Simplified guidelines for the creation of Large Scale Dialectal Arabic Annotations

Heba Elfardy and Mona Diab

Center for Computational Learning Systems
Columbia University, NY, NY 10115
{heba, mdiab}@ccls.columbia.edu

Abstract

The Arabic language is a collection of dialectal variants along with the standard form, Modern Standard Arabic (MSA). MSA is used in official Settings while the dialectal variants (DA) correspond to the native tongue of the Arabic speakers. Arabic speakers typically code switch between DA and MSA, which is reflected extensively in written online social media. Automatic processing such Arabic genre is very difficult for automated NLP tools since the linguistic difference between MSA and DA is quite profound. However, no annotated resources exist for marking the regions of such switches in the utterance. In this paper, we present a simplified Set of guidelines for detecting code switching in Arabic on the word/token level. We use these guidelines in annotating a corpus that is rich in DA with frequent code switching to MSA. We present both a quantitative and qualitative analysis of the annotations.

Keywords: Linguistic Code Switching, Dialectal Arabic, Annotation Guidelines

1. Introduction

Most languages in the world exist in some standard form and are associated with some regional varieties. Some languages exist in a state of diglossia. Diglossia (Ferguson, 1959) refers to the situation when two varieties of the same language live side-by-side and are closely related. This is the case of the Arabic language where the standard form, Modern Standard Arabic (MSA), the language used in education, scripted speech and official Settings, co-exists with the dialectal variants (DA) which is the native tongue of Arabic speakers. Even though these variants have no standard orthography, they are used in unofficial written communication and are increasingly being seen in web-forums and blogs since the language used in such forums and blogs is closer to the natural spoken language than the formal written form. Arabic dialects may be divided into five main groups: Egyptian (including Libyan and Sudanese), Levantine (including Lebanese, Syrian, Palestinian and Jordanian), Gulf, Iraqi and Moroccan. Sub-dialectal variants also exist within each dialect (Habash, 2010). Speakers of a specific Arabic Dialect normally code switch between their dialect and MSA and less frequently between different dialects. Code switching in Arabic happens both intra-sententially and inter-sententially. In fact we have evidence of code switching occurring morphemically, i.e. the stem of a word maybe dialectal and end with an MSA morpheme and vice versa. Code switching poses significant challenges to automatic processing tools and applications such as automatic translation, speech recognition, distillation, etc. The challenge is amplified with the absence of annotated resources for the different Arabic dialects. The creation of large-scale annotated corpora for code switch points can help us gain insights into the patterns of code switching and also will help with creating more robust NLP applications and tools.

In this paper, we view the problem of code switch point detection as a dialect identification task. We present a Set of simplified guidelines for the annotation of large corpora of mixed Arabic resources (DA and MSA). The current guidelines are inspired by the guidelines presented in (Habash et al., 2008), however different from the Habash et al. guidelines that focused mainly on identifying the level of dialectalness for each word, our guidelines are more coarse grained and simplified. We are specifically interested in identifying whether each word belongs to DA or MSA or is shared between the two varieties. We use these guidelines as a pilot for annotating a large scale corpus on the word/token level for both Levantine and Egyptian corpora, respectively.

2. Annotation Guidelines

We create the guidelines geared toward the annotation of a specific genre of online communication as exhibited in discussion fora. However, by design, these guidelines are general enough to be applicable to other data genres. The data comprises threads on specific topics. Each thread contains several posts by participants. Our unit of annotation is the token, however, the annotators have to consider the context of the entire post and utterance surrounding the token. The first step in the annotation process is to read the whole post. This is crucial in order to get enough context to judge the id of each word in context, not in isolation. Figure 2. shows an example of a post. We keep the handle (name/alias) of the author typically associated with these posts available for the annotator. Author information is helpful for two reasons:
(1) It can be an indicator of the dialect spoken by the poster and accordingly the dialect of the post.

