An Algerian Arabic-French Code-Switched Corpus

Ryan Cotterell[1], **Adithya Renduchintala**[1], **Naomi Saphra**[1], **Chris Callison-Burch**[2]

[1] Center for Language and Speech Processing, Johns Hopkins University
[2] Computer and Information Science Department, University of Pennsylvania

Abstract

Arabic is not just one language, but rather a collection of dialects in addition to Modern Standard Arabic (MSA). While MSA is used in formal situations, dialects are the language of every day life. Until recently, there was very little dialectal Arabic in written form. With the advent of social-media, however, the landscape has changed. We provide the first romanized code-switched Algerian Arabic-French corpus annotated for word-level language id. We review the history and sociological factors that make the linguistic situation in Algerian unique and highlight the value of this corpus to the natural language processing and linguistics communities. To build this corpus, we crawled an Algerian newspaper and extracted the comments from the news story. We discuss the informal nature of the language in the corpus and the challenges it will present. Additionally, we provide a preliminary analysis of the corpus. We then discuss some potential uses of our corpus of interest to the computational linguistics community.

Keywords: code-switching, Algerian Arabic, romanized Arabic, French

1. Introduction

Language identification systems have long operated under the assumption that text is written in a single language. As social media becomes a more prominent mode of communication, such systems are confronting text that increasingly challenges the monolingual assumption. More than half the world's population is bilingual and information communication is often code-switched, reflecting the need for a deeper understanding of code-switching in relation to NLP tasks. Recent work has proposed both supervised and unsupervised methods for *word*-level language id. Current methods, however, rely on the assumption that external resources exist, such as large non-code-switched corpora and dictionaries. These resources are not available for some languages and dialects, including Algerian vernacular Arabic, a dialect often code-switched with French. We are releasing a corpus of romanized Algerian Arabic and French scraped from the comments sections of Echorouk, an Algerian daily newspaper with the second-largest reader base of any Arabic paper. This is the only significant corpus of romanized Arabic known to the authors and additionally it is the largest corpus of code-switched data to our knowledge. Such a resource is necessary because romanized Arabic is becoming increasingly popular on the internet. The corpus will also be of use for the linguistic study of code-switching. Much of previous code-switching research has focused on data collected from field work, and a found dataset like ours could provide an interesting perspective on the use of code-switching in conversation.

2. Code-switching

Code-switching is a linguistic phenomenon wherein speakers switch between two or more languages in conversation, often within a single utterance (Bullock and Toribio, 2009). It can be viewed through a sociolinguistic lens where situation and topic influence the choice of language (Kachru, 1977). To define code-switching as a phenomenon, it is important to make the distinction between *code-switching* and *borrowing*. *Borrowing* is the act of using a foreign word without recourse to syntactic or morphological properties of that language and often occurs with phonological assimilation. *Code-switching*, on the other hand, involves switching between languages in which the speakers are fluent, and can in effect be viewed as changing the grammar in use. Some linguists have even proposed a scale of code-switching, positing the existence of a continuum between *borrowing* and *code switching* (Auer, 1999). Code-switching points (the times at which speakers change language) and the context around which switching occurs are also of interest to linguists. These points often lie within a sentence and their position is influenced by the syntax of the respective languages. Poplack (Poplack, 1988) posited that code-switching points cannot occur within a constituent. Recent work, however, has found that many speakers relax this constraint. The Matrix Language-Frame (MLF) model is one theory has gained traction (Myers-Scotton, 1993) to explain code-switching patterns. MLF proposes that there is a Matrix Language (ML) and an Embedded Language (EL). The ML is the more dominant language and is often the language which the speaker identifies as their native tongue. The EL is then inserted into the ML at certain grammatical frames. Within this framework, further work has gone into the exact syntactic and morphological contexts that allow for code-switching points (Myers-Scotton and Bolonyai, 2001).

3. Code-Switching in North Africa

Code-switching in North African Arabic is an established phenomenon that has been studied by the linguistics community (Bentahila and Davies, 1983). It dates back to the initial French colonization of North Africa. North Africa is also home to many cultures, a fact which potentially affects language use and code switching in particular. Until recently, mixed language communication has been observed mainly as a spoken phenomenon. With the widespread use of computer-mediated communication, code-switching is becoming common in North African Arabic (Salia, 2011).

