slim.abdennadher@guc.edu.eg

Abstract

Code-switching has become a prevalent phenomenon across many communities. It poses a challenge to NLP researchers, mainly due to the lack of available data needed for training and testing applications. In this paper, we introduce a new resource: a corpus of Egyptian-Arabic code-switch speech data that is fully tokenized, lemmatized and annotated for part-of-speech tags. Beside the corpus itself, we provide annotation guidelines to address the unique challenges of annotating code-switch data. Another challenge that we address is the fact that Egyptian Arabic orthography and grammar are not standardized.

1. Introduction

Code-switching (CS) is the act of using more than one language in text, or more commonly, in speech. According to Poplack (1980), there are several types of language alternations, which include:

- Intra-sentential CS: defined as using multiple languages within the same sentence. For example:
  "I do not think أني عاوز أبقى student any more." (I don't think I want to be a student anymore.)
- Inter-sentential CS: defined as switching languages from one sentence to another. For example:
  "It was really nice. اتعلمت كتير." (It was really nice. I learnt a lot.)

CS is a phenomenon commonly observed in the Arabic-speaking world. Arabic itself encompasses many varieties, most prominently modern standard Arabic (MSA) and dialectal Arabic (DA). The former is used in formal contexts such as formal writings and speeches, while the latter is the language used in everyday conversations as well as informal writings such as in social media. DA itself encompasses a number of dialects, of which the most widely used (by number of speakers) is Egyptian Arabic. Arabic speakers typically mix MSA and DA in the same utterances (Elfardy and Diab, 2012), sometimes in addition to English, French or both. Colonization and international business and education have played a major role in introducing the English and French languages into everyday conversations.
Despite Arabic being one of the most widely-used languages, there is still a huge gap in the available resources and Natural Language Processing (NLP) applications. Most of the work has focused on MSA, with comparatively less work on DA, and even less on code-switched DA. In this paper, we aim at providing a multi-lingual corpus[1] for code-switched Egyptian Arabic-English that can be used in several NLP tasks. We build on the speech data in Hamed et al. (2018) by tokenizing, lemmatizing and annotating a subset of the corpus with part-of-speech (POS) tags. The languages used in the speech data are Egyptian Arabic and English. Although code-switching between MSA and DA is a common phenomenon, we do not observe it in our data. The corpus can thus be used as a testbed for several NLP applications, ranging from speech recognition to text-based applications, and gives foundation towards building an Egyptian Arabic-English treebank.
A substantial part of our effort to develop this corpus is geared towards creating coherent and extensive guidelines that cover all the necessary areas from orthography to tokenization and POS tagging. The challenges stem in part from the fact that Egyptian Arabic (and DA in general) is not standardized, but most importantly from the fact that CS data poses unique challenges during annotation that typically do not arise when annotating the individual languages independently. Beside the corpus itself, the complete annotation guidelines are also publicly available so that future research in the same direction may benefit from them.

[1]The corpus is publicly available and can be obtained from the authors.

2. Related Work

One of the main factors hindering the development of NLP applications that handle CS is the lack of corpora, whose collection can be challenging (Çetinoğlu et al., 2016). It is unsurprising that the available corpora are scarce. In this section, we present an overview of corpora similar to ours, i.e. annotated CS corpora, then give a brief overview of other ones that are related to ours only in that they involve the Arabic language.

2.1. Morpho-Syntactically Annotated CS Corpora

A number of researchers have worked to provide morpho-syntactically annotated CS corpora, including treebanks. In all the cases we are aware of, the raw data source is either social media (Çetinoğlu, 2016; Bhat et al., 2017) or transcribed conversations (Çetinoğlu, 2017). The corpus

we present here is no exception, since it is also based on transcribed speech data. [2]An early effort was the annotation of an English-Spanish corpus with part-of-speech tags (Solorio and Liu, 2008). This was before the popularity of the Universal Dependencies (UD) scheme (Nivre et al., 2016). Hence, for each of the languages, a different tagset was used. More recent efforts have made use of the unified UD tagset and unified guidelines for morpho-syntactic annotation, in order to make available CS corpora in which both languages use the same annotation scheme, making it easier for subsequent research (e.g. in linguistics) to extract insights from the data, but also for NLP researchers to make use of monolingual UD resources. This is evidenced in the Turkish-German CS corpus (Çetinoğlu and Çöltekin, 2016) and the Hindi-English treebank (Bhat et al., 2017; Bhat et al., 2018).

