Identifying Code-switching in Arabizi

Safaa Shehadi and **Shuly Wintner**
Department of Computer Science
University of Haifa, Israel
safa.shehadi@gmail.com, shuly@cs.haifa.ac.il

Abstract

We describe a corpus of social media posts that include utterances in Arabizi, a Roman-script rendering of Arabic, mixed with other languages, notably English, French, and Arabic written in the Arabic script. We manually annotated a subset of the texts with word-level language IDs; this is a non-trivial task due to the nature of mixed-language writing, especially on social media. We developed classifiers that can accurately predict the language ID tags. Then, we extended the word-level predictions to identify sentences that include Arabizi (and code-switching), and applied the classifiers to the raw corpus, thereby harvesting a large number of additional instances. The result is a large-scale dataset of Arabizi, with precise indications of code-switching between Arabizi and English, French, and Arabic.

1 Introduction

Arabizi is a writing system for (primarily dialectal) Arabic that uses the Roman alphabet. It is ubiquitous on social media outlets, and has many characteristics of social media writings in other languages (e.g., slang, tendency towards the spoken register, spelling errors, abbreviations, character repetition, use of emoticons, etc.) The use of the Roman alphabet facilitates (and perhaps even encourages) *code-switching*: moving between Arabic (represented in Arabizi) and other languages, notably English and French, sometimes even within the same sentence.

Code-switching is becoming more and more prevalent as the world's population is becoming more multilingual (Grosjean, 1998). It is a natural phenomenon that is triggered by linguistic, sociolinguistic, psycholinguistic, demographic, and contextual prompts, and has been studied mainly in the spoken language until recently. With the ubiquity of text online, however, code-switching is beginning to be investigated also in the written language (e.g., Solorio and Liu, 2008; Solorio et al., 2014; Aguilar et al., 2018; Solorio et al., 2021). Such research has various practical applications, both for understanding and for generation of code-switched language (Sitaram et al., 2019; Doğruöz et al., 2021). Our main interest is in code-switching phenomena in Arabizi; in order to better understand them, a large dataset of Arabizi is required.

The main goal of this work is to construct a large-corpus of Arabizi utterances, potentially including instances of code-switching between Arabizi and English, French, or Arabic written in the Arabic script. The dataset is based on social media posts from two outlets: Twitter and Reddit. To collect the data, we implemented a classifier that can identify sentences containing words in Arabizi, Arabic, English, and French, and used it to filter out texts harvested from the two outlets.

We describe the dataset and the methods we used to curate it (Section 3). We then discuss the challenge of determining the language ID of words in multilingual texts, and describe classifiers that can accurately predict such language tags, based on a schema we developed for the task (Section 4). We extend the word-level classifiers to sentence-level ones, assigning a complex tag to each sentence that indicates the presence of words from various categories (i.e., languages) in it (Section 5). Finally, we use the classifiers to extract additional instances of sentences with Arabizi (and with code-switching) from our raw corpus (Section 6).

This paper makes several contributions: 1. We release a large-scale corpus of Twitter and Reddit posts that include Arabizi; 2. We introduce a novel annotation scheme that determines the language of words in multilingual utterances; specifically, we advocate a unique tag for words that can be included in more than one mental lexicon (and hence trigger code-switching); 3. We release a portion of the dataset, manually annotated according to this annotation scheme; and 4. We provide highly-accurate classifiers that can determine the language

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 194 - 204

ID tags of words in this corpus; the classifiers were used to identify hundreds of thousands of additional sentences that are very likely to include Arabizi in general and code-switching with Arabizi in particular. We expect these resources, which are all publicly available,[1] to be instrumental for future research in code-switching and in Arabizi.

2 Related work

Arabizi has attracted some interest in recent years, and various works address the tasks of detecting it and converting Arabizi to the Arabic script. Darwish (2014) used word- and sequence-level features to identify Arabizi mixed with English and achieved 98.5% accuracy on the identification task. He argued that classifying a word as Arabizi or English has to be done *in context*, and thus employed sequence labeling using Conditional Random Fields (CRF) for classification. The data were selected from Twitter, by querying (three) commonly used Arabizi words and then extracting the user IDs of all the authors of the resulting tweets, obtaining all their tweets, under the assumption that authors who use Arabizi once may use it often. Then, tweets in which most of the words contained Arabic letters were filtered out. This resulted in 522 tweets consisting of 5207 tokens, of which 1203 were in Arabizi.

Cotterell et al. (2014) compiled a corpus of more than half a million pages from an Algerian newspaper website, from which they extracted almost 7M tokens which were annotated for language, using three tags: Arabic, French, or Other. More recently, Samih and Maier (2016) compiled a corpus of Arabic mixed with Moroccan Darija, in which tokens were assigned to seven categories: three for languages, and then mixed (morphemes from more than one language in the same token), named entity, ambiguous and other. In total, 223K tokens were annotated.