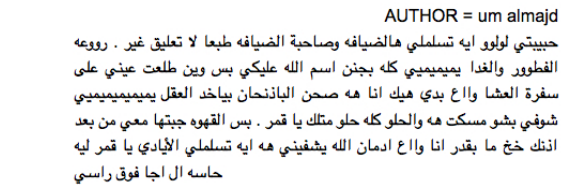
AUTHOR = um almajd
حبيبتي لولوو ايه تسلملي هالضيافه وصاحبة الضيافه طبعا لا تعليق غير . رووعه
الفطوور والغدا يميميي كله بجنن اسم الله عليكي بس وين طلعت عيني على
سفرة العشا واع بدي هيك انا هه صحن الباذنجان بياخد العقل يميميميميي
شوفي بشو مسكت هه والحلو كله حلو متلك يا قمر . بس القهوه جبتها معي من بعد
اذنك خخ ما بقدر انا واع ادمان الله يشفيني هه ايه تسلملي الأيادي يا قمر ليه
حاسه ال اجا فوق راسي

Figure 1: Example of a post

Ex. لؤلؤة الخليج *l&l&p Alxlyj*[1] which means **"The pearl of the Gulf"** is a good indicator that the post mainly belongs to Gulf Arabic.

(2) It helps the annotators identify the handles of the participants in the thread thereby, they would recognize the usage of handles as named entities when included in the body of the post. This is quite relevant in the context of the genre of data we are targeting since people get very creative with their handles and are quite often adjectival, leading to significant confusion with regular/typical nominals.

Annotators are asked to choose one of the following classes for each word; (1) MSA, (2) DA, (3) Both MSA & Dialectal, (4) Named Entity, (5) Foreign, (6) Typo, and (7) Unknown. Judging the class of each word includes a combination of context dependent and context independent considerations. While phonology could be an important criterion for judging the class of the word, we decided not to incorporate it in the decision process whenever it is not reflected in the orthography or meaning of the word. For example هنا, *hnA*, **"here"**, is pronounced in MSA هُنا *hunA*, and pronounced in DA هِنا *hina*, however if the text is undiacritized then we would consider both forms as the same word since they have the same meaning and are spelled the same orthographically when they appear in naturally occurring undiacritized text, despite the fact that they are pronounced differently depending on whether it is in a DA or MSA context. On the other hand, كتب *ktb*, **"to write"**, if pronounced كَتَب *kataba*, or it can mean **"was written"** indicating passive voice, if it is pronounced كُتِبَ *kutib*. Accordingly, we treat each case differently even if the text is undiacritized since the voice in each case is different reflecting the passive inflection. Below is a detailed description of the different classes:

A. **MSA [M]**;
A word is considered MSA if it is used ONLY in MSA and never/rarely used in DA.

[1]We use Buckwalter transliteration scheme http://www.qamus.org/transliteration.htm

Ex. 1. منذ *mn** **"since"**, مياه *myAh* **"water"**, عندما *EndmA* **"when"**

B. **Dialectal Arabic (DA) [D]**;
A word is considered a dialectal word if it is:

- **A Dialect Lexeme (Lexical entry)**;
A totally dialectal word (it will not be found in an MSA context)

  Ex. 1. شلونك *$lwnk* **"how are you (in Gulf)"**, مش *m$* **"not"**, هادا *hAdA* **"this"**, and ازيك *Azyk* **"How are you (in Egyptian)"**.

  NOTE: It is worth noting the following considerations:
  – The judgment in this case is context independent;
  – In case of totally dialectal words that appear with MSA inflectional morphology (affixes or clitics), it still should be considered dialectal and assigned a D tag: lexically, it is a dialectal word.

  Ex. 2. DA Lexeme with MSA 3MP *wn* inflectional morphology: يتشقلبون *yt$qlbwn*, **"they tumble"**

  Ex. 3. EGY dialect, the *hn* is marked as an MSA morpheme: بلوزتهن *blwzthn*, **"their shirt"**

  Ex. 4. DA Lexeme with the MSA future marker clitic *s*: سيزعل *syzEl* **"will become upSet"**

- **MSA semantic cognate (synonym) with a consistent systematic dialectal phonological variation**;
The ث *v* letter is consistently rendered as a ت *t* or س *s*, the ذ * is consistently rendered as a د *d*, etc. Ex. 1. تلاته *tlAth*, **'three'"** instead of ثلاثة *vlAvp*, كتير *ktyr*, **"much/a lot"** instead of كثير *kvyr*, ضابط *DAbT*, **"policeman"** instead of ظابط *ZAbT*, ضهر *Dhr*, **"back"** instead of ظهر *Zhr*, دهب *dhb*, **"gold"** instead of ذهب **hb*, دكر *dkr*, **"male"** instead of ذكر **kr*, and ألم *>lm*, **"pen"** instead of قلم *qlm*.