Although this is not an extensive overview, the number of morpho-syntactically annotated CS corpora is still very limited, and so is the number of language pairs that are represented in them. Our corpus is therefore a valuable addition, especially as it provides data for a previously unrepresented language pair.

2.2. Arabic CS Corpora

Recently, more attention has been given by researchers towards the CS phenomenon. Text corpora have been collected (Mustafa and Suleman, 2011; Hamed et al., 2017) by crawling documents and web pages related to computers and computer science. Some corpora have also covered the Algerian Arabic-French language. Cotterell et al. (2014) present a text corpus of 339,504 comments with romanized Arabic[3] collected from news story of an Algerian newspaper and annotated for word-level language ID. Samih and Maier (2016) also provided a text corpus for the Arabic-Moroccan Darija language. The corpus contains 233,000 token crawled from internet discussion forums and blogs. Sabty et al. (2019) have collected a text corpus of 1,331 sentences from speech transcriptions (Hamed et al., 2018) and Twitter and annotated them with named-entities. Abainia (2019) has provided an Algerian Arabic-French parallel corpus containing $2,400$ Facebook comments annotated for gender, regions and cities, named entities, emotion and level of abuse. Amazouz et al. (2018) collected the FACST Algerian Arabic-French speech corpus having 7.5 hours of read and spontaneous speech. The speech was annotated with transcription, sentence segmentation, language boundary and word- and phone-level time codes. Mohdeb-Amazouz et al. (2016) provide a 53 hours Maghrebian[4] Arabic-French speech corpus obtained from TV entertainment and talk shows. The speech was annotated for sentence segmentation and language tags. Ismail (2015) also investigated the CS behavior for Saudi-English bilinguals, where 89 minutes of informal dinner gatherings of 3 Saudi couples were recorded and transcribed. Apart from mixing Arabic with other languages, there has been also work on CS between different varieties of Arabic itself (Elfardy and Diab, 2012). To the best of our knowledge, the corpus we present here is the first to provide complete annotations across the three different annotation layers: tokenization, lemmatization and POS tagging.

3. Data Collection And Annotation

The whole corpus is based on the work of Hamed et al. (2018), whose aim was to collect Egyptian Arabic-English speech data. Interviews were conducted with 12 participants, all affiliated to the German University in Cairo. The result was 5.3 hours of recorded speech. A selected subset of those hours was divided among a number of transcribers, tasked with transcribing only the interviewees' speech and ignoring the interviewers' speech. There were, however, no clear orthographic guidelines, and thus the orthographic conventions used for Egyptian Arabic varied from one transcriber to another. A pilot study followed, where 388 sentences were manually annotated with POS tags[5]. However, this was done without proper tokenization and POS tagging guidelines. Moreover, there was no unified framework to ensure consistency across both languages and a proper handling of code-switched words.

To develop our corpus, we select the speech data for six out of the twelve participants from the speech corpus described above. Then we devise a setup composed of two phases. In the first phase, we develop orthographic guidelines, which we describe in the next section. We then assign three transcribers to transcribe the data according to the guidelines. After obtaining the transcriptions, everything is revised and improved manually by two of the authors who are both native in Arabic and near-native in English. The second phase is dedicated to the annotation. This includes three consecutive tasks -namely tokenization, POS tagging and lemmatization- conducted in that order. For each task, we first develop annotation guidelines as described in the next section. Then we provide gold-standard, manual annotations. The labor is always divided among two of the authors (the same who revise the transcriptions), then each one annotates their part and revises the part of the other annotator. Disagreements are resolved via thorough discussions among the authors until consensus is reached. Due to lack of resources, no external annotators are recruited for this phase. All the annotations are done in the CONLL-U format[6].