The task of transliterating Arabizi to Arabic was addressed by Al-Badrashiny et al. (2014), who employed finite-state transducers, a language model and morphological processors for Arabic. They used a dataset consisting of 1500 words only. This approach was then extended to the Tunisian dialect (Masmoudi et al., 2015). The transliteration task was applied to the Tunisian dialect in a more recent work (Younes et al., 2022), using contemporary machine-learning techniques, but the datasets remained relatively small. Shazal et al. (2020) addressed the joint task of identifying Arabizi and transliterating it to the Arabic script, reporting high word accuracy on a large (1M token) dataset.

Tobaili (2016) trained an SVM classifier to identify Arabizi in multilingual Twitter data. He assumed that in order to tag a tweet as Arabizi it should have more Arabizi words than English words. The best results were obtained using three features: (1) the languages as detected by *Langdetect*; (2) the language as detected by the twitter API; and (3) the count of word occurrences per tweet. The dataset used in this work is small, and has merely 465 Arabizi sentences from Lebanon and 955 from Egypt. Tobaili (2016) also found that the use of Arabizi differed between Egypt and Lebanon (for example, more omission of vowels in the former, and more mixed language in the latter).

Two Arabizi datasets were recently compiled and released (Baert et al., 2020): LAD, a corpus of 7.7M tweets written in Arabizi; and SALAD, a randomly-selected subset of LAD, containing 1700 tweets, manually annotated for sentiment analysis. The tweets were harvested using Twint: Twitter Intelligence Tool, by setting 48 common words in Egyptian as seeds. This work focused mainly on the Egyptian dialect, and the manually-annotated dataset is rather small.

Seddah et al. (2020) built the first North-African Arabizi treebank. It contains 1500 sentences, fully annotated with morpho-syntactic and Universal Dependency codes, with full translation at both the word and the sentence levels. It is also supplemented by 50K unlabeled sentences collected using web-crawling. The texts reflect the Algerian dialect, and contain 36% French tokens. Recently, this dataset was extended by adding transliterations of all the Arabizi tokens, as well as sentence-level annotations of sentiment and topic (Touileb and Barnes, 2021).

Adouane et al. (2016) focused on the task of identifying Arabizi (and Romanized Bereber) in social media texts, reporting near-perfect accuracy using very simple character-ngram features. The data were collected from North-African sources and reflect these dialects. More recently, Younes et al. (2020) used deep learning methods to identify the language of words in Tunisian social media texts. They defined five categories for the classification (Tunisian dialect words, foreign langauge words, punctuation, symbols, and emoticons) and

[1] Available from https://github.com/HaifaCLG/Arabizi.

reported almost perfect accuracy on this task.

One of our goals in this work is to create a large dataset of sentences containing Arabizi, potentially mixed with words in other languages, focusing on the Egyptian and Lebanese dialects. Unlike much existing work, we annotate our dataset at the word level, thereby yielding a richer annotation that clearly outlines sentences with code-switching. Our language ID annotation scheme acknowledges the difficulty of assigning language ID tags to words that may be shared by more than one mental lexicon; such words, which include proper names and cognates, are assumed to trigger code-switching (Clyne, 2003; Broersma and De Bot, 2006; Broersma, 2009; Soto et al., 2018; Soto and Hirschberg, 2019). We then use our annotated dataset to train classifiers that we employ to extract more code-switched Arabizi instances from Reddit and Twitter, thereby extending the scope of our dataset significantly.

3 Data collection

We conjectured that social media outlets, particularly Reddit and Twitter, would include a sizable amount of Arabizi utterances. To identify them, we modified the method suggested by Rabinovich et al. (2018), which has subsequently been used also to harvest code-switched data from Reddit (Rabinovich et al., 2019).

First, we identified some Reddit fora ('subreddits') where we expected to find Arabizi used. These included `r/arab`, `r/arabs`, `r/egypt`, `r/jordan`, `r/lebanon`, and `r/syria`. We downloaded the entire collection of the above subreddits. The resulting (raw) Reddit dataset consisted of 3,584,915 sentences, 59,593,594 words and 72,305 authors.

For twitter, we followed Darwish (2014) and defined a few dialectal Arabic seed words that we expected to occur with high frequency in Arabizi texts, focusing on the Egyptian dialect (where we expected to find code-switching with English) and the Lebanese dialect (where we expected mixed French). These seed terms are listed in Appendix A. We located and retained tweets that included any of the seed words in our list. We then extracted the user IDs of authors of such texts, under the assumption that authors that use Arabizi in some tweets are likely to use it elsewhere, too; and we included all tweets authored by these users in our corpus. The resulting (raw) Twitter dataset consisted of 2,466,642 sentences (22,530,044 words) authored by 1090 users: 936 Egyptians and 154 Lebanese.