- **Faux amis: An MSA undiacritized homograph but contextually semantically DA with a different meaning from MSA**;
These words are undiacritized homographs in DA and MSA, however given the context they have different meanings in MSA and DA. Accordingly, in a DA context they are judged as DA, i.e. faux-amis

  Ex. 1. عم *Em* in MSA means **"uncle"** while in DA is used as a progressive particle, جدا *jdA* in MSA means **"grandfather"** while in DA means

"much", and يعني *yEny* in MSA means **"to mean"** while in DA means **"to some extent"** and is used a discourse/pragmatic marker.

Ex. 2. كتب كتابه يوم الجمعة اللي فاتت *ktb ktAbh ywm AljmEp Ally fAtt* can mean **"He got married last Friday"** in DA and **"He wrote his book last Friday"** in MSA. In this context, the intended meaning is *"He got married"* since we have "last Friday" indicating a short term event and also *AljmEp Ally fAtt* is a DA context namely using the Egyptian DA relative particle *Ally* "That/which".

Ex. 3. عربية *Erbyp* **"car or Arab "** لو اشتريت عربية نصف نقل اشغلها فين وازاى وعند مين ؟ *lw A$tryt Erbyp nSf nql A$glhA fyn wAzAY wEnd myn?* meaning **"If I bought a truck where and how and with whom can I use it?"**
أطلقت جوجل ١١ صفحة لخرائط ١١ دولة عربية *>Tlqt jwjl 11 SfHp lxrA}T 11 dwlp Erbyp* meaning **"Google launched 11 pages for the maps of 11 Arab countries."**
In the first sentence it means **"car"** while in the second sentence it means **"Arab**.

Ex. 4. If it exists as a part of a DA Multi-Word-Expression such as كبر دماغك *kbr dmAgk*, **"forget about it"**, individually each of the words in this expression is both MSA and DA but their meaning as an MWE, is non compositional and more frequent in the dialect (in fact is not an MSA MWE).

- **MSA and DA semantic cognate and undiacritized homograph but with DA clitics or inflectional morphology**;
  The word/stem is deemed DA and it has a semantic orthographic cognate – with the same meaning, not necessarily a homophone, especially with vowels – in MSA but there is some added DA suffix/prefix clitics:
  ex: حيكتب، ديكتب *dyktb, Hyktb* **"will write"**, بيكتب *byktb* **"is writing"**, and هالسؤال *hAls&Al* **"this question"**

- **Dialect lexeme with MSA morphology**;
  The word/stem is DA but it occurs with some MSA suffix/prefix/clitics morphology:
  Ex. سيزعل *syzEl*, **"will be upSet"**. The word يزعل *yzEl* which means **"be upSet"** is DA and the prefix س *"s"* meaning **"will"** is MSA.

C. **Both MSA and DA [B]**;
If the word is used in both MSA and Dialectal Arabic with the same meaning and same orthography/spelling – a synonym and undiacritized homograph. In this case, the context does not need to be considered and even if the phonology is different, as long as the word satisfies the orthographic and meaning conditions, it should be judged as B.
NOTE: It is worth noting that if such words occur with MSA inflectional morphology or clitics, then they should be judged as MSA or if they occur with DA inflectional morphology or clitics then they should be deemed DA.
Ex. 1. حساب *HsAb* **"account"**, ، مقيم *mqym* **"resident"** and سريع *sryE* **"fast"**

D. **Foreign Words [F]**;
These are words that are not part of the Arabic Language whether written in Arabic or Latin Script. It's worth mentioning that Foreign named Entities don't belong to this class but rather the next one **"Named-Entities"**
Ex. 1. كاتشب *kAt$b* **ketchup**, and كانيلوني *kAnylwny* **cannelloni**
NOTE: We make a distinction between borrowings and nonces (words that have become part of the language such as **"merci"** and **"gateau"**, for all intents and purposes these words are considered DA and should be judged as such, we provide a list of such words to the annotators.[2]

E. **Named Entities [N]**;
Names of people, organizations, companies, countries, titles and handles. The named entities could be Arabic or foreign.
A rule of thumb is that if a word is rendered in English with a word initial capital letter then it is a named entity.
Ex. 1. د. محمد البرادعي *d. mHmd AlbrAdEy* **"Dr. Mohamed ElBaradei"**, and وايت هاوس *wAyt hAws* **"White House"**