4. Guidelines

4.1. Orthography

Unlike MSA, DA orthography is not standardized, which poses a challenge to NLP tasks, as the same word is represented differently across various texts. Our first goal is to tackle this problem by presenting guidelines for Egyptian Arabic writing. We abide by these writings when obtaining corpus transcriptions to insure standardization throughout the text. There have been attempts to standardize DA orthography, most notably the work of Habash et al. (2012), which attempts to remain as close to MSA orthography as

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

[3]Arabic written in the Latin script

[4]Algerian, Moroccan and Tunisian

[5]Not using the UD tagset.

[6]https://universaldependencies.org/format.html

possible, and diverges only when needed. However, we decided to base our orthography guidelines on the ones developed and used by the Egyptian Arabic Wikipedia community [7].

There are two main reasons for preferring the Wikipedia guidelines over guidelines used in previous academic studies. First, we keep in mind that future researchers might want to use our corpus not in separation from, but rather in conjunction with data freely available online. Wikipedia is a great source of data, and our choice makes our corpus orthographically compatible with Wikipedia data. Second, our personal observation is that the Wikipedia conventions have gained some popularity among the online community since their first publication. Therefore, by opting for the Wikipedia conventions we remain close to other online content. One advantage of the mentioned guidelines is that they diverge a lot from MSA orthography in order to stay close to the actual pronunciation. It is important to note however that the Wikipedia guidelines form only the basis of our guidelines. We make some extensions, mainly in cases that are left ambiguous in Wikipedia or when many orthographic varieties are still permitted. Here, we make coherent decisions to restrict the number of possibilities -usually to one variant only- so as to reduce ambiguity.

4.2. Annotation

For the different annotation layers, we base our guidelines on the Universal Dependencies (UD) scheme. We are confronted with two challenges, which we have to resolve in order to develop coherent guidelines in the spirit of the UD community.

The first challenge is the fact that, unlike similar code-switch corpora for which UD annotated corpora existed for the individual languages, there are no UD-annotated treebanks for Egyptian Arabic at the time of this writing. There exist, however, multiple treebanks for MSA (Taji et al., 2017; Hajic et al., 2004). We have therefore to adapt the annotation principles used for MSA to our concrete use case, namely Egyptian Arabic. Just as for orthography, we diverge a lot from MSA in order to reflect the true morpho-syntactic properties of Egyptian Arabic. For instance, consider the annotation of the first participle (called in Arabic اسم الفاعل) in both varieties. In MSA corpora, this has always to be considered a noun or an adjective. However, we observe it also used as a verb in Egyptian Arabic, that can take direct, indirect or prepositional objects:

- أنا مش جايبهولك (I am not bringing it for you)
- لسه باعتاها امبارح (I just sent it yesterday)

The second challenge is common to all CS corpora, namely the need to resolve conflicts that arise when unifying the morpho-syntactic guidelines of two distinct languages. This challenge has been discussed in Çetinoğlu and Çöltekin (2019) in the context of Turkish-German CS. Even though the goal of UD is to annotate similar structures across different languages in a unified fashion, we observe that this does not entirely hold true, perhaps especially for languages pertaining to different language families. One benefit of annotating code-switched corpora is that such discrepancies between languages become noticed. That is, if the same structure had been annotated differently for two different languages, say MSA and English, for which the treebanks had been developed independently, it now becomes clear when the two languages are juxtaposed in CS data that the annotation principles for at least one of the languages have to change in order to achieve the goal of having a unified annotation across languages.

Clearly, that second challenge is particularly important to solve in order to annotate CS data. Before we embark on developing the guidelines, we establish two guiding principles: (a) to be as coherent and consistent as possible, and (b) to be accurate in capturing the linguistic phenomena that emerge from CS. We identify three core strategies -which we here call operations- that are needed to annotate code-switch data:

1. unification
2. discrimination
3. disambiguation

The need for **unification** arises, as mentioned earlier, when the annotation principles for two languages diverge for a syntactic structure that can be considered linguistically similar. Here we propose retaining the annotation principles of the language that is closest in spirit to the UD scheme in general[8] and carrying them over to the other language.

