Arabic Dialect Identification

Omar F. Zaidan*
Microsoft Research

Chris Callison-Burch**
University of Pennsylvania

The written form of the Arabic language, Modern Standard Arabic *(MSA), differs in a non-trivial manner from the various spoken regional dialects of Arabic—the true "native languages" of Arabic speakers. Those dialects, in turn, differ quite a bit from each other. However, due to MSA's prevalence in written form, almost all Arabic data sets have predominantly MSA content. In this article, we describe the creation of a novel Arabic resource with dialect annotations. We have created a large monolingual data set rich in dialectal Arabic content called the* Arabic Online Commentary Data set *(Zaidan and Callison-Burch 2011). We describe our annotation effort to identify the dialect level (and dialect itself) in each of more than 100,000 sentences from the data set by crowdsourcing the annotation task, and delve into interesting annotator behaviors (like over-identification of one's own dialect). Using this new annotated data set, we consider the task of Arabic dialect identification: Given the word sequence forming an Arabic sentence, determine the variety of Arabic in which it is written. We use the data to train and evaluate automatic classifiers for dialect identification, and establish that classifiers using dialectal data significantly and dramatically outperform baselines that use MSA-only data, achieving near-human classification accuracy. Finally, we apply our classifiers to discover dialectical data from a large Web crawl consisting of 3.5 million pages mined from on-line Arabic newspapers.*

1. Introduction

The Arabic language is a loose term that refers to the many existing varieties of Arabic. Those varieties include one "written" form, *Modern Standard Arabic* (MSA), and many "spoken" forms, each of which is a regional dialect. MSA is the only variety that is standardized, regulated, and taught in schools, necessitated by its use in written communication and formal venues. The regional dialects, used primarily for day-to-day dealings and spoken communication, remain somewhat absent from written communication compared with MSA. That said, it is certainly possible to produce dialectal Arabic text, by using the same letters used in MSA and the same (mostly phonetic) spelling rules of MSA.

* E-mail: ozaidan@gmail.com.
** Computer and Information Science Department University of Pennsylvania, Levine Hall, room 506, 3330 Walnut Street, Philadelphia, PA 19104. E-mail: ccb@cis.upenn.edu.

Submission received: 12 March 2012; revised version received: 14 March 2012; accepted for publication: 17 April 2013.

doi:10.1162/COLI_a_00169

One domain of written communication in which both MSA and dialectal Arabic are commonly used is the on-line domain: Dialectal Arabic has a strong presence in blogs, forums, chatrooms, and user/reader commentary. Harvesting data from such sources is a viable option for computational linguists to create large data sets to be used in statistical learning setups. However, because all Arabic varieties use the same character set, and furthermore much of the vocabulary is shared among different varieties, it is not a trivial matter to distinguish and separate the dialects from each other.

In this article, we focus on the problem of Arabic dialect identification. We describe a large data set that we created by harvesting a large amount of reader commentary on on-line newspaper content, and describe our annotation effort on a subset of the harvested data. We crowdsourced an annotation task to obtain sentence-level labels indicating what proportion of the sentence is dialectal, and which dialect the sentence is written in. Analysis of the collected labels reveals interesting annotator behavior patterns and biases, and the data are used to train and evaluate automatic classifiers for dialect detection and identification. Our approach, which relies on training language models for the different Arabic varieties, greatly outperforms baselines that use (much more) MSA-only data: On one of the classification tasks we considered, where human annotators achieve 88.0% classification accuracy, our approach achieves 85.7% accuracy, compared with only 66.6% accuracy by a system using MSA-only data.

The article is structured as follows. In Section 2, we provide an introduction to the various Arabic varieties and corresponding data resources. In Section 3, we introduce the dialect identification problem for Arabic, discussing what makes it a difficult problem, and what applications would benefit from it. Section 4 provides details about our annotation set-up, which relied on crowdsourcing the annotation to workers on Amazon's Mechanical Turk. By examining the collected labels and their distribution, we characterize annotator behavior and observe several types of human annotator biases. We introduce our technique for automatic dialect identification in Section 5. The technique relies on training separate language models for the different Arabic varieties, and scoring sentences using these models. In Section 6, we report on a large-scale Web crawl that we performed to gather a large amount of Arabic text from on-line newspapers, and apply our classifier on the gathered data. Before concluding, we give an overview of related work in Section 7.

2. Background: The MSA/Dialect Distinction in Arabic

Although *the Arabic language* has an official status in over 20 countries and is spoken by more than 250 million people, the term itself is used rather loosely and refers to different varieties of the language. Arabic is characterized by an interesting linguistic dichotomy: the written form of the language, MSA, differs in a non-trivial fashion from the various *spoken varieties* of Arabic, each of which is a regional dialect (or a *lahjah, lit.* "accent"; also *darjah, lit.* "modern"). MSA is the only variety that is standardized, regulated, and taught in schools. This is necessitated because of its use in written communication in formal venues.[1] The regional dialects, used primarily for day-to-day dealings and spoken communication, are not taught formally in schools, and remain somewhat absent from traditional, and certainly official, written communication.

[1] The term *MSA* is used primarily by linguists and in educational settings. For example, constitutions of countries where Arabic is an official language simply refer to *The Arabic Language*, the reference to the standard form of Arabic being implicit.

Unlike MSA, a regional dialect does not have an explicit written set of grammar rules regulated by an authoritative organization, but there is certainly a concept of *grammatical* and *ungrammatical*.[2] Furthermore, even though they are "spoken" varieties, it is certainly possible to produce dialectal Arabic *text*, by spelling out words using the same spelling rules used in MSA, which are mostly phonetic.[3]

There is a reasonable level of mutual intelligibility across the dialects, but the extent to which a particular individual is able to understand other dialects depends heavily on that person's own dialect and their exposure to Arab culture and literature from outside of their own country. For example, the typical Arabic speaker has little trouble understanding the Egyptian dialect, thanks in no small part to Egypt's history in movie-making and television show production, and their popularity across the Arab world. On the other hand, the Moroccan dialect, especially in its spoken form, is quite difficult to understand by a Levantine speaker. Therefore, from a scientific point of view, the dialects can be considered separate languages in their own right, much like North Germanic languages (Norwegian/Swedish/Danish) and West Slavic languages (Czech/Slovak/Polish).[4]

2.1 The Dialectal Varieties of Arabic

One possible breakdown of regional dialects into main groups is as follows (see Figure 1):

- **Egyptian**: The most widely understood dialect, due to a thriving Egyptian television and movie industry, and Egypt's highly influential role in the region for much of the 20th century (Haeri 2003).
- **Levantine**: A set of dialects that differ somewhat in pronunciation and intonation, but are largely equivalent in written form; closely related to Aramaic (Bassiouney 2009).
- **Gulf**: Folk wisdom holds that Gulf is the closest of the regional dialect to MSA, perhaps because the current form of MSA evolved from an Arabic variety originating in the Gulf region. Although there are major differences between Gulf and MSA, Gulf has notably preserved more of MSA's verb conjugation than other varieties have (Versteegh 2001).

[2] There exist resources that describe grammars and dictionaries of many Arabic dialects (e.g., Abdel-Massih, Abdel-Malek, and Badawi 1979; Badawi and Hinds 1986; Cowell 1964; Erwin 1963; Ingham 1994; Holes 2004), but these are compiled by individual linguists as one-off efforts, rather than updated regularly by central regulatory organizations, as is the case with MSA and many other world languages.

[3] Arabic speakers writing in dialectal Arabic mostly follow MSA spelling rules in cases where MSA is *not* strictly phonetic as well (e.g., the pronunciation of the definite article *Al*). Habash, Diab, and Rambow (2012) have proposed CODA, a Conventional Orthography for Dialectal Arabic, to standardize the spelling of Arabic dialect computational models.

[4] Note that such a view is not widely accepted by Arabic speakers, who hold MSA in high regard. They consider dialects, including their own, to be simply imperfect, even "corrupted," versions of MSA, rather than separate languages (Suleiman 1994). One exception might be the Egyptian dialect, where a nationalistic movement gave rise to such phenomena as the Egyptian Wikipedia, with articles written exclusively in Egyptian, and little, if any, MSA. Another notable exception is the Lebanese poet Said Akl, who spearheaded an effort to recognize Lebanese as an independent language, and even proposed a Latin-based Lebanese alphabet.

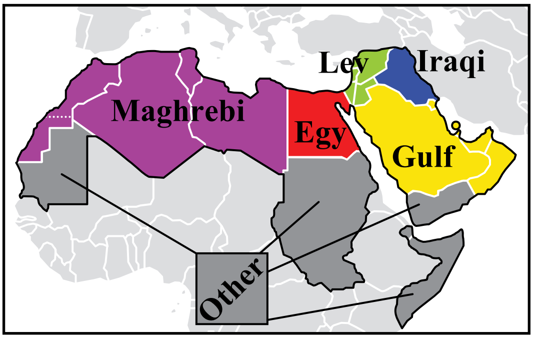


Figure 1
One possible breakdown of spoken Arabic into dialect groups: Maghrebi, Egyptian, Levantine, Gulf, and Iraqi. Habash (2010) and Versteegh (2001) give a breakdown along mostly the same lines. Note that this is a relatively coarse breakdown, and further division of the dialect groups is possible, especially in large regions such as the Maghreb.

- **Iraqi**: Sometimes considered to be one of the Gulf dialects, though it has distinctive features of its own in terms of prepositions, verb conjugation, and pronunciation (Mitchell 1990).
- **Maghrebi**: Heavily influenced by the French and Berber languages. The Western-most varieties could be unintelligible by speakers from other regions in the Middle East, especially in spoken form. The Maghreb is a large region with more variation than is seen in other regions such as the Levant and the Gulf, and could be subdivided further (Mohand 1999).