F. **Typos [T]**;
These are words that are misspelled either because:
(1) a letter is used instead of another letter
Ex. 1. عاير *EAyr* as opposed to عايز *EAyz* which means **"I want"** and استحدم *AstHdm* as opposed to استخدم *Astxdm* which means **"uses"**
(2) if a word is split into several words (i.e., the word has extra spaces).
Ex. 2. نتخا نأ *ntxA n>* instead of نتخاناً *ntxAn>*, **"we fight"** and صبا حاً *SbA HAF* as opposed to صباحاً *SbAHAF* which means **"in the morning"**
(3) If multiple words are stuck together.
Ex. 3. ازعلاوي *AzElAwy* instead of ازعل اوي *AzEl Awy* meaning **"I'll become very upSet"** and

[2]We acknowledge that this list is not comprehensive and quite subjective depending on the dialect being annotated, moreover there are several cases of words where it is quite difficult to judge whether they are borrowings vs. nonces such as technical terms, Ex. *"radio, fax, email, etc."*

كتبتبواسطة *ktbtbwAsTp* instead of كتبت بواسطة *ktbt bwAsTp* which means **"written by"**

NOTE: The following cases are not considered typos:
– Missing/Extra hamza, inconsistent (with respect to an MSA cognate) use of ة *p* and ه *h*, or phonological variants which are used with synonymous MSA cognates.
Ex. 4. انبا *AnbA* instead of انباء *AnbA'* which means **"news"**, and مدرسه *mdrsh* instead of مدرسة *mdrsp* meaning **"school"**
– Words that have speech effects (consecutive repeated characters)
Ex. 5. ووحشتووني *wwH$twwny* instead of وحشتوني *wH$twny* meaning **"I missed you"** and مشكوور *m$kwwr* instead of مشكور *m$kwr* meaning **"thank you"**
Ex. 6. الم *Alm* instead of قلم *qlm* which means **"pen"** and ضهر *Dhr* instead of ظهر *Zhr* meaning **"back"**

G. **Unknown Words [U]**;
These are words that are totally unknown.
ex. بتتكفخ *bttkfx* and شويقه *$wyqh*

3. Annotation Using Guidelines

In order to test the efficacy of our proposed guidelines, we recruited two native speakers for each of Egyptian (EGY) and Levantine (LEV) DA. We asked them to annotate a total of 30K tokens. Below are the details of our study.

3.1. Corpus Collection

We use a corpus that is crawled from Levantine (LEV), Egyptian (EGY) and Iraqi (IRQ) forums for the purposes of the COLABA project (Diab et al., 2010). These forums are rich in dialectal content with frequent code switching to MSA. The collection comprises various threads covering topics ranging from social issues, to religion and politics. Each thread in these forums includes a collection of posts with each post consisting of one or more sentences by a single author/participant. After the data is crawled, the three different collection URLs (Egyptian, Levantine and Iraqi) are further manually scrutinized for majority dialect affiliation, i.e. a manual check to examine each URL's dialect, whether the majority of the postings indeed pertain to the indicated dialect. We sample roughly 10% of each of the URLs crawled for manual inspection.

All threads are then cleaned by removing HTML/XML tags, separating punctuation and numbers from tokens, and tagging words written in Latin script. It is worth mentioning that when digits are interleaved with romanized characters, we did not separate out the digits as this could be an indicator of transliteration of Arabic words. Table 1 shows the number of threads, posts, tokens, types, Latin tokens and types in the three collection URLs of our corpus.

	EGY	LEV	IRQ
Threads	944	2,199	266
Posts	10,857	27,646	1,357
Tokens	2,464,241	3,302,407	427,761
Types	207,594	264,569	66,607
Latin Tokens	50,029	59,329	8,099
Latin Types	11,075	9,190	3,994

Table 1: COLABA Collection Statistics

3.1.1. Ranking the posts

After cleaning the data, we rank all posts according to their estimated dialectal content, where the higher the dialectal content, the higher rank a post would be. Our ranking is based on a formula that uses three criteria: number of words that are non-analyzable by an MSA Morphological Analyzer, number of words present in DA Dictionaries, and number of words written in pure Latin script (we consider transliteration with inter digits among the letters as an indication of Romanized Arabic). This initial ranking is automatically generated.

Criterion 1: Percentage of words analyzable by an MSA morphological analyzer Our initial filter uses an MSA morphological Analyzer (Habash, 2007) that is based on SAMA version 3.1. (SAMA, 2010). The hypothesis is that if a word is found in SAMA then it is considered MSA.
Ex. المشاركة *Alm$Arkp* **"the contribution"** and بواسطة *bwAsTp* **"by"**.