Sometimes, however, two syntactic structures that at first glance appear to be similar turn out to represent genuinely different structures in the respective languages. In this case, instead of unifying, we need to **discriminate** those different structures. One example is the use of the definite article in English and Arabic, where in the former it is considered a separate token with the POS tag DET (determiner), while in the latter it is considered a morpheme that attaches as a prefix to nouns and adjectives when they are in the definite state. To clarify this, consider the following examples that are the same in MSA and DA:

- بيت كبير (a big house, literally: big house)
- البيت الكبير (the big house, literally: the.big the.house)

The examples show that in Arabic, definiteness is not marked by an article that attaches to a noun phrase (as in many European languages), but rather by a morpheme that attaches to nouns (as well as adjectives, because of the concord with nouns), even though superficially, this morpheme resembles a definite article like the English *the*.

Yet, an interesting question arises here, namely: do speakers who code-switch always maintain the difference between such superficially similar but actually different structures across languages? We show here two examples from our corpus before we answer that question:

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

[8] https://universaldependencies.org/guidelines.html

1. الfactor الأولاني (the first factor, literally: the.factor the.first)
2. ال perspective camera (the perspective camera)

In phrase 1, it is apparent that the speaker uses Arabic grammar rules, even though the noun is in English. That is, the noun as well as the adjective that is in concord with it are both in definite state and carry the Arabic definiteness morpheme. However, in phrase 2, things are reversed, at least partially. The overall structure follows English grammar rules, with a definite article attaching to a whole noun phrase. Surprisingly, the definite article is not the English one, but the Arabic definite morpheme. This illustrates what we mean by discriminating different structures. It is not correct to say that the Arabic definite marker has always to be annotated in a certain way, regardless of the conrete usage. What matters is the structure. To answer the question from above: the sentences in our corpus suggest that the answer is no. The speakers do not keep the different structures of both languages independent from each other. They mix them in sometimes rather unexpected ways.

This brings us to the third and last operation, namely **disambiguation**. How can we deal with a code-switched word or phrase which can be interpreted differently according to which grammar we use to explain it. For instance should we consider الtext as one token (an English noun in Arabic definite state) or as two tokens, a definite article followed by a noun? Without further evidence, it appears to the authors that in fact both explanations are plausible. Although there exist frameworks that can handle such uncertainty (Plank et al., 2014), it is conventional for treebanks that are encoded in CONLL-U format to list only one annotation for each sentence. In this case, as we observe that it is more often the case that Arabic grammar is thrust upon English words than vice versa, we decide to consider the Arabic interpretation the "correct" one.

5. Corpus Overview

In this initial release, the corpus consists of 1153 transcribed sentences, pertaining to 6 different participants. Around 68% of the sentences are code-switched, that is, there are 788 sentences with intra-sentential code-switching. The remainder of the sentences exhibit a lot of inter-sentential code-mixing, i.e. switching from one sentence in a certain language to a sentence in another language. The sentences contain a total of 11,286 tokens, but only 2152 unique tokens. Roughly 4% of the tokens are code-switched, meaning they contain morphemes from both Egyptian Arabic and English. In this context, it is also important to point at the fact that the UD scheme distinguishes between (regular) tokens and multiword tokens. A multiword token is a single orthographic token that actually corresponds to multiple syntactic ones. For instance, in the phrase "the student's grade" the word *student's* can be considered a multiword token, comprised of the two tokens *student* and *'s*. Arabic, especially DA, relies heavily on cliticization, thus we encounter a significant number of multiword tokens in our corpus. Around 5% of those are code-switched. Note that multiword tokens are not always necessarily marked explicitly in English treebanks. However, in order to be consistent with the Egyptian Arabic annotation, we mark them explicitly in our corpus. Regarding the lemmas, there are 1407 unique lemmas for both languages together. Table 1 gives a summary of the number of sentences and tokens.

count	English	Arabic	CS	total
sentences	62	303	788	1153
tokens	2752	8097	437	11286
multiword tokens	51	1109	62	1222

Table 1: Overview of the number of sentences and tokens.

POS	English	Arabic	CS
NOUN	1073	853	335
VERB	197	1251	74
PROPN	111	79	16
ADJ	231	232	12

Table 2: Breakdown of POS counts by language. Only POS tags for which at least one code-switched token exists are shown in the table.