Comments on news-feeds and social media outlets like Twitter and Facebook often contain code-switching. North African Arabic is not the only language to appear in code-switching writing. A body of recent sociolinguistics work has considered the phenomenon in various settings. Swiss-German and German code-switching in chat rooms was analyzed in Siebenhaar (2006) and Callahan (2004) considered Spanish and English code-switching in a written corpus.

4. Related Work

From a computational perspective, code-switching has received relatively little attention. Joshi (1982) provides a tool for parsing mixed sentences. More recently, Rosner and Farrugia (2007) focused on processing code-switched SMS. Solorio and Liu (2008) trained classifiers to predict code-switching points in Spanish and English. Nguyen and Dogruoz (2013) also focused on word level language identification in Dutch-Turkish news commentary. To our knowledge, Elfardy and Diab (2012) is the only computational work on Arabic code-switching done to date. That work does not include romanized Arabic. Many languages written in a non-Roman script are *romanized* on the internet. This practice presents a problem for standard NLP tools that are trained on the language with its standard orthography (Irvine et al., 2012). We believe that this type of romanized data will become more pervasive as more users employ computer-mediated communication globally. More information will be generated in such settings, and it is critical for future NLP systems to be able to process the data produced. The corpus we are presenting is a step in this direction.

5. Data Collection

We used a corpus crawled from an Algerian newspaper website. We scraped 598,047 pages in September 2012. These fora are rich in both dialectal Arabic and French content. The corpus contains discussion on a wide-ranging set of issues including domestic politics, international relations, religion, and sporting events. We extracted 6,949 comments, containing 150,000 words in total. We separated the comments section from the main article on each page and stripped HTML tags and other non-user generated content. The metadata was stripped in an attempt to preserve anonymity. The Arabic portion of the corpus was annotated for sentence level dialect on Mechanical Turk (Cotterell and Callison-Burch, 2014).

We separated all the comments, in which more than half the non-white space characters were in the Roman alphabet, determining these to be romanized. We did no further processing, e.g. tokenization. The final data set contains 339,504 comments with an average length of 19 tokens, as determined by separating on white space and punctuation. 1,000 of the comments are annotated using the guidelines described below. Our corpus has 493,038 types and 6,718,502 tokens, and is formatted in JSON.

6. Romanization of Arabic

This corpus is unique in that it is the first large corpus to the authors' knowledge that is composed of of an Arabic dialect written in romanized form. Romanized Arabic is particularly difficult because there is no standard form of romanization used across the Arab world. In order to use standard NLP tools on such corpora, it is often necessary to *deromanize* the corpus. In the case of Urdu, this task has been successfully completed using standard Machine Translation software (Irvine et al., 2012).

Arabic	Arabizi	Arabic	Arabizi
ا	a	ب	b, p
ت	t	ث	th, s
ج	j, g	ح	7, h
خ	7', 5	د	d
ذ	th z	ر	r
ز	z	س	s c
ش	sh, ch	ص	9
ض	9', d	ط	t
ظ	th	ع	3
غ	gh, 3'	ف	f
ق	8, 2, k, q	ك	k
ل	l	م	m
ن	n	ه	h
و	w, o, ou	ي	y, i, e

Figure 1: Correspondence Between Arabic Letters and Romanized Arabic (Yaghan, 2008)

Arabic written in the Latin alphabet, often dubbed *arabizi*, is extremely common on the internet and SMS. The exact mapping from the Arabic script onto the Latin alphabet varies significantly between regions. The specific case of romanization by young speakers of Gulf Arabic in the United Arab Emirates is discussed thoroughly in Palfreyman and Khalil (2003).

Figure 1 expresses the most common mappings across the Arab world. Algerians, and North Africans in general, tend to use romanizations that reflect French orthography: for instance و ↦ ou, ش ↦ ch, ج ↦ dj and ا ↦ è or é. To illustrate this difference consider the frequency of the common transcriptions of إن شاء الله (God willing); we see RL ↦ ch about an order of magnitude more often than ش ↦ sh.

This transcription variation makes it unlikely that a single, general-purpose Arabic deromanization tool will be enough, and such romanized corpora will need to be developed for other dialects as well in order to analyze the users romanization preferences on dialectal basis.