There are a large number of linguistic differences between MSA and the regional dialects. Some of those differences do not appear in written form if they are on the level of short vowels, which are omitted in Arabic text anyway. That said, many differences manifest themselves textually as well:

- MSA's morphology is richer than dialects' along some dimensions such as case and mood. For instance, MSA has a *dual* form in addition to the singular and plural forms, whereas the dialects mostly lack the dual form. Also, MSA has two plural forms, one masculine and one feminine, whereas many (though not all) dialects often make no such gendered distinction.[5] On the other hand, dialects have a more complex cliticization system than MSA, allowing for circumfix negation, and for attached pronouns to act as indirect objects.
- Dialects lack grammatical case, whereas MSA has a complex case system. In MSA, most cases are expressed with diacritics that are rarely explicitly written, with the accusative case being a notable exception, as it is expressed using a suffix (+*A*) in addition to a diacritic (e.g., on objects and adverbs).

[5] Dialects may preserve the dual form for nouns, but often lack it in verb conjugation and pronouns, using plural forms instead. The same is true for the gendered plural forms, which exist for many nouns (e.g., 'teachers' is either *mςlmyn* [male] or *mςlmAt* [female]), but not used otherwise as frequently as in MSA.

- There are lexical choice differences in the vocabulary itself. Table 1 gives several examples. Note that these differences go beyond a lack of orthography standardization.
- Differences in verb conjugation, even when the triliteral root is preserved. See the lower part of Table 1 for some conjugations of the root *š-r-b* (to drink).

This list, and Table 1, deal with differences that are expressed at the inidividual-word level. It is important to note that Arabic varieties differ markedly in style and sentence composition as well. For instance, all varieties of Arabic, MSA, and otherwise, allow both SVO and VSO word orders, but MSA has a higher incidence of VSO sentences than dialects do (Aoun, Benmamoun, and Sportiche 1994; Shlonsky 1997).

2.2 Existing Arabic Data Sources

Despite the fact that speakers are usually less comfortable communicating in MSA than in their own dialect, MSA content significantly dominates dialectal content, as MSA is the variant of choice for formal and official communication. Relatively little printed material exists in local dialects, such as folkloric literature and some modern poetry, but the vast majority of published Arabic is in MSA. As a result, MSA's dominance is also apparent in data sets available for linguistic research. The problem is somewhat mitigated in the speech domain, since dialectal data exists in the form of phone conversations and television program recordings, but, in general, dialectal Arabic data sets are hard to come by.

Table 1
A few examples illustrating similarities and differences across MSA and three Arabic dialects: Levantine, Gulf, and Egyptian. Even when a word is spelled the same across two or more varieties, the pronunciation might differ due to differences in short vowels (which are not spelled out). Also, due to the lack of orthography standardization, and variance in pronunciation even within a single dialect, some dialectal words could have more than one spelling (e.g., Egyptian "I drink" could be *bAšrb*, Levantine "He drinks" could be *byšrb*). (We use the Habash-Soudi-Buckwalter transliteration scheme to represent Arabic orthography, which maps each Arabic letter to a single, distinct character. We provide a table with the character mapping in Appendix A.)

English	MSA	LEV	GLF	EGY
Book	*ktAb*	*ktAb*	*ktAb*	*ktAb*
Year	*snħ*	*snħ*	*snħ*	*snħ*
Money	*nqwd*	*mSAry*	*flws*	*flws*
Come on!	*hyA!*	*ylA!*	*ylA!*	*ylA!*
I want	*Aryd*	*bdy*	*Abγý*	*ςAyz*
Now	*AlĀn*	*hlq*	*AlHyn*	*dlwqt*
When?	*mtý?*	*Aymtý?*	*mtý?*	*Amtý?*
What?	*mAA?*	*Ayš?*	*wš?*	*Ayh?*
I drink	*Âšrb*	*bšrb*	*Ašrb*	*bšrb*
He drinks	*yšrb*	*bšrb*	*yšrb*	*byšrb*
We drink	*nšrb*	*bnšrb*	*nšrb*	*bnšrb*

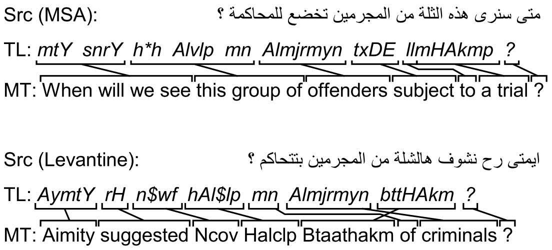

Src (MSA):
متى سنرى هذه الثلة من المجرمين تخضع للمحاكمة ؟
TL: mtY snrY h*h Alvlp mn Almjrmyn txDE llmHAkmp ?
MT: When will we see this group of offenders subject to a trial ?
Src (Levantine):
ايمتى رح نشوف هالشلة من المجرمين بتتحاكم ؟
TL: AymtY rH n$wf hAl$lp mn Almjrmyn bttHAkm ?
MT: Aimity suggested Ncov Halclp Btaathakm of criminals ?

Figure 2
Two roughly equivalent Arabic sentences, one in MSA and one in Levantine Arabic, translated by the same MT system (Google Translate) into English. An acceptable translation would be *When will we see this group of criminals undergo trial* (or *tried*)?. The MSA variant is handled well, whereas the dialectal variant is mostly transliterated.

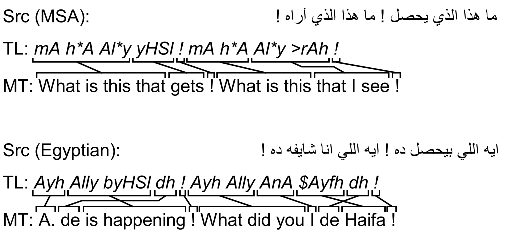

Src (MSA):
ما هذا الذي يحصل ! ما هذا الذي أراه !
TL: mA h*A Al*y yHSl ! mA h*A Al*y >rAh !
MT: What is this that gets ! What is this that I see !
Src (Egyptian):
ايه اللي بيحصل ده ! ايه اللي انا شايفه ده !
TL: Ayh Ally byHSl dh ! Ayh Ally AnA $Ayfh dh !
MT: A. de is happening ! What did you I de Haifa !

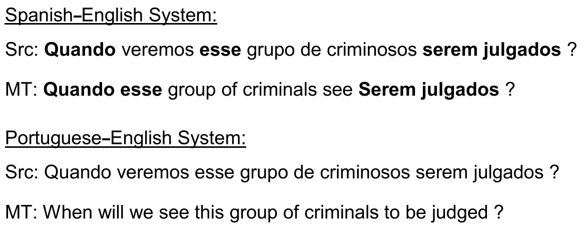
Figure 3
Two roughly equivalent Arabic sentences, one in MSA and one in Egyptian Arabic, translated by the same MT system (Google Translate) into English. An acceptable translation would be *What is this that is happening? What is this that I'm seeing?*. As in Figure 2, the dialectal variant is handled quite poorly.

The abundance of MSA data has greatly aided research on computational methods applied to Arabic, but only the MSA variant of it. For example, a state-of-the-art Arabic-to-English machine translation system performs quite well when translating MSA source sentences, but often produces incomprehensible output when the input is dialectal. For example, most words of the dialectal sentence shown in Figure 2 are transliterated, whereas an equivalent MSA sentence is handled quite well. The high transliteration rate is somewhat alarming, as the first two words of the dialectal sentence are relatively frequent function words: *Aymtý* means 'when' and *rH* corresponds to the modal 'will'.

Figure 3 shows another dialectal sentence, this time in Egyptian, which again causes the system to produce a poor translation even for frequent words. Case in point, the system is unable to consistently handle any of *Ayh* ('what'), *Ally* (the conjunction 'that'), or *dh* ('this'). Granted, it is conceivable that processing dialectal content is more difficult than MSA, but the main problem is the lack of dialectal training data.[6]

This is an important point to take into consideration, because the dialects differ to a large enough extent to warrant treating them as more or less different languages. The behavior of machine translation systems translating *dialectal* Arabic when the system

[6] In the context of machine translation in particular, additional factors make translating dialectal content difficult, such as a general mismatch between available training data and the topics that are usually discussed dialectally.



Spanish–English System:

Src: **Quando** veremos **esse** grupo de criminosos **serem julgados** ?

MT: **Quando esse** group of criminals see **Serem julgados** ?

Portuguese–English System:

Src: Quando veremos esse grupo de criminosos serem julgados ?

MT: When will we see this group of criminals to be judged ?

Figure 4
The output of a Spanish-to-English system when given a Portuguese sentence as input, compared with the output of a Portuguese-to-English system, which performs well. The behavior is very similar to that in Figures 2 and 3, namely, the failure to translate out-of-vocabulary words when there is a language mismatch.

has been trained exclusively on **MSA** data is similar to the behavior of a **Spanish**-to-English MT system when a user inputs a *Portuguese* sentence. Figure 4 illustrates how MT systems behave (the analogy is not intended to draw a parallel between the linguistic differences MSA-dialect and Spanish-Portuguese). The MT system's behavior is similar to the Arabic example, in that words that are shared in common between Spanish and Portuguese are translated, while the Portuguese words that were never observed in the Spanish training data are left untranslated.

This example illustrates the need for dialectal data to train MT systems to handle dialectal content properly. A similar scenario would arise with many other NLP tasks, such as parsing or speech recognition, where dialectal content would be needed in large quantities for adequate training. A robust dialect identifier could sift through immense volumes of Arabic text, and separate out dialectal content from MSA content.