There is significant lexical overlap between DA and MSA. However, we acknowledge the caveat that many words look orthographically similar to MSA while they could be faux amis. For example an MSA word such as بكره *bkrh*, **'I hate/with hate', or 'by a ball'** also occurs as an undiacritized homograph in DA to mean 'tomorrow' in EGY. Moreover, the morphological analyzer SAMA contains dialectal entries. Accordingly, we create a more pure 'MSA' version of SAMA which excludes highly dialectal lemma entries (which occur with much higher frequency in DA rather than MSA based on native speaker intuition of the annotators). We also exclude them from our filtering process in addition to excluding all the lemmas that have **"FOREIGN"** part-of-speech tag in SAMA.) Out of the 42,334 lemmas in SAMA, 1,725 entries(4% of all lem-

mas) are identified as more dialectal hence excluded from this criterion, i.e. we used the resulting filtered SAMA dictionary as the base for ALMOR.

Criterion 2: Percentage of words present in DA dictionaries The next criterion we consider is the number of dialectal words we observe in our dictionaries. We have several Machine Readable Dictionaries (COLABA dictionaries) for the dialects. We created a single word list of all the dialectal entries in our dictionaries (which might comprise orthographically similar words to MSA – undiacritized MSA homographs) and use it as a word look up table. The entries in these dictionaries included a variety of Lemmas and Surface forms some of which are diacritized and some not. Table 2 shows the number of entries in each dictionary.

Examples of words in the DA dictionaries are the following: Ex.دي *dy* **"this"**, and دغري *dgry* **"straight"**.

Dict.	No. of Entries
EGY	31,458
LEV	5,450
IRQ	1,402

Table 2: Statistics of Dialectal Dictionaries

Criterion 3: Number of Latin encoded words We identify all the words that are written in a post using Latin encoding as foreign words. This is a crude initial filter since people can write Arabic in Latin transliteration. We use the English Gigaword to create a word-list of English words and check all Latin words against this list. The majority of which is found in the English word-list. The rest are mostly mis-spelled foreign words, and transliterated names. This suggests that in this kind of forums participants tend to use Arabic and not Latin script.
For example we found the following words in the posts *chemist, fern, leader,..etc*

We note that often participants in the thread tend to use Romanization for their handles for example, *n3na3aah* which means **"mint"** in Arabic is used as a web-name/handle even though the post itself was entirely in Arabic script.

3.2. Annotation Process

Each annotator is presented with three Sets of posts: (i) Set A: posts identified as being mostly DA; (ii) Set B: posts identified as containing a relatively equal number of DA and MSA tokens; and, (iii) Set C: posts identified as being mostly MSA. The identification process is performed automatically using our automated dialect identification tool based on the three criteria described in Section 3.1.1.. They are instructed to read each post in its entirety at least once before starting the annotation on the token level to get a sense of the context. The word-level annotations are performed regardless of whether the words are written in Arabic or Romanized script. All annotators are bilingual speakers of Arabic with at least a college level education or higher. The Egyptian posts are annotated by Egyptian annotators and the Levantine ones are annotated by Syrian annotators.

	No. Posts	No. of Tokens	No. of Types
EGY	331	15,057	6,926
LEV	846	14,195	7,127

Table 3: Statistics of the annotated data-Sets

4. Annotation Results and Discussion

Our current annotation results is based on a pilot annotation. It comprises three Sets of EGY posts and 3 Sets of LEV posts. Table 3 illustrates the statistics of the data used for annotations in both the EGY and the LEV data chosen for annotation. All the data is doubly annotated.
In general, we note that the posts have a majority of the tokens deemed as B indicating that there is a majority semantic cognates in the data. This is not surprising given the domain of the posts which is political and religious.
Comparing Set A to Set C, in both dialects EGY and LEV, we note that Set A has a higher percentage of DA (relative DA content as per the annotators) and Set C has a higher percentage of MSA, indicating that our automatic post ranking is doing a good job of coarse grained classification of the posts. Set B for both dialects have a majority of the tokens ranked as B. In the LEV we note that the number of tokens deemed DA vs. MSA is almost equivalent. In the EGY case, we note that the number of tokens deemed MSA is higher than those deemed DA. Table 4 shows the percentage of tokens assigned to each of the 7 classes by each annotator for Sets A, B, and C of the EGY and LEV data.
Table 6 shows the statistics with the inter-annotator agreements. For the first Levantine Set, Set A which is mostly dialectal, the inter-annotator agreement was 77%. For the second Set, Set B, which is roughly equivalent MSA and DA, the inter-annotator agreement is 53.1%. For Set C, which is deemed mostly MSA, the inter-annotator agreement is 84.3%.
For EGY, the inter-annotator agreement for Set A is 74.5%. For Set B the inter-annotator agreement is 66%, and Set C the inter-annotator agreement is 64%.
In general Set B obtained low inter-annotator agreement across both EGY and LEV. We believe this is the trickier data Set. EGY results overall are relatively worse than LEV results. This maybe attributed to the fact that many of the posts in the EGY collection were not actually EGY consistently.
Ex. بس فعلا حال أمتنا يبجي الصخر يادكتور *bs fElA HAl >mtnA ybjy AlSxr yAdktwr*, in this case the whole post