We used NLTK (Bird et al., 2009) for sentence boundary detection and tokenization. As the tokenizer did not split emojis from other tokens, we added a simple post-processing step to make sure all emojis were standalone tokens. We removed extra spaces and separated Arabic letters from non-Arabic ones. We also shortened adjacent repeated letters to only two (e.g., we converted '*ahhhhh edaaa thankkk youuuu*' to '*ahh edaa thankk youu*').

Next, we aimed to identify sentences containing Arabizi in the raw dataset. We first utilized a number of language identification tools, including *Spacy* (Honnibal et al., 2020), Google's *LangDetect* (we used the Python port), *langid* (Lui and Baldwin, 2011, 2012), and *FastText* (Joulin et al., 2017). Unsurprisingly, they all failed to detect Arabizi with acceptable accuracy.

To evaluate the accuracy of existing language ID tools on Arabizi we selected 100 sentences from the annotated Arabizi dataset of Tobaili (2016): the first 50 sentences containing only Arabizi words from the Egypt dataset, and the first 50 from the Lebanon dataset. We applied the above-mentioned classifiers to these 100 sentences; since none of the tools was trained on Arabizi data, none predicted Arabizi. But they did not predict Arabic, either: instead, *Langdetect* defaulted to Somali 43 times, (and Indonesian 25 times); *Langid* detected English, Spanish-Castilian, Indonesian, and Swahili for 50 of the sentences; *Fasttext* preferred English and Spanish; and *Spacy* identified half of the sentences as Somali or Indonesian.

We therefore resorted to defining our own language ID detection model, which we specifically tuned to identifying Arabizi (in addition to English and French). We developed a dedicated scheme for tagging words in a mixed-language dataset (Section 4.1), manually tagged a sizeable number of sentences reflecting the various language combinations witnessed in the dataset (Section 4.2), and then used the manually annotated subset to train classifiers (Sections 4.3–4.4) that can assign language ID tags to words in unseen texts. Finally, we extended the annotation from words to sentences (Section 5) in order to devise an efficient extractor for more instances of code-switched Arabizi from our corpus. We now detail these stages.

4 Word level classification

Some existing work on Arabizi focused on identifying the language of a sentence, or a larger chunk of text. For example, Tobaili (2016) defined a tweet as Arabizi if it contained at least 50% Arabizi tokens. In contrast, we focus on identifying the language of each individual token in the corpus, as our main motivation is to prepare a dataset suitable for research on code-switching, which may of course be intra-sentential. As mentioned above, existing tools for word-level language ID fail miserably when Arabizi is concerned.

We begin by discussing the challenges involved in word-level annotation of multilingual texts (Section 4.1), detail the manual annotation (Section 4.2), and then discuss our classifiers, both statistical (Section 4.3) and neural (Section 4.4).

4.1 Annotation of language ID

Annotating multilingual data for language is challenging, especially where named entities are involved. Much work on code-switching assumes that a switch is defined when two consecutive words come from two different languages; and much cognitive linguistic work focuses on understanding what facilitate such switches. Specifically, it has been suggested that *cognates* (words in two languages that share a similar form and a similar meaning) facilitate code-switching (Clyne, 2003; Broersma and De Bot, 2006; Soto and Hirschberg, 2019). However, assigning a clear language tag to words in multilingual texts may not always be possible (Clyne, 2003, Chapter 3).

Consider the case of *borrowing*: a French word may be borrowed by Arabic, and sound like a foreign word initially, during which period its use in an otherwise Arabic sentence may be considered an insertional switch (e.g., `balcon` '*balcony*'). With time, this word may obtain properties of the borrowing language (its phonology might be adapted to Arabic, it may obtain Arabic morphological affixes, etc.), until finally it may be considered by native Arabic speakers, including monolinguals, a common Arabic word. How should such words be tagged during various stages of their assimilation?

Similarly, *culturally-specific words* in one language may be borrowed into another language simply because they have no translation equivalents in the borrowing language. For example, Arabic `alhamdulillah` '*thank God*' can be used verbatim in an otherwise English (or French) text. This may extend also to common nouns, for example `mjadara` '*mujadara, a lentil-based dish*'.

A particularly challenging case is *named entities* (which are often the extreme case of cognates). They can have identical forms in the two languages (e.g., '*Beirut*' in Arabic and in English); but they may also be adapted to the phonology of each language, and thus drift apart from each other (e.g., `Amreeca` '*America*', `Surya` '*Syria*', `Alqahirah` '*Cairo*'). The distance between the two forms may be significant (e.g., `al-Jazair` '*Algeria*'). Sometimes, proper names are translated rather than adapted (e.g., `al-welayat al-muttahida` '*United States*'), or use different words altogether (e.g., `masr` '*Egypt*'). What language ID tag should we assign to such tokens in multilingual texts?