2.3 Harvesting Dialect Data from On-line Social Media

One domain of written communication in which MSA and dialectal Arabic are both commonly used is the on-line domain, because it is more individual-driven and less institutionalized than other venues. This makes a dialect much more likely to be the user's language of choice, and dialectal Arabic has a strong presence in blogs, forums, chatrooms, and user/reader commentary. Therefore, on-line data is a valuable resource of dialectal Arabic text, and harvesting this data is a viable option for computational linguists for purposes of creating large data sets to be used in statistical learning.

We created the *Arabic On-line Commentary* Data Set (AOC) (Zaidan and Callison-Burch 2011) a 52M-word monolingual data set by harvesting reader commentary from the on-line versions of three Arabic newspapers (Table 2). The data is characterized by the prevalence of dialectal Arabic, alongside MSA, mainly in Levantine, Gulf, and Egyptian. These correspond to the countries in which the three newspapers are published: *Al-Ghad* is from Jordan, *Al-Riyadh* is from Saudi Arabia, and *Al-Youm Al-Sabe'* is from Egypt.[7]

Although a significant portion of the AOC's content is dialectal, there is still a very large portion of it that is in MSA. (Later analysis in Section 4.2.1 shows dialectal content is roughly 40%.) In order to take full advantage of the AOC (and other Arabic data sets

[7] URLs: www.alghad.com, www.alriyadh.com, and www.youm7.com.

Table 2
A summary of the different components of the AOC data set. Overall, 1.4M comments were harvested from 86.1K articles, corresponding to 52.1M words.

News Source	*Al-Ghad*	*Al-Riyadh*	*Al-Youm Al-Sabe'*	ALL
# articles	6.30K	34.2K	45.7K	86.1K
# comments	26.6K	805K	565K	1.4M
# sentences	63.3K	1,686K	1,384K	3.1M
# words	1.24M	18.8M	32.1M	52.1M
comments/article	4.23	23.56	12.37	16.21
sentences/comment	2.38	2.09	2.45	2.24
words/sentence	19.51	11.14	23.22	16.65

with at least some dialectal content), it is desirable to separate dialectal content from non-dialectal content automatically. The task of dialect identification (and its automation) is the focus for the remainder of this article. We next present the task of Arabic dialect identification, and discuss our effort to create a data set of Arabic sentences with their dialectal labels. Our annotation effort relied on crowdsourcing the annotation task to Arabic-speakers on Amazon's Mechanical Turk service (Section 3).

3. Arabic Dialect Identification

The discussion of the varieties of Arabic and the differences between them gives rise to the task of automatic **dialect identification** (DID). In its simplest form, the task is to build a learner that can, given an Arabic sentence S, determine whether or not S contains dialectal content. Another form of the task would be to determine in *which* dialect S was written, which requires identification at a more fine-grained level.

In many ways, DID is equivalent to *language* identification. Although language identification is often considered to be a "solved problem," DID is most similar to a particularly difficult case of language ID, where it is applied to a group of closely related languages that share a common character set. Given the parallels between DID and language identification, we investigate standard statistical methods to establish how difficult the task is. We discuss prior efforts for Arabic DID in Section 7.

3.1 The Difficulty of Arabic DID

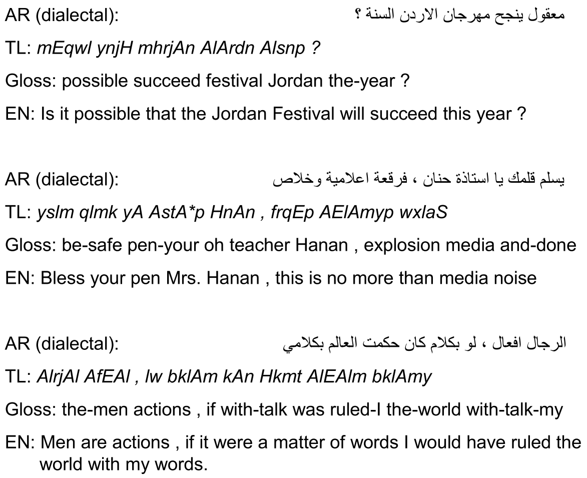
Despite the differences illustrated in the previous section, in which we justify treating the different dialects as separate languages, it is not a trivial matter to *automatically* distinguish and separate the dialects from each other. Because all Arabic varieties use the same character set, and because much of the vocabulary is shared among different varieties, identifying dialect in a sentence is not simply a matter of, say, compiling a dialectal dictionary and detecting whether or not a given sentence contains dialectal words.

This word-level source ambiguity is caused by several factors:

- A dialectal sentence might consist entirely of words that are used across all Arabic varieties, including MSA. Each of the sentences in Figure 5 consists of words that are used both in MSA and dialectally, and an MSA-based dictionary would not (and should not) recognize those words as out of vocabulary (OOV). Nevertheless, the sentences are heavily dialectal.

- Some words are used across the varieties with *different* functions. For example, *Tyb* is used dialectally as an interjection, but is an adjective in MSA. (This is similar to the English usage of *okay*.)
- Primarily due to the omission of short vowels, a dialectal word might have the same spelling as an MSA word with an entirely different meaning, forming pairs of heteronyms. This includes strongly dialectal words such as *dwl* and *nby*: *dwl* is either Egyptian for *these* (pronounced *dowl*) or the MSA for *countries* (pronounced *duwal*); *nby* is either the Gulf for *we want* (pronounced *nibi*) or the MSA for *prophet* (pronounced *nabi*).

It might not be clear for a non-Arabic speaker what makes certain sentences, such as those of Figure 5, dialectal, even when none of the individual words are. The answer lies in the *structure* of such sentences and the particular *word order* within them, rather than the individual words themselves taken in isolation. Figure 6 shows MSA sentences that express the same meaning as the dialectal sentences from Figure 5. As one could see, the two versions of any given sentence could share much of the vocabulary, but in ways that are noticeably different to an Arabic speaker. Furthermore, the differences would be starker still if the MSA sentences were composed from scratch, rather than by modifying the dialectal sentences, since the tone might differ substantially when composing sentences in MSA.



AR (dialectal): معقول ينجح مهرجان الاردن السنة ؟

TL: *mEqwl ynjH mhrjAn AlArdn Alsnp ?*

Gloss: possible succeed festival Jordan the-year ?

EN: Is it possible that the Jordan Festival will succeed this year ?

AR (dialectal): يسلم قلمك يا استاذة حنان ، فرقعة اعلامية وخلاص

TL: *yslm qlmk yA AstA*p HnAn , frqEp AElAmyp wxlaS*

Gloss: be-safe pen-your oh teacher Hanan , explosion media and-done

EN: Bless your pen Mrs. Hanan , this is no more than media noise

AR (dialectal): الرجال افعال ، لو بكلام كان حكمت العالم بكلامي

TL: *AlrjAl AfEAl , lw bklAm kAn Hkmt AlEAlm bklAmy*

Gloss: the-men actions , if with-talk was ruled-I the-world with-talk-my

EN: Men are actions , if it were a matter of words I would have ruled the world with my words.

Figure 5
Three sentences that were identified by our annotators as dialectical, even thought they do not contain individually dialectal words. A word-based OOV-detection approach would fail to classify these sentences as being dialectal, because all these words could appear in an MSA corpus. One might argue that a distinction should be drawn between informal uses of MSA versus dialectical sentences, but annotators consistently classify these sentences as dialect.

AR (dialectal): معقول ينجح مهرجان الاردن السنة ؟
TL: *mEqwl ynjH mhrjAn AlArdn Alsnp ?*
Gloss: possible succeed festival Jordan the-year ?
AR (MSA): هل من الممكن أن ينجح مهرجان الأردن هذه السنة ؟
TL: *hl mn Almmkn >n ynjH mhrjAn Al>rdn h*h Alsnp ?*
Gloss: is? of the-possible that succeed festival Jordan this the-year ?
EN: Is it possible that the Jordan Festival will succeed this year ?

AR (dialectal): يسلم قلمك يا استاذة حنان ، فرقعة اعلامية وخلاص
TL: *yslm qlmk yA AstA*p HnAn , frqEp AElAmyp wxlaS*
Gloss: be-safe pen-your oh teacher Hanan , explosion media and-done
AR (MSA): سلم قلمك يا أستاذة حنان ، هذه مجرد ضجة إعلامية
TL: *slm qlmk yA >stA*p HnAn , h*h mjrd Djp <ElAmyp*
Gloss: was-safe pen-your oh teacher Hanan , this only noise media
EN: Bless your pen Mrs. Hanan , this is no more than media noise

AR (dialectal): الرجال افعال ، لو بكلام كان حكمت العالم بكلامي
TL: *AlrjAl AfEAl , lw bklAm kAn Hkmt AlEAlm bklAmy*
Gloss: the-men actions , if with-talk was ruled-I the-world with-talk-my
AR (MSA): الرجال بالأفعال ، لو بالكلام لحكمت العالم بكلامي
TL: *AlrjAl bAl>fEal , lw bAlklAm lHkmt AlEAlm bklAmy*
Gloss: the-men with-the-actions , if with-the-talk would-ruled-I the-world with-talk-my
EN: Men are actions , if it were a matter of words I would have ruled the world with my words.

Figure 6
The dialectal sentences of Figure 5, with MSA equivalents.