EGY	Class	Ann. 1	Ann. 2	LEV	Class	Ann. 2	Ann. 2
Set A	M	3%	0.87%	Set A	M	2.1%	6%
	D	41.3%	44%		D	37.3%	31.6%
	B	44%	41.2%		B	45.8%	49.1%
	F	2.1 %	1.2%		F	1%	1.2%
	N	8.3%	7.9%		N	8.9%	4.4%
	U	0.4%	3.2%		U	1.2%	2.7%
	T	0.84%	1.2%		T	3.7%	5.1%
Set B	M	27.8%	9.8%	Set B	M	43.3%	20.3%
	D	1.8%	0.35%		D	15.1%	12.9%
	B	57.8%	70.4%		B	32.9%	60.2%
	F	0.5%	0.16%		F	0.44%	1%
	N	12.4%	18.6%		N	6.8%	4.2%
	U	0.03%	0.28%		U	0.1%	0.4%
	T	0.27%	0.33%		T	1.3%	.9%
Set C	M	34.7%	9.8%	Set C	M	79.7%	77.3%
	D	0.22%	0.35%		D	1.25%	0.7%
	B	46.5%	70.4%		B	3.41%	9.8%
	F	0.16%	0.16%		F	0.02%	0.24%
	N	18.3%	18.6%		N	15.2%	11.8%
	U	0%	0.28%		U	0%	0.04%
	T	0.13%	0.41%		T	0.37%	0.13%

Table 4: Percentage of each class label assigned by each of the EGY and LEV annotators in Sets A, B, and C

Annot. 1	Annot. 2	Ex. (UTF-8)	BW	ENG
M	D	مرحب	mrHb	Welcome
M	B	مصطلح	mSTlH	Expression
M	N	الرجل	Alrjl	The man
M	T	لديهاء	ldyhA’	She has
M	U	العنادل	AlEnAdl	Nightingales
M	F	بكربونات	bkrbwnAt	Bicarbonate
D	B	فاهمة	fAhmp	Understanding
D	N	حفيدو	Hfydw	His grandchild
D	T	علطول	ElTwl	Straight
D	U	هاهاها	hAhAhA	Laugh
D	F	تاكس	tAks	Taxi
B	N	ابو	Abw	Father of
B	T	سال	sAl	Asked
B	U	ضفضعة	DfDEp	Frog
B	F	البوتاجاز	AlbwtAjAz	Butane
F	T	والأسانيد	wAl>sAnyd	And the proofs
F	U	الهافان	AlhAfAn	Havana
T	U	ثويتين	vwytyn	-

Table 5: Examples of disagreements between the Egyptian annotators

is not Egyptian and يبجي *ybjy* is a non-egyptian phonological variant of يبكي *ybky* meaning **"to make it cry"** so it can be difficult for the Egyptian annotators to identify it.

The highest inter-annotator agreement is obtained for Sets C and A in LEV. In Set B, the majority of the confusion is between the N and M label as well as the B and other labels.
In general most disagreements are attributed to:
(1) confusing Named Entities (NE) with DA. Since participants in forums normally use fake names, the task of determining whether a given word is DA or a name becomes more challenging. For example, in the following text the fourth word which means "suits me" in LEV Arabic is used by the writer to refer to another participant who used this word as her web-name, handle alias.