7. Text Analysis

Because our Algerian corpus is from the length-constrained informal domain of online forum comments, it would be difficult to process meaningfully without normalizing beforehand. It exhibits extreme variation in spelling and grammar. Many forms of the same word may appear throughout our corpus. For example, we identified 69 variants of the common word إن شاء الله alone. 70% of all token types in our corpus appeared only once, so the OOV rate in this forum corpus can confound language processing systems without text normalization. Several typical sources of variation for Arabic identified by (Darwish et al., 2012) were found in our corpus.

- The use of elongations, especially in the form of vowel repetition.

  > we ki ta3arfou wach rah testfadou ? **hhhhh** cha3ab **kar3adjiiiiiiiii**

- Spelling mistakes, such as dropped or transposed characters.
- Abbreviations.

  > rah thablona bel **BAC** had al3am !!!!!

- Emotional tokens and ejaculative abbreviations, such as the abbreviation "lol" borrowed from English web speech, or emoticons.

In addition to these irregularities, our corpus contains variations particular to romanized Arabic text because there is no standardized way to transcribe Arabic orthography in this informal domain, Arabic words can be represented by multiple spellings.

8. Example Posts

We present below a few example sentences that we have collected use the data collection methodology described above.

- bezaf m3a saifi oalah mnkalifoha mairbahch
  Had enough with Saifi.
- la howla wa la kowata il bi lah el3alier l3adim wa la yassa3oni an akoul anaho kllo chaye momkin ma3a ljazairyine
  For Gods Sake! I can just say that anything is possible with Algerians .
- we ki ta3arfou wach rah testfadou ? hhhhh cha3ab kar3adjiiiiiiiii
  Don't try to know everything because it does not matter to you.
- 7ade said nchalah li lmontakhabina el3assekari nchalah yjibo natija mli7a bitawfiiiiiiiiiiiiiiik . . . onchoriya chorouk stp
  Good luck to our military team, I hope they get a good score. Good Luck! Say it chorouk!
- ya khawti tt simplement c bajiou l3arab o makanech fi tzayer kamla joueur kima lhadji et on vai ras le 27 03 2011 chkoun houma rjal liyestahlouha
  My brothers he's simply the Baggio of Arabs, and there is none like Lhadji in Algeria, and on 27th of march 2011 we will see who wins.
- mais les filles ta3na ysedkou n'import quoi ana hada face book jamais cheftou khlah kalbi
  Our girls believe anything, I have never seen this Facebook before.

9. Annotation Guidelines

Annotating for word level language identification in code-switched text is a difficult task because whether a word is code-switched is often more of a continuum than a binary decision. Place names form a simple example: باريس (Paris) is an MSA word in that it is found in most Arabic dictionaries, but it is clearly of French origin. On the other hand, فيديو (video) is an example of a recent borrowing from European languages that should be considered an Arabic word. To simplify the decision, we made use of guidelines for dialectal Arabic annotation provided in (Elfardy and Diab, 2012). Their guidelines were created for annotating world level language id in a corpus composed of mixed dialectal Arabic and MSA both written in the Arabic script. As we annotating two linguistically dissimilar languages, the same level of ambiguity does not arise.

Further research in area of code-switching should focus on richer annotation schemata that are both linguistically motivated, i.e. taking into account the continuum of code-switching, and serve practical NLP needs. Another interesting area to focus on could be to annotate broad categories describing the type of code-switching. Kecskes (2006) describes three prominent patterns in code-switching (Insertion, Alternation and Congruent Lexicalization) based on Gibraltar data. Insertion involves adding lexical items from one language into the structure of the other. Alternation is similar to insertion except that larger chunks are inserted, rather than single tokens. Congruent Lexicalization is adding lexical items from different lexical inventories into a common grammar structure. The dataset could annotate each point of code-switching with these patterns of code-switching as well. This data set, however, is focused on only the points of code-switching.

The annotators were presented with posts and asked to label each word, split on white space and punctuation. They were given the choice of Arabic (A), French (F) and Other (O). We excluded punctuation from the annotation. Figure 4 shows the distribution of these tags in our dataset. The annotation was conducted with an interactive Python script.