3.2 Applications of Dialect Identification

Being able to perform automatic DID is interesting from a purely linguistic and experimental point of view. In addition, automatic DID has several useful applications:

- Distinguishing dialectal data from non-dialectal data would aid in creating a large monolingual dialectal data set, exactly as we would hope to do with the AOC data set. Such a data set would aid many NLP systems that deal with dialectal content, for instance, to train a language model for an Arabic dialect speech recognition system (Novotney, Schwartz, and Khudanpur 2011). Identifying dialectal content can also aid in creating parallel data sets for machine translation, with a dialectal source side.
- A user might be interested in content of a specific dialect, or, conversely, in strictly non-dialectal content. This would be particularly relevant in fine-tuning and personalizing search engine results, and could allow for better user-targeted advertising. In the same vein, being able to recognize dialectal content in user-generated text could aid in characterizing communicants and their biographic attributes (Garera and Yarowsky 2009).

- In the context of an application such as machine translation (MT), identifying dialectal content could be quite helpful. Most MT systems, when faced with OOV words, either discard the words or make an effort to transliterate them. If a segment is identified as being dialectal first, the MT system might instead attempt to find equivalent MSA words, which are presumably easier to process correctly (e.g., as in Salloum and Habash [2011] and, to some degree, Habash [2008]). Even for non-OOV words, identifying dialectal content before translating could be critical to resolve the heteronym ambiguity of the kind mentioned in Section 3.1.

4. Crowdsourcing Arabic Dialect Annotation

In this section, we discuss crowdsourcing Arabic dialect annotation. We discuss how we built a data set of Arabic sentences, each of which is labeled with whether or not it contains dialectal content. The labels include additional details about the **level** of dialectal content (i.e., how much dialect there is), and of which **type** of dialect it is. The sentences themselves are sampled from the AOC data set, and we observe that about 40% of sentences contain dialectal content, with that percentage varying between 37% and 48%, depending on the news source.

Collecting annotated data for speech and language applications requires careful quality control (Callison-Burch and Dredze 2010). We present the annotation interface and discuss an effective way for quality control that can detect spamming behavior. We then examine the collected data itself, analyzing annotator behavior, measuring agreement among annotators, and identifying interesting biases exhibited by the annotators. In Section 5, we use the collected data to train and evaluate statistical models for several dialect identification tasks.

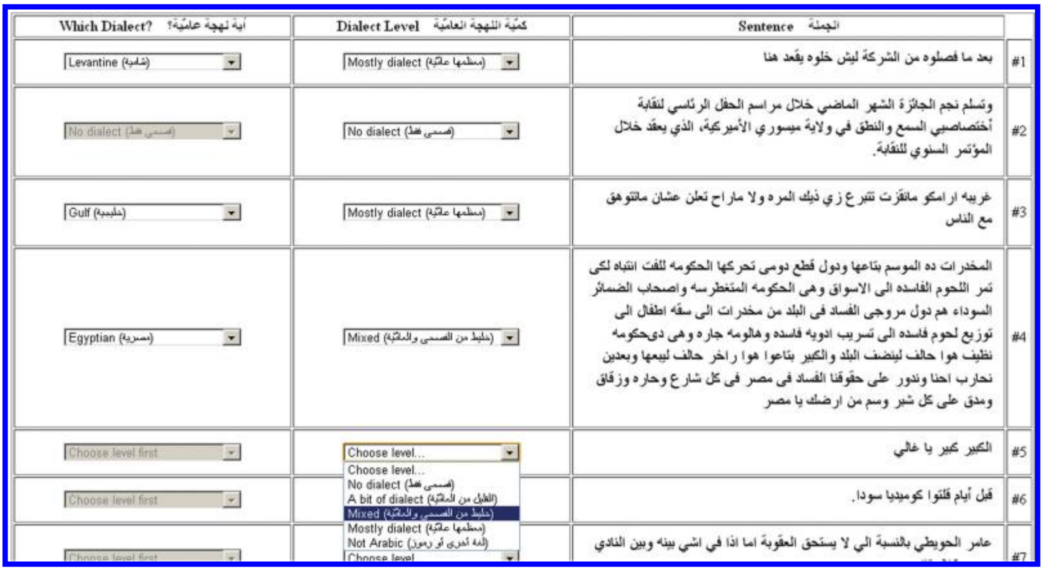
4.1 Annotation Interface

The annotation interface displayed a group of Arabic sentences, randomly selected from the AOC. For each sentence, the annotator was instructed to examine the sentence and make two judgments about its dialectal content: the **level** of dialectal content, and its **type**, if any. The instructions were kept short and simple:

> This task is for Arabic speakers who understand the different local Arabic dialects, and can distinguish them from *Fusha*[8] Arabic.
>
> Below, you will see several Arabic sentences. For each sentence:
>
> 1. Tell us how much dialect is in the sentence, and then
> 2. Tell us which Arabic dialect the writer intends.

The instructions were accompanied by the map of Figure 1, to visually illustrate the dialect breakdown. Figure 7 shows the annotator interface populated with some actual examples, with labeling in progress. We also collected self-reported information such as native Arabic dialect and age (or number of years speaking Arabic for non-native speakers). The interface also had built-in functionality to detect each annotator's geographic location based on their IP address.

[8] *Fusha* is the Arabic word for MSA, pronounced *foss-ha*.



Which Dialect? أية لهجة عاميّة؟	Dialect Level كميّة اللهجة العاميّة	Sentence الجملة	
Levantine (شامية)	Mostly dialect (معظمها عامّية)	بعد ما فصلوه من الشركة ليش خلوه يقعد هنا	#1
No dialect (فصحى فقط)	No dialect (فصحى فقط)	وتسلم نجم الجائزة الشهر الماضي خلال مراسم الحفل الرئاسي لنقابة أختصاصيي السمع والنطق في ولاية ميسوري الأميركية، الذي يعقد خلال المؤتمر السنوي للنقابة.	#2
Gulf (خليجية)	Mostly dialect (معظمها عامّية)	غريبه ارامكو مانقزت تتبرع زي ذيك المره ولا ماراح تعلن عشان ماتتوهق مع الناس	#3
Egyptian (مصرية)	Mixed (خليط من الفصحى والعامّية)	المخدرات ده الموسم بتاعها ودول قطع دومى تحركها الحكومه للفت انتباه لكى تمر اللحوم الفاسده الى الاسواق وهى الحكومه المتغطرسه واصحاب الضمائر السوداء هم دول مروجى الفساد فى البلد من مخدرات الى سقه اطفال الى توزيع لحوم فاسده الى تسريب ادويه فاسده وهالومه جاره وهى دىحكومه نظيف هوا حالف لينضف البلد والكبير بتاعوا هوا راخر حالف لبيعها وبعدين نحارب احنا وندور على حقوقنا الفساد فى مصر فى كل شارع وحاره وزقاق ومدق على كل شبر وسم من ارضك يا مصر	#4
Choose level first	Choose level... Choose level... No dialect (فصحى فقط) A bit of dialect (القليل من العامّية) Mixed (خليط من الفصحى والعامّية) Mostly dialect (معظمها عامّية) Not Arabic (لغة أخرى أو رموز)	الكبير كبير يا غالي	#5
Choose level first		قبل أيام قلتوا كوميديا سودا.	#6
Choose level first	Choose level	عامر الحويطي بالنسبة الي لا يستحق العقوبة اما اذا في اشي بينه وبين النادي	#7

Figure 7
The interface for the dialect identification task. This example, and the full interface, can be viewed at the `http://bit.ly/eUtiO3`.

Of the 3.1M sentences in the AOC, we randomly[9] selected a "small" subset of about 110,000 sentences to be annotated for dialect.

For each sentence shown in the interface, we asked annotators to label which dialect the segment is written in and the level of dialect in the segment. The dialect labels were Egyptian, Gulf, Iraqi, Levantine, Maghrebi, other dialect, general dialect (for segments that could be classified as multiple dialects), dialect but unfamiliar (for sentences that are clearly dialect, but are written in a dialect that the annotator is not familiar with), no dialect (for MSA), or not Arabic (for segments written in English or other languages). Options for the level of dialect included no dialect (for MSA), a small amount of dialect, an even mix of dialect and MSA, mostly dialect, and not Arabic. For this article we use only the dialect labels, and not the level of dialect. Zaidan (2012) incorporates finer-grained labels into an "annotator rationales" model (Zaidan, Eisner, and Piatko 2007).

The sentences were randomly grouped into sets of 10 sentences each, and when Workers performed our task, they were shown the 10 sentences of a randomly selected set on a single HTML page. As a result, each screen contained a mix of sentences across the three newspapers presented in random order. As control items, each screen had two additional sentences that were randomly sampled from the *article bodies*. Such sentences are almost always in MSA Arabic, and so their expected label is MSA. Any worker who frequently mislabeled the control sentences with a non-MSA label was considered a spammer, and their work was rejected. Hence, each screen had twelve sentences in total.

[9] There are far fewer sentences available from *Al-Ghad* commentary than the other two sources over any given period of time (third line of Table 2). We have taken this imbalance into account and heavily oversampled *Al-Ghad* sentences when choosing sentences to be labeled, to obtain a subset that is more balanced across the three sources.

We offered a reward of $0.05 per screen (later raised to $0.10), and had each set redundantly completed by three distinct Workers. The data collection lasted about 4.5 months, during which 33,093 Human Intelligence Task (HIT) Assignments were completed, corresponding to 330,930 collected labels (excluding control items). The total cost of annotation was $3,050.52 ($2,773.20 for rewards, and $277.32 for Amazon's commission).

4.2 Annotator Behavior

With the aid of the embedded control segments (taken from article bodies) and expected dialect label distribution, it was possible to spot spamming behavior and reject it. Table 3 shows three examples of workers whose work was rejected on this basis, having clearly demonstrated they are unable or unwilling to perform the task faithfully. In total, 11.4% of the assignments were rejected on this basis. In the approved assignments, the embedded MSA control sentence was annotated with the MSA label 94.4% of the time. In the remainder of this article, we analyze only data from the approved assignments.