ومرحبا بك اختي دلوعة «وبيلبقلي» وسعداء *wmrHbA bk Axty dlwEp wbylbqly wsEdA'* **"Welcome my sisters Daloo'a, Beyelba'ly and Sa'daa"**

(2) Deciding upon whether a word is used in both MSA and DA as a B class or is used in MSA only (especially if it was part of a multi-word expression). These two observations indicate that we need more refined characterizations of both categories for the annotation as well as the D class. It is worth noting that due to the infiltration of DA into MSA it has become quite complex judging the boundaries between MSA and DA. In the absence of the phonological signal, the context sensitivity is relatively high in judging the class of the tokens. Table 5 shows examples of the disagreements between the two Egyptian annotators.

Below is a detailed analysis with some of our observations.

MSA: Set A, in EGY, 201 tokens are deemed MSA by either annotator, they agree only on 15% of the data. The majority of the words is considered B, 76%. In LEV, for Set A, 359 tokens are considered MSA by the annotators, they agree only on 14.5% of the data. However in this case, the majority of the disagreement falls into the N category with 52.6% followed by 28.1% annotated as B by one of the annotators. *(The participants in the Levantine data were very creative in choosing their names which made it at many times hard to identify)* For Set B, 1,505 in EGY tokens are deemed MSA, 12.2% of which the annotators agreed upon while in LEV 2,253 tokens are considered MSA, and the annotators agreed on 27.2% of them.

For Set C, in EGY 1,198 tokens are deemed MSA, with a 18.2% agreement between the annotators and in LEV 3,875 tokens with 84.1% agreement are considered MSA.

DA: Set A has 3018 DA tokens in EGY and 2094 DA tokens in LEV with an agreement between annotators of 63% and 66% on EGY and LEV respectively. For Set B, 312 tokens are DA in EGY and 836 in LEV with an agreement between annotators of 6% and 51% on EGY and LEV respectively. While in Set C 13 tokens in EGY are considered dialectal with an inter-annotator agreement of 23% and 59 DA tokens in LEV with an inter-annotator agreement of 46%.

Both In general, the majority of the words were annotated as (B), in set A, we have 3,147 (B) tokens in EGY and 2,771 (B) tokens in LEV with an agreement between annotators of 65% and 74% on EGY and LEV, respectively. For Set B, 4,229 tokens are (B) in EGY and 3,023 in LEV with an agreement between annotators of 39% and 64% on EGY and LEV, respectively. While in Set C 2,360 tokens in EGY and 513 in LEV are annotated as B with an inter-annotator agreement of 58% and 17% for the EGY and LEV data respectively. If we want to choose posts that have less (B) words we should probably use a new criteria in the ranking process by taking into consideration the number of dictionaries (ex. Egyptian, Levantine, MSA ..etc) the word occurred in.

Foreign: The Foreign tokens in all sets were very few; this is mainly attributed to the ranking process in which the score of the posts that have many Latin words was penalized, hence were not selected in the annotation process. In Set A, only 133 tokens are considered foreign in EGY and 78 in LEV data. While in Set B, 33 and 48 tokens are deemed foreign in EGY and LEV data, respectively. And in Set C, 6 and 4 tokens in the EGY and LEV data are considered foreign.

Named Entities: Set A has 564 named-entity tokens in EGY and 438 tokens in LEV with an agreement between annotators of 60.5% and 40.4% on EGY and LEV, respectively. For Set B, 511 tokens are named-entities in EGY and 235 in LEV with an agreement between annotators of 60% and 41.7% on EGY and LEV, respectively. While in Set C 334 tokens in EGY are named-entities with an inter-annotator agreement of 82.9% as opposed to 247 tokens in LEV with an inter-annotator agreement of 25%.

Unknown: Set A in both dialects has the highest number of overall unknowns (197 tokens in EGY and 154 tokens in LEV) followed by Set B (35 tokens in EGY and 22 tokens in LEV) then Set C (9 tokens in EGY and 2 tokens in LEV). In LEV in general we find less U class annotation (relative percentage). The U is mostly confused with D (dialectal).

Typo: Similar to unknown words, set A in both dialects has the highest number of overall typos (103 tokens in EGY and 445 tokens in LEV) followed by Set B (24 tokens in EGY and 74 tokens in LEV) then Set C (17 tokens in EGY and 17 tokens in LEV).

From the detailed analysis, it seems that the LEV has more confusability with N for all three data Sets A, B, and C with MSA.