Several decisions must be taken in order for the annotation to be consistent, and not all decisions can always be fully justified. Our motivation in devising the annotation scheme was to facilitate consistency by providing clear and easy-to-apply guidelines. We thus defined the following categories:

0: Arabizi including any form variant that may be considered Arabizi;

1: English including common social media variants of words such as spelling errors, shorthand (`Idk` '*I don't know*', `plz` '*please*'), letter repetition (`nooooo` '*no*', `Cuuute` '*cute*'), etc.;

2: French with similar social media accommodations;

3: Arabic written in the Arabic script;

4: Shared see below;

5: Other tokens that are either non-linguistic or common to several languages. These include punctuation marks, numbers, emoticons and emojis, etc. As we focus only on Arabic, English and French, we also mark tokens in other languages as 'other'. Examples include '*Bhag hindu ka baccha*', '*Eww!*', '*12k?*', and '*ahahahaha*'. Notice that morphological indications of language may change a token from 'Other' to that language; e.g., '*1st*' or '*3rd*' are considered English.

In light of our focus on code-switching, we defined the category *shared* to include words that we have reasons to believe may belong to more than one mental lexicon (or, alternatively, to a shared mental lexicon). In the linguistic literature, *trigger words* are defined as words that are positively associated with code-switching, either because they are

cognates or because they increase the facilitation of the other language (Clyne, 2003; Broersma and De Bot, 2006). Our annotation guidelines were the following; notice that in all these cases, the annotation is context-independent: the same token will be tagged uniformly independently of where it occurs.

- Arabizi named entities which have different (translated) counterparts in English are tagged as Arabizi, and their translation equivalents are considered English; e.g., `Al-Emirat Al-Arabiya Al-mutahida` '*United Arab Emirates*', `masr` '*Egypt*', `al-maghrib` '*Morocco*'.
- Named entities in Arabizi and English that are *not* translated, and hence are written in a similar way in both languages, are considered as shared words; e.g., `al-ordon` '*Jordan*', `alqahirah` '*Cairo*', `Lubnan` '*Lebanon*'.
- Culturally-dependent terms that have no translation equivalent in the other language are tagged as shared; e.g., `mjadara` '*mujadara*', `alhamdulillah` '*thank God*', `ramadan` '*ramadan*', `muezzin` '*muezzin*'.
- This also extends to loan words that do not have translation equivalents in the borrowing languages e.g., `video` '*video*', or where the loan word is commonly used even if a translation exists; e.g., `taxi` '*taxi*', `mobile` '*cellphone*'.

To demonstrate the word-level annotation, consider the following examples:

- `Ask for Mjadara Hamra`

Here, the first two tokens are obviously English ('1'), while the third token is tagged '4' for shared. The fourth token, `Hamra` '*red*', raises a question: is it the adjective '*red*', in which case it should be tagged '0' for Arabizi, or is it part of a named entity that includes `Mjadara` '*mujadara*', in which case it should be '4' for shared? We opted for the former. In contrast, in

- `even the humble kibbe nayeh`

We tagged the first 3 tokens as English ('1'), and `kibbe nayeh` '*raw kibbe*', where `kibbe` is a popular dish consisting of meat and bulgur, but `nayeh` '*raw*' changes its meaning to a different dish made from raw meat, were both tagged '4' for shared as we considered them part of a single named entity.

A particularly interesting example is

- `Nis-har youm el sabt 3al Balcon`

which means '*We stay up Saturday night on the balcony*'. The verb `nishar` '*we spend the evening*' was probably spelled with a dash in order to prevent the '*sh*' from being pronounced as English [sh]. We tagged all tokens '0' for Arabizi, except the last one which was tagged '4' for shared.

Finally, some cases involved intra-word code-switching. In

- `ma2darsh a subtweet u da mabda2yan` '*I can't subtweet you, this is tentative*'

the English '*subtweet*' is used as a verb, with the Arabic prefix '*a*' which is a derivational morpheme that converts nouns to verbs; the result is `asubtweet` '*to subtweet*'. In this case, the author introduced a space between the two mporphemes so we could tag '*a*' as Arabizi and '*subtweet*' as English. In another example, `ana ba-act` '*I act*', the author used a dash between the Arabizi prefix '*ba*' and the English verb '*act*', so again we could tag both morphemes separately. We do not have a special tag for tokens that involve morphemes in more than one language because no such case was witnessed in our dataset.

4.2 Manual annotation

From the raw datasets we described in Section 3, we initially manually annotated 1050 sentences (roughly 500 each from Reddit and Twitter) at the word level, assigning a tag of '0' to '5' to each token.[2] We then used the classifier described below (Section 4.3) to identify more "interesting" samples in the entire dataset (the vast majority of the sentences in the dataset are naturally plain English sentences). Of those, we manually selected more sentences that reflected as best as possible the diversity of sentence types in the dataset, and manually corrected the predictions of the classifier. This process resulted in 2643 manually annotated sentences, over 1000 of which including Arabizi words, which constitute the final word-level annotated dataset on which we train and evaluate our classifiers. The details are summarized in Table 1 (note that not all sentences in a given post were annotated).