Next, we look at the relation between CS and the part of speech of a word. Table 2 breaks down the code-switched tokens by POS tag. As can be seen, the number of code-switched tokens is highest for nouns, moderate for verbs, low for adjectives and proper nouns, and zero for all other POS tags. Apart from this, it may be also of interest to study the dominance of one language over the other depending on the part of speech of a word. One observation here is that English is used more often than Arabic when it comes to nouns (including proper nouns), but Arabic is preferred for verbs (including auxiliaries, which are not shown in the table). Regarding adjectives, there is a tie between both languages, as the number of adjectives in both Arabic and English is roughly similar. Another category where this is also the case is numbers (NUM).

6. Summary And Future Work

We presented a novel Egyptian Arabic-English corpus, containing 1153 sentences that have been tokenized, lemmatized and tagged with POS. The main contributions are (a) developing guidelines to address the challenges of annotating CS data, and (b) developing the corpus itself, which we find is a valuable addition given the scarcity of annotated CS data. Currently, new data is being transcribed and annotated, and a second, extended version of our corpus is being planned. Additionally, we are developing guidelines for the syntactic annotation of dependencies according to the UD scheme. In the future, we would like to add the dependency relations to our corpus, so that it grows into a treebank. Finally, researchers interested in the structural analysis of CS data can use our corpus for linguistic studies, or as a testbed for various NLP tasks.

7. Acknowledgement

This project has benefited from financial support to Mohamed Balabel and Özlem Çetinoğlu by DFG via project CE 326/1-1 "Computational Structural Analysis of German-Turkish Code-Switching" (SAGT).

8. Bibliographical References

Abainia, K. (2019). Dzdc12: a new multipurpose parallel algerian arabizi–french code-switched corpus. *Language Resources and Evaluation*, pages 1–37.

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

Bhat, I., Bhat, R. A., Shrivastava, M., and Sharma, D. (2017). Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 324–330.

Bhat, I., Bhat, R. A., Shrivastava, M., and Sharma, D. (2018). Universal dependency parsing for hindi-english code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 987–998.

Çetinoğlu, Ö. (2016). A turkish-german code-switching corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Çetinoğlu, Ö. and Çöltekin, Ç. (2016). Part of speech annotation of a Turkish-German code-switching corpus. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 120–130, Berlin, Germany, August. Association for Computational Linguistics.

Çetinoğlu, Ö. and Çöltekin, Ç. (2019). Challenges of annotating a code-switching treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90.

Çetinoğlu, Ö., Schulz, S., and Vu, N. T. (2016). Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11.

Çetinoğlu, Ö. (2017). A code-switching corpus of Turkish-German conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 34–40, Valencia, Spain, April. Association for Computational Linguistics.

Cotterell, R., Renduchintala, A., Saphra, N., and Callison-Burch, C. (2014). An algerian arabic-french code-switched corpus. In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Programme*, page 34.

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

Habash, N., Diab, M. T., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *LREC*, pages 711–718.

Hajic, J., Smrz, O., Zemánek, P., Šnaidauf, J., and Beška, E. (2004). Prague arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pages 110–117.

Hamed, I., Elmahdy, M., and Abdennadher, S. (2017). Building a first language model for code-switch arabic-english. *Procedia Computer Science*, 117:208–216.

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)*.

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

Mohdeb-Amazouz, D., Martine, A.-D., and Lamel, L. (2016). Arabic-french code-switching across maghreb arabic dialects: a quantitative analysis.

Mustafa, M. and Suleman, H. (2011). Building a multilingual and mixed arabic-english corpus. In *Proceedings Arabic Language Technology International Conference (ALTIC)*.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Plank, B., Hovy, D., and Søgaard, A. (2014). Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751.

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

Sabty, C., Elmahdy, M., and Abdennadher, S. (2019). Named entity recognition on arabic-english code-mixed data. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 93–97. IEEE.

Samih, Y. and Maier, W. (2016). An arabic-moroccan darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4170–4175.

Solorio, T. and Liu, Y. (2008). Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060. Association for Computational Linguistics.

Taji, D., Habash, N., and Zeman, D. (2017). Universal dependencies for arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176.