We note here that we only rejected assignments where the annotator's behavior was *clearly* problematic, opting to *approve* assignments from workers mentioned later in Section 4.2.3, who exhibit systematic biases in their labels. Although these annotators' behavior is non-ideal, we cannot assume that they are not working faithfully, and therefore rejecting their work might not be fully justified. Furthermore, such behavior might be quite common, and it is worth investigating these biases to benefit future research.

Table 3
Some statistics over the labels provided by three spammers. Compared with the typical worker (right-most column), all workers perform terribly on the MSA control items, and also usually fail to recognize dialectal content in commentary sentences. Other red flags, such as geographic location and "identifying" unrepresented dialects, are further proof of the spammy behavior.

	A29V7OGM2C6205	A3SZLM2NK8NUOG	A8EF1I6CO7TCU	Typical
MSA in control items	0%	14%	33%	>90%
LEV in *Al-Ghad*	0%	0%	15%	25%
GLF in *Al-Riyadh*	8%	0%	14%	20%
EGY in *Al-Youm Al-Sabe'*	5%	0%	27%	33%
Other dialects	56%	0%	28%	<1%
Incomplete answers	13%	6%	1%	<2%
Worker location	Romania	Philippines	Jordan	Middle East
Claimed native dialect	Gulf	"Other"	Unanswered	(Various)

4.2.1 Label Distribution. Overall, 454 annotators participated in the task, 138 of whom completed at least 10 HITs. Upon examination of the provided labels for the commentary sentences, 40.7% of them indicate some level of dialect, and 57.1% indicate no dialectal content (Figure 8a). Note that 2.14% of the labels identify a sentence as being non-Arabic, non-textual, or as being left unanswered. The label breakdown is a strong confirmation of our initial motivation, which is that a large portion of reader commentary contains dialectal content.[10]

Figure 8 also illustrates the following:

- The most common dialectal label within a given news source matches the dialect of the country of publication. This is not surprising, since the readership for any newspaper is likely to mostly consist of the local population of that country. Also, given the newspapers' countries of publication, there is almost no content that is in a dialect other than Levantine, Gulf, or Egyptian. For this reason, other dialects such as Iraqi and Maghrebi, all combined, correspond to less than 0.01% of our data, and we mostly drop them from further discussion.
- The three news sources vary in the prevalence of dialectal content. The Egyptian newspaper has a markedly larger percentage of dialectal content (46.6% of labels) compared with the Saudi newspaper (40.1%) and the Jordanian newspaper (36.8%).
- A nontrivial amount of labels (5–8%) indicate `General` dialectal content. The `General` label was meant to indicate a sentence that is dialectal but lacks a strong indication of a *particular* dialect. Although many of the provided `General` labels seem to reflect an intent to express this fact, there is evidence that some annotators used this category in cases where choosing the label `Not sure` would have been more appropriate but was ignored (see Section 4.2.3).
- Non-Arabic content, although infrequent, is not a rare occurrence in the Jordanian and Egyptian newspapers, at around 3%. The percentage is much lower in the Saudi newspaper, at 0.8%. This might reflect the deeper penetration of the English language (and English-only keyboards) in Jordan and Egypt compared with Saudi Arabia.

We can associate a label with each segment based on the majority vote over the three provided labels for that segment. If a sentence has at least two annotators choosing a dialectal label, we label it as `dialect`. If it has at least two annotators choosing the MSA label, we label it as `MSA`.[11] In the remainder of the article, we will report classification accuracy rates that assume the presence of gold-standard class labels. Unless otherwise noted, this majority-vote label set is used as the gold-standard in such experiments.

[10] Later analysis in Section 4.2.3 shows that a non-trivial portion of the labels were provided by MSA-biased annotators, indicating that dialectal content could be even more prevalent than what is initially suggested by the MSA/dialect label breakdown.

[11] A very small percentage of sentences (2%) do not have such agreement; upon inspection these are typically found to be sentences that are in English, e-mail addresses, romanized Arabic, or simply random symbols.

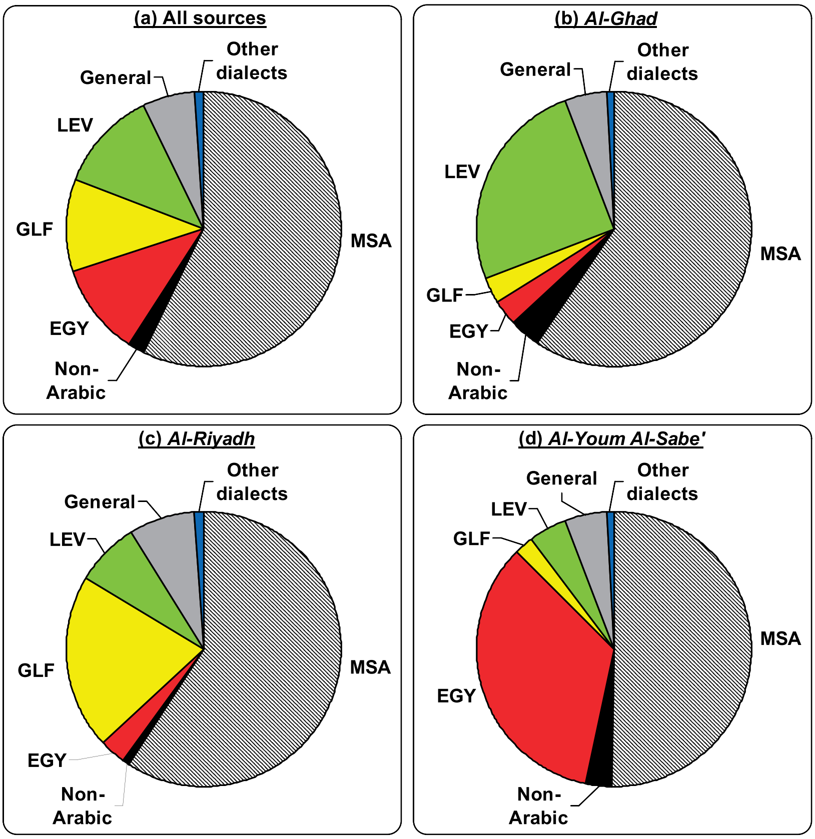


Figure 8
The distribution of labels provided by the workers for the dialect identification task, over all three news sources (a) and over each individual news source (b–d). *Al-Ghad* is published in Jordan, *Al-Riyadh* in Saudi Arabia, and *Al-Youm Al-Sabe'* in Egypt. Their local readerships are reflected in the higher proportion of corresponding dialects. Note that this is *not* a breakdown on the sentence level, and does *not* reflect any kind of majority voting. For example, most of the LEV labels on sentences from the Saudi newspaper are trumped by GLF labels when taking a majority vote, making the proportion of LEV-majority sentences smaller than what might be deduced by looking at the label distribution in (c).

In experiments where the dialectal label set is more fine-grained (i.e., LEV, GLF, and EGY instead of simply dialect), we assign to the dialectal sentence the label corresponding to the news source's country of publication. That is, dialectal sentences in the Jordanian (respectively, Saudi, Egyptian) are given the label LEV (respectively, GLF, EGY). We could have used dialect labels provided by the annotators, but chose to override those using the likely dialect of the newspaper instead. It turns out that sentences with an EGY majority, for instance, are extremely unlikely to appear in either the Jordanian or Saudi newspaper—only around 1% of those sentences have an EGY majority. In the case of the Saudi newspaper, 9% of all dialectal sentences were originally annotated as LEV but were transformed to GLF. Our rationales for performing the transformation is that no context was given for the sentences when they were annotated, and annotators had a bias towards their own dialect. We provide the original annotations for other researchers to re-analyze if they wish.

Table 4
The specific-dialect label distribution (given that a dialect label was provided), shown for each speaker group.

	Group size	% LEV	% GLF	% EGY	% GNRL	% Other dialects
All speakers	454	26.1	27.1	28.8	15.4	2.6
Levantine speakers	181	**35.9**	28.4	21.2	12.9	1.6
Gulf speakers	32	21.7	**29.4**	25.6	21.8	1.4
Egyptian speakers	121	25.9	19.1	**38.0**	10.9	6.1
Iraqi speakers	16	18.9	**29.0**	23.9	18.2	10.1
Maghrebi speakers	67	20.5	28.0	**34.5**	12.7	4.3
Other/Unknown	37	17.9	18.8	27.8	**31.4**	4.1

Even when a sentence *would* receive a majority-vote label that differs from the news source's primary dialect, inspection of such sentences reveals that the classification was usually unjustified, and reflected a bias towards the annotator's native dialect. Case in point: Gulf-speaking annotators were in relatively short supply, whereas a plurality of annotators spoke Levantine (see Table 4). Later in Section 4.2.3, we point out that annotators have a native-dialect bias, whereby they are likely to label a sentence with their native dialect even when the sentence has no evidence of being written in that particular dialect. This explains why a non-trivial number of LEV labels were given by annotators to sentences from the Saudi newspaper (Figure 8). In reality, most of these labels were given by Levantine speakers over-identifying their own dialect. Even if we were to assign dialect labels based on the (Levantine-biased) majority votes, Levantine would only cover 3.6% of the sentences from the Saudi newspaper.[12]

Therefore, for simplicity, we assume that a dialectal sentence is written in the dialect corresponding to the sentence's news source, without having to inspect the specific dialect labels provided by the annotators. This not only serves to simplify our experimental set-up, but also contributes to partially reversing the native dialect bias that we observed.