4.3 Statistic classification

We begin with more conservative statistic classification. Since the tag of a given token is highly

[2]Manual annotation was performed by the first author, who is a native speaker of Palestinian Arabic and fluent in English. The main challenge was the identification of shared words, which required discussion between the two authors, as well as with colleagues.

Dataset	Posts	Sents.	Tokens
Reddit	922	980	13752
Twitter	1653	1663	16061
Total	2575	2643	29813

Table 1: Word-level annotated dataset.

depepndent on the tags of its predecessors, we used CRF (Lafferty et al., 2001) to train a sequence-to-sequence classifier. We used the following features to represent each instance (token):

- The word itself in lowercase;
- Are all the word's letters uppercase?;
- Is only the first letter uppercase?;
- Is the word in the (freely-available list of) 5050 most frequent English words, taken from the one billion word Corpus of Contemporary American English?;
- Is the word in the 930 most frequent French words?;
- Is it an Arabic word? We used CAMel tools (Obeid et al., 2020) in order to detect Arabic words;
- Does the word contain numerals? This is useful because digits are used to represent Arabic letters in Arabizi;
- All the features above, with respect to the previous word;
- Is it the first word in the sentence?;
- Is it the last word in the sentence?

Here and elsewhere, we used ten-fold cross-validation for evaluation. Table 2 lists the evaluation results (precision, recall and F1) for each category separately, as well as the number of words of each category in the test set ("support"). It also shows the total evaluation metrics, averaged over all categories (we report micro-, macro- and weighted averages). The total accuracy, over the entire test set, is 0.949.

4.4 Neural classification

We also experimented with more contemporary neural classification. We defined a deep neural network consisting of three layers: (1) An embedding layer which is the concatenation of the last 4 layers of a BERT (Devlin et al., 2019) model (we used the multilingual uncased version); (2) A bidirectional LSTM (Hochreiter and Schmidhuber, 1997) layer: 2 hidden layers of size 400, and dropout of 0.5; (3) A CRF layer (Huang et al., 2015).

We used the BERT tokenizer, in the multilingual uncased version, to tokenize the text. As the tokenizer is different, the number of tokens differs slightly from the case of statistical classification (this explains the differences in the support size between Tables 2 and 3). More importantly, BERT's predictions are provided for units (sub-tokens) that we did not manually annotate. As is common in such cases, for each original token that was split by BERT we selected the tag of the first sub-token and induced it over the other sub-tokens to which the original token was split. Of course, this may harm the accuracy of the neural classifier.

We used the Adam optimizer with a learning rate of 0.001 and cross-entropy loss. We trained the model for four epochs and chose a batch size of 32. The results are listed in Table 3. The total accuracy, over the entire test set, is 0.952, almost identical to the accuracy of the statistic classifier.

Tag	Prc.	Rcl.	F1	Support
Arabizi	0.90	0.95	0.92	4865
English	0.96	0.98	0.97	16563
French	0.74	0.64	0.69	149
Arabic	0.99	0.99	0.99	2671
Shared	0.81	0.51	0.63	1401
Other	0.97	0.94	0.95	4164
Micro avg.	0.95	0.95	0.95	29813
Macro avg.	0.90	0.84	0.86	29813
Weighted avg.	0.95	0.95	0.95	29813

Table 2: Results: word-level statistic classification.

Tag	Prec.	Rcl.	F1	Support
Arabizi	0.91	0.95	0.93	4869
English	0.97	0.98	0.97	16938
French	0.56	0.43	0.49	167
Arabic	0.98	0.99	0.98	2680
Shared	0.77	0.66	0.71	1406
Other	0.97	0.94	0.95	4385
Micro avg.	0.95	0.95	0.95	30445
Macro avg.	0.86	0.82	0.84	30445
Weighted avg.	0.95	0.95	0.95	30445

Table 3: Results: word-level neural classification.

5 Identifying code-switching

The word-level annotation immediately facilitates the identification of code-switching: a sentence with at least one word in Arabizi and one in either English or French necessarily includes a switch. To

simplify this task, we now annotate full sentences: we assign complex tags to sentences that reflect the existence of each of our six word categories in a given sentence. The tags consist of six bits, each referring to the presence in the sentence of words categorized as Arabizi, English, French, Arabic, shared, and Other. This Table 4 lists the number of samples associated with each 6-bit tag in the annotated dataset.