5. Conclusion and Future Direction

In this paper we present a simplified Set of guidelines for creating a large scale corpus of Dialectal Arabic annotations focusing on code switch points targeting informal

EGY Set A								LEV Set A							
	M	D	B	F	T	U	N		M	D	B	F	T	U	N
M	30	-	-	-	-	-	-	M	52	-	-	-	-	-	-
D	11	1,899	-	-	-	-	-	D	12	1,387	-	-	-	-	-
B	153	860	2,045	-	-	-	-	B	101	515	2,054	-	-	-	-
F	0	29	15	60	-	-	-	F	0	27	8	31	-	-	-
T	3	39	11	2	18	-	-	T	5	70	62	1	143	-	-
U	3	132	10	7	8	17	-	U	0	74	3	0	17	40	-
N	1	48	112	20	22	20	341	N	189	9	28	11	4	20	177
EGY Set B								LEV Set B							
	M	D	B	F	T	U	N		M	D	B	F	T	U	N
M	183	-	-	-	-	-	-	M	611	-	-	-	-	-	-
D	84	53	-	-	-	-	-	D	65	427	-	-	-	-	-
B	1,197	171	2,687	-	-	-	-	B	1,464	312	1,172	-	-	-	-
F	0	0	6	6	-	-	-	F	7	9	11	15	-	-	-
T	3	3	5	2	7	-	-	T	16	9	20	1	23	-	-
U	9	0	10	6	2	2	-	U	3	8	6	1	3	1	-
N	29	1	153	13	2	6	307	N	87	6	38	4	2	0	98
EGY Set C								LEV Set C							
	M	D	B	F	T	U	N		M	D	B	F	T	U	N
M	219	-	-	-	-	-	-	M	3,258	-	-	-	-	-	-
D	4	3	-	-	-	-	-	D	17	27	-	-	-	-	-
B	950	6	1,367	-	-	-	-	B	407	13	88	-	-	-	-
F	1	0	0	4	-	-	-	F	5	0	0	1	-	-	-
T	3	0	4	0	0	-	-	T	9	0	3	0	4	-	-
U	5	0	3	0	0	0	-	U	0	2	0	0	0	0	-
N	16	0	30	1	9	1	277	N	179	0	2	3	1	0	62

Table 6: Token level confusion matrix for EGY and LEV data

genres. We used a corpus that was crawled from forums pertaining to Egyptian, Levantine and Iraqi dialects. The corpus is automatically cleaned and the posts in the forums are ranked according to their automatically estimated level of dialectalness. We annotated 3 Levantine Sets and 3 Egyptian Sets according to these guidelines. The annotators agreed with our automatic post classification for both dialects. However, we note that two classes seem to cause a significant amount of confusion for the annotators namely B and N. We plan on further refinement of the guidelines to alleviate the points of confusion among the annotators aiming for higher inter-annotator agreements. We plan on using at least 3 annotators per dialect with adjudication. In the near future, we plan on annotating more data and adding a level of DA class annotation where the annotator further specifies the dialect of the token in context such as EGY or LEV or some other DA.

6. Acknowledgement

We would like to thank the feedback received from 3 anonymous reviewers. Also, we thank Reem Faraj, Abdelati Hawwari, and Nizar Habash on their useful feedback on the guidelines and annotation process. This work was partially funded by DARPA project number HR0011-12-C-0014. This material is based upon work partially supported by the National Science Foundation under Grant No. 0958440. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

7. References

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Al Tantawy and Yassine Benajiba. 2010. *COLABA: Arabic Dialect Annotation and Processing*. In Proceedings of LREC Workshop on Semitic Language Processing, Malta, May 2010.

Ferguson. 1959 *Diglossia*. Word 15. 325340.

Nizar Habash. *Arabic Morphological Representations for Machine Translation*. 2007. Book chapter in Arabic Computational Morphology: Knowledge-based and Empirical Method, 2009-07-03 17:45:32 -0400 Editors: A. van den Bosch and A. Soudi

Nizar Habash. 2010 *Introduction to Arabic Natural Language Processing*, Morgan & Claypool Publishers

Nizar Habash, Owen Rambow, Mona Diab and Reem Farraj. 2008. *Guidelines for Annotating Arabic Dialect*. In Proceedings of Workshop on Arabic and its local languages, LREC, Marrakech, Morocco. 2008.

Mohamed Maamouri, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, Seth Kulick 2010: 2010. *LDC Standard Arabic Morphological Analyzer (SAMA) Version 3.1* `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2010L01`