4.2.2 Annotator Agreement and Performance. The annotators exhibit a decent level of agreement with regard to whether a segment is dialectal or not, with full agreement (i.e., across all three annotators) on 72.2% of the segments regarding this binary dialect/MSA decision. This corresponds to a kappa value of 0.619 (using the definition of Fleiss [1971] for multi-rater scenarios), indicating very high agreement.[13] The full-agreement percentage decreases to 56.2% when expanding the classification from a binary decision to a fine-grained scale that includes individual dialect labels as well. This is still quite a reasonable result, since the criterion is somewhat strict: It does not include a segment labeled, say, {Levantine, Levantine, General}, though there is good reason to consider that annotators are in "agreement" in such a case.

[12] Note that the distributions in Figure 8 are on the label level, *not* on the sentence level.

[13] Although it is difficult to determine the significance of a given kappa value, Landis and Koch (1977) characterize kappa values above 0.6 to indicate "substantial agreement" between annotators.

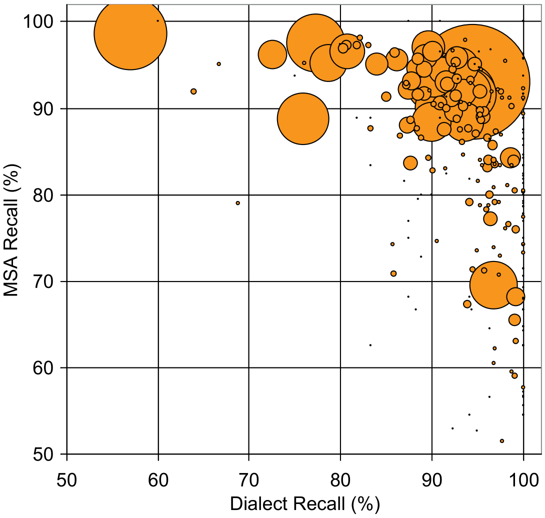


Figure 9
A bubble chart showing workers' MSA and dialect recall. Each data point (or bubble) in the graph represents one annotator, with the bubble size corresponding to the number of assignments completed by that annotator.

So how good are humans at the classification task? We examine their classification accuracy, dialect recall, and MSA recall. The classification accuracy is measured over all sentences, both MSA and dialectal. We define dialect (MSA) recall to be the number of sentences labeled as being dialectal (MSA), over the total number of sentences that have dialectal (MSA) labels based on the majority vote. Overall, human annotators have a classification accuracy of 90.3%, with dialect recall at 89.0%, and MSA recall at 91.5%. Those recall rates do vary across annotators, as shown in Figure 9, causing some accuracy rates to drop as low as 80% or 75%. Of the annotators performing at least five HITs, 89.4% have accuracy rates greater than or equal to 80%.

Most annotators have both high MSA recall and high dialect recall, with about 70% of them achieving at least 80% in both MSA and dialect recall. Combined with the general agreement rate measure, this is indicative that the task is well-defined—it is unlikely that many people would agree on something that is incorrect.

We note here that the accuracy rate (90.3%) is a slight overestimate of the human annotators' accuracy rate, by virtue of the construction of the gold labels. Because the correct labels are based on a majority vote of the annotators' labels themselves, the two sets are not independent, and an annotator is inherently likely to be correct. A more informative accuracy rate disregards the case where only two of the three annotators agreed and the annotator whose accuracy was being evaluated contributed one of those two votes. In other words, an annotator's label would be judged against a majority vote that is independent of that annotator's label. Under this evaluation set-up, the human accuracy rate slightly decreases, to 88.0%.

4.2.3 Annotator Bias Types. Examining the submitted labels of individual workers reveals interesting annotation patterns, and indicates that annotators are quite diverse in their

Table 5
Two annotators with a `General` label bias, one who uses the label liberally, and one who is more conservative. Note that in both cases, there is a noticeably smaller percentage of `General` labels in the Egyptian newspaper than in the Jordanian and Saudi newspapers.

	All workers	A1M50UV37AMBZ3	A2ZNK1PZOVIECD
% `General`	6.3	12.0	2.3
% `General` in *Al-Ghad*	5.2	14.2	3.1
% `General` in *Al-Riyadh*	7.7	13.1	2.6
% `General` in *Al-Youm Al-Sabe'*	4.9	7.6	1.0
Native dialect	(Various)	Maghrebi	Egyptian

behavior. An annotator can be observed to have one or more of the following bias types:[14]

- **MSA bias/dialect bias**: Figure 9 shows that annotators vary in how willing they are to label a sentence as being dialectal. Whereas most workers (top right) exhibit both high MSA and high dialect recall, other annotators have either a MSA bias (top left) or a dialect bias (bottom right).
- **Dialect-specific bias**: Many annotators over-identify a particular dialect, usually their native one. If we group the annotators by their native dialect and examine their label breakdown (Table 4), we find that Levantine speakers over-identify sentences as being Levantine, Gulf speakers over-identify Gulf, and Egyptian speakers over-identify Egyptian. This holds for speakers of other dialects as well, as they over-identify other dialects more often than most speakers. Another telling observation is that Iraqi speakers have a bias for the Gulf dialect, which is quite similar to Iraqi. Maghrebi speakers have a bias for Egyptian, reflecting their unfamiliarity with the geographically distant Levantine and Gulf dialects.
- **The `General` bias**: The `General` label is meant to signify sentences that cannot be definitively classified as one dialect over another. This is the case when enough evidence exists that the sentence is not in MSA, but contains no evidence for a specific dialect. In practice, some annotators make very little use of this label, even though many sentences warrant its use, whereas other annotators make extensive use of this label (see, for example, Table 5). One interesting case is that of annotators whose `General` label seems to mean they are unable to identify the dialect,

[14] These biases should be differentiated from *spammy* behavior, which we already can deal with quite effectively, as explained in Section 4.2.

> and a label like `Not sure` might have been more appropriate. Take the case of the Maghrebi worker in Table 5, whose `General` bias is much more pronounced in the Jordanian and Saudi newspapers. This is an indication she might have been having difficulty distinguishing Levantine and Gulf from each other, but that she is familiar with the Egyptian dialect.

5. Automatic Dialect Identification

From a computational point of view, we can think of dialect identification as language identification, though with finer-grained distinctions that make it more difficult than typical language ID. Even languages that share a common character set can be distinguished from each other at high accuracy rates using methods as simple as examining character histograms (Cavnar and Trenkle 1994; Dunning 1994; Souter et al. 1994), and, as a largely solved problem, the one challenge becomes whether languages can be identified for very short segments.

Due to the nature and characteristics and high overlap across Arabic dialects, relying on character histograms alone is ineffective (see Section 5.3.1), and more context is needed. We will explore higher-order letter models as well as word models, and determine what factors determine which model is best.

5.1 Smoothed *n*-Gram Models

Given a sentence S to classify into one of k classes $C_1, C_2, \ldots, C_k$, we will choose the class with the maximum conditional probability:

$$C^* = \underset{C_i}{\operatorname{argmax}}\, P(C_i|S) = \underset{C_i}{\operatorname{argmax}}\, P(S|C_i) \cdot P(C_i) \tag{1}$$

Note that the decision process takes into account the prior distribution of the classes, which is estimated from the training set. The training set is also used to train probabilistic models to estimate the probability of S given a particular class. We rely on training n-gram language models to compute such probabilities, and apply Kneser-Ney smoothing to these probabilities and also use that technique to assign probability mass to unseen or OOV items (Chen and Goodman 1998). In language model scoring, a sentence is typically split into words. We will also consider *letter*-based models, where the sentence is split into sequences of characters. Note that letter-based models would be able to take advantage of clues in the sentence that are not complete words, such as prefixes or suffixes. This would be useful if the amount of training data is very small, or if we expect a large domain shift between training and testing, in which case content words indicative of MSA or dialect might not still be valuable in the new domain.

Although our classification method is based only on language model scoring, and is thus relatively simple, it is nevertheless very effective. Experimental results in Section 5.3 (e.g., Figure 10) indicate that this method yields accuracy rates above 85%, only slightly behind the human accuracy rate of 88.0% reported in Section 4.2.2.

5.2 Baselines

To properly evaluate classification performance trained on dialectal data, we compare the language-model classifiers to two baselines that do not use the newly collected data.

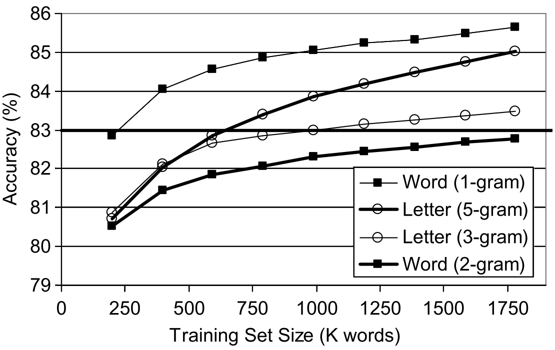


Figure 10
Learning curves for the general `MSA` vs. `dialect` task, with all three news sources pooled together. Learning curves for the individual news sources can be found in Figure 11. The 83% line has no significance, and is provided to ease comparison with Figure 11.

Rather, they use available MSA-only data and attempt to determine how MSA-like a sentence is.

The first baseline is based on the assumption that a dialectal sentence would contain a higher percentage of "non-MSA" words that cannot be found in a large MSA corpus. To this end, we extracted a vocabulary list from the Arabic Gigaword Corpus, producing a list of 2.9M word types. Each sentence is given a score that equals the OOV percentage, and if this percentage exceeds a certain threshold, the sentence is classified as being dialectal. For each of the cross validation runs in Section 5.3.1, we use the threshold that yields the optimal accuracy rate on the test set (hence giving this baseline as much a boost as possible). In our experiments, we found this threshold to be usually around 10%.