For example, the sentence

- `good luck albi, have a nice dayy <3` '*good luck my love, have a nice day* ♡'

is associated with the tag 110001, reflecting the presence of English, Arabizi and an emoticon (note that we treat the misspelled '*dayy*' as a valid English word). More example sentences include:

- `"Khalas tamam , you know best"` '*Okay, you know best*' . (110000)
- `happiest birthday ya hussein :)` '*happiest birthday oh hussein :)*' (110011)
- `Take a flight to Jeddah w ishtiri al baik` '*Take a flight to Jeddah and buy the bike*' (110010, as '*Jeddah*' is shared)

Note that we do not commit on the precise location of the switch; when a sentence contains shared words, they may serve as wildcards for determining this location. For example, in the last sentence above, the switch may occur before or after the shared word '*Jeddah*'.

5.1 Direct classification

First, we trained a statistic classifier to directly predict the 6-bit tags. We experimented with various statistic classification models, including SVM, logistic regression, KNN, and random forest. The latter yielded the best accuracy, so the results we report below were obtained with random forest. We used the following features:

- Character uni-gram, bi-gram and tri-gram counts, normalized by the number of characters in the sentence. We only used the most frequent 250 n-grams;
- Number of English, Arabic and French words, all normalized by the number of tokens in the sentence (excluding emojis);
- The number of tokens that contain numeric digits, normalized by the number of tokens in the sentence;
- The normalized number of emojis, punctuation and numbers in the sentence, to help identify the category *Other*;

Arabizi	English	French	Arabic	Shared	Other	Occurrences
0	1	0	0	0	1	604
0	1	0	0	1	1	297
0	1	0	0	0	0	233
1	1	0	0	0	1	187
1	0	0	0	0	0	184
1	1	0	0	1	1	155
1	0	0	0	0	1	154
0	0	0	1	0	1	153
1	1	0	0	0	0	115
1	1	0	0	1	0	109
0	0	0	1	0	0	91
0	1	0	0	1	0	71
0	0	0	0	0	1	65
1	0	0	0	1	0	55
1	0	0	0	1	1	43
0	1	0	1	0	1	36
0	0	0	0	1	1	22
0	0	0	0	1	0	14
0	0	1	0	0	1	10
0	1	0	1	0	0	8
0	1	1	0	0	1	5
0	1	0	1	1	1	5
0	0	1	0	0	0	4
1	0	1	0	0	1	4
1	0	1	0	0	0	4
0	0	1	0	1	1	3
1	1	0	1	0	1	2
0	0	0	1	1	0	2
0	0	0	1	1	1	2
0	1	0	1	1	0	2
1	1	0	1	0	1	1
0	1	1	0	1	1	1
1	0	1	0	1	1	1
1	1	1	0	0	1	1
1	1	1	0	1	0	1
1	0	0	1	1	1	1
1	1	1	0	1	1	1
1	1	0	1	1	1	1

Table 4: Distribution of sentence-level tags in the annotated dataset.

- The number of English words detected by *fastText* with confidence score greater than 0.95;
- The number of French words detected by *fastText* with confidence score greater than 0.5;
- The number of words that do not belong to any of the previous categories, which helps detect *Arabizi* and *Other*;

- A binary flag which checks whether the whole sentence was detected by fastText as English with confidence score greater than 0.8. We observed that sentences with score greater than 0.8 tend to actually include English words, but pure Arabizi sentences are sometimes erroneously classified as English with lower confidence;
- A binary flag which checks whether the whole sentence was detected as French with confidence score greater than 0.3;
- A binary flag which checks whether the whole sentence was detected as some language other than French, English, or Arabic. This helps detecting *Arabizi* and other languages.

We used ten-fold cross-validation and evaluated the accuracy of the model in predicting each of the bits in the tag vector independently (i.e., predicting whether a given sentence includes words in English, Arabizi, French, etc.) The accuracy results on each category are listed in Table 5. The total accuracy of assigning the exact 6-bit tag to each sentence is 0.62.

Tag	Acc.	Prec.	Rcl.	F1
Arabizi	0.90	0.91	0.83	0.87
English	0.92	0.95	0.94	0.95
French	0.99	0.10	0.03	0.05
Arabic	1.00	1.00	1.00	1.00
Shared	0.75	0.67	0.34	0.45
Other	0.96	0.99	0.96	0.97

Table 5: Results: sentence-level direct classification.

5.2 Indirect classification

As an alternative to direct classification, it is possible to combine the predictions of the word-level classifiers (Section 4) and create 6-bit tags for each sentence. Recall that tags at the sentence level only indicate the existence of words from a given category in the sentence (rather than whether *all* words in the sentence are annotated correctly). The results of inducing sentence-level tags from the word-level ones (as obtained by the statistic classifier, Section 4.3) are listed in Table 6. The total accuracy of correctly identifying the complex, 6-bit tag is 0.78, much better than with the direct classifier.

Note that in both approaches, the identification of Arabic is perfect, most likely owing to the different character set of Arabic; and in both cases, shared words are the most challenging to identify (recall that they were also hard to manually annotated). The accuracy on French is low, probably because of the small number of sentences with French words in the training data.