10. Potential Uses

The corpus provided is the first of its kind in that it is the first large corpus of romanized Arabic that is code-switched with another language. This has numerous potential uses in both the NLP community and the linguistics community. In the NLP community, the processing of informal text is becoming and increasingly popular task among researchers (Yang and Eisenstein, 2013). This corpus adds another complication to informal text processing with the addition of code-switching. In the linguistics community, a corpus based analysis of a code-switched corpus offers the possibility to test various hypotheses on large number of documents. The MLF hypothesis has already been studied in bilingual speech corpora from Miami, Patagonia, and Wales (Carter, 2010), but it will be necessary to study such theories in the context of many distinct languages and cultures to gain deeper insight into the code-switching phenomenon.

11. Conclusion and Future Work

The primary contribution of this paper is the release of a Algerian Arabic-French code-switched corpus. We have made use of a previously proposed annotation scheme for word level language identification and highlighted the unusual qualities of this corpus that make it a significant contribution to field. Future work in this line should largely focus on experiments using the corpus both in NLP and linguistics. It would also be of interest to construct and annotate similar corpora for other informal code-switched Arabic dialects.

Acknowledgements

This material is based on research sponsored by a DARPA Computer Science Study Panel phase 3 award entitled "Crowdsourcing Translation" (contract D12PC00368). The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements by DARPA or the U.S. Government. This research was supported by the Johns Hopkins University Human Language Technology Center of Excellence.

12. References

Peter Auer. 1999. From codeswitching via language mixing to fused lects toward a dynamic typology of bilingual speech. *International Journal of Bilingualism*, 3(4):309–332.

Abdelali Bentahila and Eirlys E Davies. 1983. The syntax of Arabic-French code-switching. *Lingua*, 59(4):301–330.

Barbara E Bullock and Almeida Jacqueline Toribio. 2009. *The Cambridge handbook of linguistic code-switching*. Cambridge University Press Cambridge.

Laura Callahan. 2004. *Spanish/English codeswitching in a written corpus*, volume 27. John Benjamins Publishing.

Ryan Cotterell and Chris Callison-Burch. 2014. The extended Arabic online commentary dataset. In *LREC*.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for Arabic microblog retrieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2427–2430, New York, NY, USA. ACM.

Heba Elfardy and Mona T Diab. 2012. Token level identification of linguistic code switching. In *COLING (Posters)*, pages 287–296.

Ann Irvine, Jonathan Weese, and Chris Callison-Burch. 2012. Processing informal, Romanized Pakistani text messages. In *Proceedings of the Second Workshop on Language in Social Media*, pages 75–78. Association for Computational Linguistics.

Aravind K Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th conference on Computational linguistics-Volume 1*, pages 145–150. Academia Praha.

Braj B Kachru. 1977. Linguistic schizophrenia and language ensus: A note on the Indian situation. *Linguistics*, 15(186):17–32.

I Kecskes. 2006. A dual language model to explain code-switching: A cognitive-pragmatic approach. *Intercultural Pragmatics*, 3:257–283.

Carol Myers-Scotton and Agnes Bolonyai. 2001. Calculating speakers: Codeswitching in a rational choice model. *Language in Society*, 30(1):1–28.

Carol Myers-Scotton. 1993. Common and uncommon ground: Social and structural factors in codeswitching. *Language in Society*, 22:475–475.

Dong-Phuong Nguyen and AS Dogruoz. 2013. Word level language identification in online multilingual communication. Association for Computational Linguistics.

David Palfreyman and Muhamed al Khalil. 2003. a funky language for teenzz to use: representing gulf Arabic in instant messaging. *Journal of Computer-Mediated Communication*, 9(1):0–0.

Shana Poplack. 1988. Contrasting patterns of code-switching in two communities. *Codeswitching: Anthropological and Sociolinguistic Perspectives. New York: Mouton de Gruyter*, pages 215–244.

Mike Rosner and Paulseph-John Farrugia. 2007. A tagging algorithm for mixed language identification in a noisy domain. In *INTERSPEECH*, pages 190–193.

R. Salia. 2011. *Between Arabic and French Lies the Dialect: Moroccan Code-Weaving on Facebook*. Undergraduate thesis, Columbia University.

Beat Siebenhaar. 2006. Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics*, 10(4):481–506.

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

Mohammad Ali Yaghan. 2008. Arabizi: A contemporary style of Arabic slang. *Design Issues*, 24(2):39–52.

Yi Yang and Jacob Eisenstein. 2013. A log-linear model for unsupervised text normalization. In *Proc. of EMNLP*.