The second approach uses a more fine-grained approach. We train a language model using MSA-only data, and use it to score a test sentence. Again, if the perplexity exceeds a certain threshold, the sentence is classified as being dialectal. To take advantage of domain knowledge, we train this MSA model on the sentences extracted from the article bodies of the AOC, which corresponds to 43M words of highly relevant content.

5.3 Experimental Results

In this section, we explore using the collected labels to train word- and letter-based DID systems, and show that they outperform other baselines that do not utilize the annotated data.

5.3.1 Two-Way, MSA vs. Dialect Classification. We measure classification accuracy at various training set sizes, using 10-fold cross validation, for several classification tasks. We examine the task both as a general `MSA` vs. `dialect` task, as well as when restricted within a particular news source. We train unigram, bigram, and trigram (word-based) models, as well as unigraph, trigraph, and 5-graph (letter-based) models. Table 6 summarizes the accuracy rates for these models, and includes rates for the baselines that do not utilize the dialect-annotated data.

Generally, we find that a unigram word model performs best, with a 5-graph model slightly behind (Figure 11). Bigram and trigram word models seem to suffer from the sparseness of the data and lag behind, given the large number of parameters they

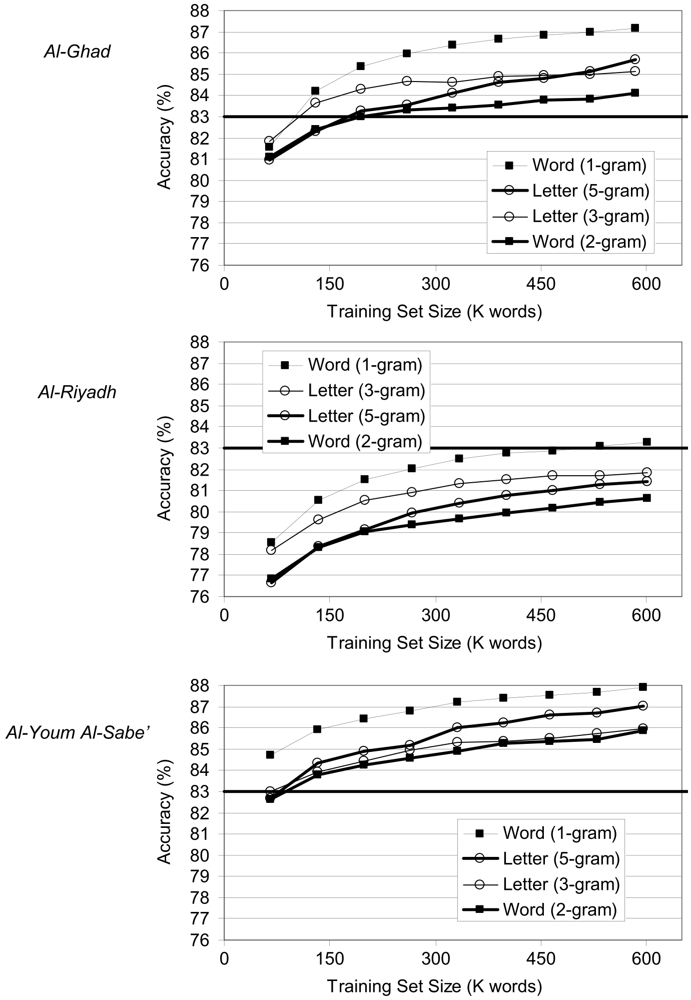


Figure 11
Learning curves for the `MSA` vs. `dialect` task, for each of the three news sources. The 83% line has no significance, and is provided to ease comparison across the three components, and with Figure 10.

would need to estimate (and instead resort to smoothing heavily). The letter-based models, with a significantly smaller vocabulary size, do not suffer from this problem, and perform well. This is a double-edged sword though, especially for the trigraph model, as it means the model is less expressive and converges faster.

Overall though, the experiments show a clear superiority of a supervised method, be it word- or letter-based, over baselines that use existing MSA-only data. Whichever model we choose (with the exception of the unigraph model), the obtained accuracy rates show a significant dominance over the baselines.

It is worth noting that a classification error becomes less likely to occur as the length of the sentence increases (Figure 12). This is not surprising given prior work on the language identification problem (Řehůřek and Kolkus 2009; Verma, Lee, and Zakos 2009), which points out that the only "interesting" aspect of the problem is performance on short segments. The same is true in the case of dialect identification: a short sentence

Table 6
Accuracy rates (%) on several two-way classification tasks (`MSA` vs. `dialect`) for various models. Models in the top part of the table do not utilize the dialect-annotated data, whereas models in the bottom part do. (For the latter kind of models, the accuracy rates reported are based on a training set size of 90% of the available data.)

Model	`MSA` vs. `dialect`	*Al-Ghad* `MSA` vs. `dialect` (Levantine)	*Al-Riyadh* `MSA` vs. `dialect` (Gulf)	*Al-Youm Al-Sabe'* `MSA` vs. `dialect` (Egyptian)
Majority Class	58.8	62.5	60.0	51.9
OOV % vs. Gigaword	65.5	65.1	65.3	66.7
MSA LM-scoring	66.6	67.8	66.8	65.2
Letter-based, 1-graph	68.1	69.9	68.0	70.4
Letter-based, 3-graph	83.5	85.1	81.9	86.0
Letter-based, 5-graph	85.0	85.7	81.4	87.0
Word-based, 1-gram	**85.7**	**87.2**	**83.3**	**87.9**
Word-based, 2-gram	82.8	84.1	80.6	85.9
Word-based, 3-gram	82.5	83.7	80.4	85.6

that contains even a single misleading feature is prone to misclassification, whereas a long sentence is likely to have other features that help identify the correct class label.[15]

One could also observe that distinguishing MSA from dialect is a more difficult task in the Saudi newspaper than in the Jordanian paper, which in turn is harder than in the Egyptian newspaper. This might be considered evidence that the Gulf dialect is the closest of the dialects to MSA, and Egyptian is the farthest, in agreement with the conventional wisdom. Note also that this is not due to the fact that the Saudi sentences tend to be significantly shorter—the ease of distinguishing Egyptian holds even at higher sentence lengths, as shown by Figure 12.

5.3.2 Multi-Way, Fine-Grained Classification. The experiments reported earlier focused on distinguishing MSA from dialect when the news source is known, making it straightforward to determine which of the Arabic dialects a sentence is written in (once

[15] The accuracy curve for the Egyptian newspaper has an outlier for sentence lengths 10–12. Upon inspection, we found that over 10% of the sentences in that particular length subset were actually repetitions of a single 12-word sentence. (A disgruntled reader, angry about perceived referee corruption, essentially bombarded the reader commentary section of several articles with that single sentence.) This created an artificial overlap between the training and test sets, hence increasing the accuracy rate beyond what would be reasonably expected due to increased sentence length alone.

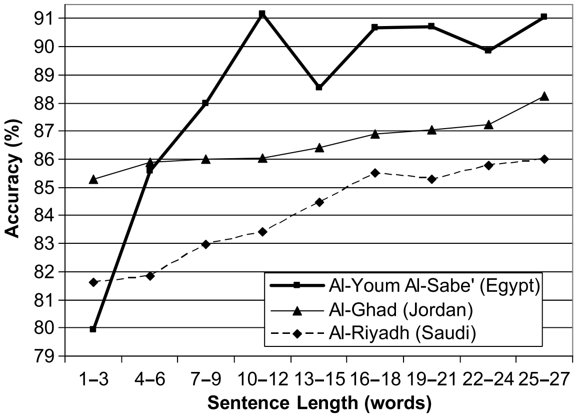


Figure 12
Accuracy rates vs. sentence length in the general MSA vs. dialect task. Accuracy rates shown are for the unigram word model trained on 90% of the data.

the sentence is determined to be dialectal). If the news source is *not* known, we do not have the luxury of such a strong prior on the specific Arabic dialect. It is therefore important to evaluate our approach in a multi-way classifiation scenario, where the class set is expanded from {MSA, dialect} to {MSA, LEV, GLF, EGY}.

Under this classification set-up, the classification accuracy decreases from 85.7% to 81.0%.[16] The drop in performance is not at all surprising, since four-way classification is inherently more difficult than two-way classification. (Note that the classifier is trained on exactly the same training data in both scenarios, but with more fine-grained dialectal labels in the four-way set-up.)

Table 7 is the classifier's confusion matrix for this four-way set-up, illustrating when the classifier tends to make mistakes. We note here that most classification errors on dialectal sentences occur when these sentences are mislabeled as being MSA, not when they are misidentified as being in some other incorrect dialect. In other words, dialect→dialect confusion constitutes a smaller proportion of errors than dialect→MSA confusion. Indeed, if we consider a three-way classification setup on dialectal sentences alone (LEV vs. GLF vs. EGY), the classifier's accuracy rate shoots up to 88.4%. This is a higher accuracy rate than for the general two-way MSA vs. dialect classification (85.7%), despite involving more classes (three instead of two), and being trained on less data (0.77M words instead of 1.78M words). This indicates that the dialects deviate from MSA in various ways, and therefore distinguishing dialects from each other can be done even more effectively than distinguishing dialect from MSA.

5.3.3 Word and Letter Dialectness. Examining the letter and word distribution in the corpus provides valuable insight into what features of a sentence are most dialectal. Let *DF*(*w*) denote the *dialectness factor* of a word *w*, defined as:

$$DF(w) \overset{\text{def}}{=} \frac{f(w|D)}{f(w|MSA)} = \frac{count_D(w)/count_D(.)}{count_{MSA}(w)/count_{MSA}(.)} \tag{2}$$

[16] For clarity, we report accuracy rates only for the unigram classifier. The patterns from Section 5.3.1 mostly hold here as well, in terms of how the different *n*-gram models perform relative to each other.