Tag	Acc.	Prec.	Rcl.	F1
Arabizi	0.94	0.91	0.95	0.93
English	0.95	0.96	0.96	0.96
French	0.99	0.75	0.55	0.63
Arabic	1.00	1.00	1.00	1.00
Shared	0.86	0.89	0.62	0.73
Other	0.98	0.99	0.98	0.98

Table 6: Results: indirect sentence-level classification.

6 Harvesting more data

With the highly accurate classifiers described above, we set out to extend our corpus of Arabizi in general and Arabizi code-switching in particular. We applied the statistic word-level classifier (Section 4.3) to the entire dataset we collected from Reddit and Twitter (Section 3). We extracted all the sentences that included at least one Arabizi word, and associated each token in these sentences with its language ID tag; we also decorated the entire sentence with the complex 6-bit tag that indicates which languages are included in it. This resulted in a set of over 880K sentences, which constitutes our automatically-obtained dataset of Arabizi (see Table 7). This dataset, we trust, will be an invaluable resource for research in Arabizi and in code-switching.

	Reddit	Twitter	Total
With Arabizi	218619	668208	886827
Arabizi	67566	479317	546883
Ar–En CS	165982	277032	443014
Ar–Fr CS	1165	1913	3078

Table 7: The automatically-annotated dataset. Number of sentences with at least one Arabizi token (*With Arabizi*); with a majority of Arabizi tokens (*Arabizi*); and with code-switching between Arabizi and English (*Ar–En CS*) and between Arabizi and French (*Ar–Fr CS*).

As an additional verification of the dataset, we randomly chose 100 sentences (50 each from Reddit and Twitter) that were annotated as including at least two tokens each in both Arabizi and English (hence, that included code-switching) and manually

inspected them. Of the 100, 77 (42 from Twitter, 35 from Reddit) indeed included code-switching between English and Arabizi.

A qualitative analysis of the errors revealed several cases in which a nonstandard spelling of English was erroneously considered Arabizi. For example, in the fully English `wtf yo where da love go`, our classifier identified '*da*' as Arabizi, probably because it is a common Egyptian word meaning '*this*'. Similarly, in `I ' m sorry 4 ya loss` the classifier unsurprisingly identified '*ya*' as Arabizi.

Some proper nouns that we tagged as *shared*, especially those whose origin is Arabic, were predicted as Arabizi. E.g., in `They also mentioned a new location ; somewhere in sin el fil`, the last three tokens were predicted Arabizi, but we tagged them as shared (the name of a suburb of Beirut). Finally, tokens that involve both letters and digits were sometimes erroneously tagged as Arabizi (e.g., `I have the 20GB 2Mbps plan`).

7 Conclusion

We described a classifier that identifies words in Arabizi, English, Arabic, and French in multilingual sentences from social media. We applied the classifier to a large set of sentences collected from Twitter and Reddit, and produced a huge dataset of more than 880K automatically-annotated Arabizi sentences, of which over 446K include code-switching with either English or French.

We are now ready to use this dataset for a large-scale corpus-based investigation of theoretical research questions in cognitive linguistics. Specifically, we are interested in the correlation between shared words, as defined in our annotation scheme, and code-switching. We leave such investigations for future work.

8 Ethical considerations and limitations

This research was approved by the University of Haifa IRB. We collected data from two social media outlets, Reddit and Twitter, in compliance with their terms of service (Reddit, Twitter). For the latter, we distribute tweet IDs and sentence IDs instead of the actual sentences, in line with Twitter's terms of use. For anonymity, we systematically replaced all user IDs (in both datasets) by unique IDs; we do not have, and therefore do not distribute, any personal information of the authors. With this additional level of anonymization, we anticipate very minimal risk of abuse or dual use of the data.

Like any other dataset, the corpus we report on here is not representative. In particular, it probably includes Arabizi as used mainly in Egypt and in Lebanon but not elsewhere in the Arab-speaking world. It is very likely unbalanced in terms of any demographic aspect of its authors. Clearly, the automatic annotation of language IDs is not perfect, and may introduce noise. Use of this corpus for linguistic research must therefore be done with caution. Nevertheless, we trust that the sheer size of the dataset would make it instrumental for research on code-switching in general and in Arabizi in particular.

Acknowledgements

We thank Melinda Fricke, Yulia Tsvetkov, Yuli Zeira, and the anonymous reviewers for their valuable feedback and suggestions. This work was supported in part by grant No. 2019785 from the United States-Israel Binational Science Foundation (BSF), and by grants No. 2007960, 2007656, 2125201 and 2040926 from the United States National Science Foundation (NSF).

A Lists of seed words

We collected data from Reddit and Twitter based on texts that included the following words.

Lebanese `bya3ref` '*he knows*', `ma3leh` '*never mind*', `be7ke` '*to say*', `halla2` '*now*', `ma32ool` '*reasonable*', `3shen` '*in order to*', `3am` (present tense particle) `mazboot` '*alright*' `kteer` '*many/much*' `3lay/3layki` '*on me/on you*$_{fem}$'.

Egyptian `awy` '*very/very much*', `kwayes` '*OK*', `ezai` '*how*', `5ales` '*never*', `7a2ee2y` '*really*', `m3lesh` '*never mind*', `howa-eh` '*what*'.

Interestingly, the word `mazboot` '*alright*' means '*strong*' in Hindi, so it yielded many false positives. However, since it also resulted in having many relevant Lebanese tweets, we manually scanned them and removed irrelevant users. Similarly, the world `awy` '*very*' is highly indicative of the Egyptian dialect, but it is also used as an abbreviation of the English word '*away*'. Attempting to use the seed words `baddi` '*I want*' and `balki` '*maybe*', both highly widespread in Lebanon, resulted in harvesting many irrelevant texts; upon inspection we revealed that these words are frequent proper names in India. They were therefore removed from the seed word list.

References

Wafia Adouane, Nasredine Semmar, and Richard Johansson. 2016. Romanized Berber and Romanized Arabic automatic language identification using machine learning. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 53–61, Osaka, Japan. The COLING 2016 Organizing Committee.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the CALCS 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147.

Mohamed Al-Badrashiny, Ramy Eskander, Nizar Habash, and Owen Rambow. 2014. Automatic transliteration of Romanized dialectal Arabic. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 30–38, Ann Arbor, Michigan. Association for Computational Linguistics.

Gaétan Baert, Souhir Gahbiche, Guillaume Gadek, and Alexandre Pauchet. 2020. Arabizi language models for sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 592–603, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Sebastopol, CA.

Mirjam Broersma. 2009. Triggered codeswitching between cognate languages. *Bilingualism: Language and Cognition*, 12(4):447–462.

Mirjam Broersma and Kees De Bot. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and cognition*, 9(1):1–13.

Michael G. Clyne. 2003. *Dynamics of language contact: English and immigrant languages*. Cambridge approaches to language contact. Cambridge University Press, Cambridge.

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An Algerian Arabic-French code-switched corpus. In *Proceedings of the First Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*. Association for Computational Linguistics.

Kareem Darwish. 2014. Arabizi detection and conversion to Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 217–224, Doha, Qatar. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1654–1666. Association for Computational Linguistics.

François Grosjean. 1998. Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, 1(2):131 – 149.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-01)*, pages 282–289, San Francisco. Morgan Kaufmann.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Marco Lui and Timothy Baldwin. 2012. Langid.Py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, page 25–30, USA. Association for Computational Linguistics.

Abir Masmoudi, Nizar Habash, Mariem Ellouze, Yannick Estève, and Lamia Hadrich Belguith. 2015. Arabic transliteration of Romanized Tunisian dialect text: A preliminary investigation. In *Computational Linguistics and Intelligent Text Processing*, pages 608–619, Cham. Springer International Publishing.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Ella Rabinovich, Masih Sultani, and Suzanne Stevenson. 2019. CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4776–4786, Hong Kong, China. Association for Computational Linguistics.

Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Transactions of the Association for Computational Linguistics*, 6:329–342.

Younes Samih and Wolfgang Maier. 2016. An Arabic-Moroccan Darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4170–4175, Portorož, Slovenia. European Language Resources Association (ELRA).

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Javier Ortiz Suárez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1139–1150, Online. Association for Computational Linguistics.

Ali Shazal, Aiza Usman, and Nizar Habash. 2020. A unified model for Arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 167–177, Barcelona, Spain (Online). Association for Computational Linguistics.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. 2019. A survey of code-switched speech and language processing.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Julia AlGhamdi, Fahadand Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors. 2021. *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Online.

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

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, POS tags and entrainment in code-switching. In *Proceedings of Interspeech 2018, the 19th Annual Conference of the International Speech Communication Association*, pages 1938–1942. ISCA.

Victor Soto and Julia Hirschberg. 2019. Improving code-switched language modeling performance using cognate features. In *Proceedings of Interspeech 2019, the 20th Annual Conference of the International Speech Communication Association*, pages 3725–3729. ISCA.

Taha Tobaili. 2016. Arabizi identification in Twitter data. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 51–57, Berlin, Germany. Association for Computational Linguistics.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online. Association for Computational Linguistics.

Jihene Younes, Hadhemi Achour, Emna Souissi, and Ahmed Ferchichi. 2020. A deep learning approach for the Romanized Tunisian dialect identification. *The International Arab Journal of Information Technology*, 17(6):935–946.

Jihene Younes, Hadhemi Achour, Emna Souissi, and Ahmed Ferchichi. 2022. Romanized Tunisian dialect transliteration using sequence labelling techniques. *J. King Saud Univ. Comput. Inf. Sci.*, 34(3):982–992.