Table 7
Confusion matrix in the four-way classification setup. Rows correspond to actual labels, and columns correspond to predicted labels. For instance, 6.7% of MSA sentences were given a `GLF` label (first row, third column). Note that entries within a single row sum to 100%.

Class label	`MSA`	`LEV`	`GLF`	`EGY`
`MSA` Sentences	**86.5%**	4.2%	6.7%	2.6%
`LEV` Sentences	20.6%	**69.1%**	8.6%	1.8%
`GLF` Sentences	24.2%	2.4%	**72.0%**	1.4%
`EGY` Sentences	14.4%	2.2%	4.6%	**78.8%**

where $count_D(w)$ (respectively, $count_{MSA}(w)$) is the number of times w appeared in the dialectal (respectively, MSA) sentences, and $count_D(.)$ is the total number of words in those sentences. Hence, $DF(w)$ is simply a ratio measuring how much more likely w is to appear in a dialectal sentence than in an MSA sentence. Note that the dialectness factor can be easily computed for letters as well, and can be computed for bigrams/bigraphs, trigrams/trigraphs, and so forth.

Figure 13 lists, for each news source, the word types with the highest and lowest dialectness factor. The most dialectal words tend to be function words, and they also tend to be *strong* indicators of dialect, judging by their very high *DF*. On the other hand, the MSA word group contains several content words, relating mainly to politics and religion.

One must also take into account the actual frequency of a word, as *DF* only captures relative frequencies of dialect/MSA, but does not capture how often the word occurs in the first place. Figure 14 plots both measures for the words of *Al-Ghad* newspaper. The

Al-Ghad				*Al-Riyadh*				*Al-Youm Al-Sabe'*			
w		Gloss	*DF* (*w*)	*w*		Gloss	*DF* (*w*)	*w*		Gloss	*DF* (*w*)
$w	شو	what	139.0	*Ay$*	ايش	what	247.7	*AwY*	اوى	very	116.6
xly	خلي	let	132.8	*w$*	وش	what	105.9	*dy*	دي	this (f.)	108.8
xlS	خلص	enough	117.5	*lyn*	لين	until	92.9	*glT*	غلط	wrong	106.6
AlHky	الحكي	the-talk	115.8	*xl*	خل	let	83.8	*dlwqtY*	دلوقتى	now	75.7
Endw	عندو	he has	95.3	*ly$*	ليش	why	82.4	*E$An*	عشان	so that	70.2
bdy	بدي	I will/want	93.6	*jAy*	جاي	coming	72.2	*mAfy$*	مافيش	none	65.7
AzA	ازا	if	93.6	*ybwn*	يبون	they will/want	69.7	*dY*	دى	this (f.)	64.5
mnyH	منيح	good	93.6	*blA$*	بلاش	lest	65.8	*tAny*	تاني	another/again	62.3
$wy	شوي	little	92.8	*El$An*	علشان	so that	64.5	*dh*	ده	this (m.)	61.4
<nw	إنو	that	90.2	*lyh*	ليه	why	48.6	*Antw*	انتو	you (pl.)	59.8
hAy	هاي	this (f.)	80.0	*wbs*	وبس	that's all	46.0	*AntwA*	انتوا	you (pl.)	59.3
bEdyn	بعدين	then	70.2	*yby*	يبي	he will/wants	44.4	*AntA*	انتا	you (s.)	58.8
mw	مو	not	65.5	*$wy*	شوي	little	43.9	*jAY*	جاى	coming	58.8
Ay$	ايش	what	63.8	*mArAH*	ماراح	will not	39.6	*EAwz*	عاوز	I want	57.8
bdw	بدو	he will/wants	60.3	*wyn*	وين	where	38.8	*El$An*	علشان	so that	57.5
	⋮				⋮				⋮		
Ebr	عبر	through	0.146	*<lyh*	إليه	to him	0.154	*<lY*	إلى	to	0.133
w>n	وأن	and-that	0.145	*w>n*	وأن	and-that	0.149	*AlmsyH*	المسيح	Christ	0.125
Al<slAm	الإسلام	Islam	0.138	*rswl*	رسول	messenger	0.131	*<dArp*	إدارة	management	0.125
tEAlY	تعالى	almighty	0.138	*ns>l*	نسأل	we ask	0.130	*flmA*A*	فلماذا	so-why	0.113
SlY	صلى	blessed	0.127	*fymA*	فيما	whilst	0.127	*$y}A*	شيئا	(any)thing	0.111
AldymqrATyp	الديمقراطية	democratic	0.108	*y>ty*	يأتي	comes	0.127	*AlfADl*	الفاضل	esteemed	0.092
Alljnp	اللجنة	the-committee	0.095	*tEAlY*	تعالى	almighty	0.122	*ljmAl*	لجمال	to-Jamal	0.090
f<n	فإن	(declarative)	0.062	*fmn*	فمن	who	0.117	*Al>stA**	الأستاذ	mister	0.078
AlmfAwDAt	المفاوضات	the-negotiations	0.038	*tlk*	تلك	that (f.)	0.105	*<lyh*	إليه	to-him	0.055
AlmbA$rp	المباشرة	the-direct	0.029	*lqd*	لقد	(declarative)	0.090	*lldktwr*	للدكتور	to-the-doctor	0.051

Figure 13
Words with the highest and lowest dialectness factor values in each of the three news sources.

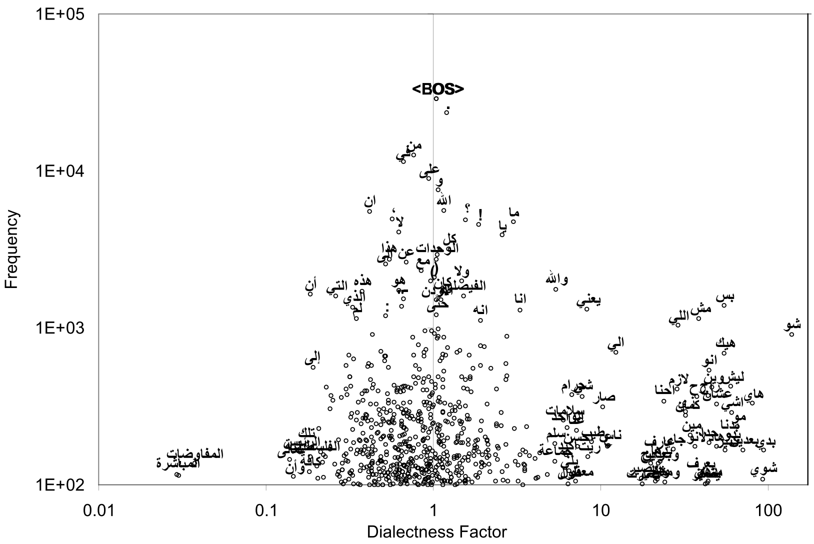


Figure 14
A plot of the most common words in the *Al-Ghad* sentences, showing each word's *DF* and corpus frequency. The right- and left-most words here also appear in Figure 13. Not every word from that list appears here though, since some words have counts below 100. For clarity, not all points display the words they represent.

plot illustrates which words are most important to the classifier: the words that are farthest away from the point of origin, along both dimensions.

As for letter-based features, many of the longer ones (e.g., 5-graph features) are essentially the same words important to the unigram word model. The letter-based models are, however, able to capture some linguistic phenomenon that the word model is unable to: the suffixes *+š* (*not* in Levantine) and *+wn* (plural conjugation in Gulf), and the prefixes *H+* (*will* in Egyptian), *bt+* (present tense conjugation in Levantine and Egyptian), and *y+* (present tense conjugation in Gulf).

Figure 15 sheds some light on why even the unigraph model outperforms the baselines. It picks up on subtle properties of the MSA writing style that are lacking

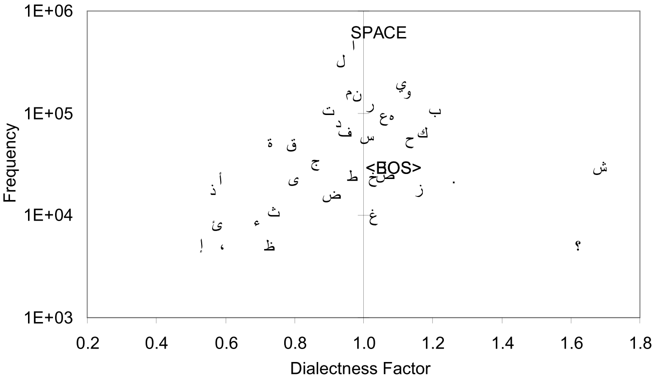


Figure 15
A plot of the most common letters in the *Al-Ghad* sentences, showing each letter's *DF* and corpus frequency.

when using dialect. Namely, there is closer attention to following *hamza* rules (distinguishing *A*, *Â*, and *Ǎ* from each other, rather than mapping them all to *A*), and better adherence to (properly) using +*ħ* instead of +*h* at the end of many words. There is also a higher tendency to use words containing the letters that are most susceptible to being transformed when pronounced dialectally: ð (usually pronounced as *z*), *Ď* (pronounced as *D*), and θ (pronounced as *t*).

On the topic of spelling variation, one might wonder if nomalizing the Arabic text before training language models might enhance coverage and therefore improve performance. For instance, would it help to map all forms of the *alef hamza* to a single letter, and all instances of *ħ* to *h*, and so on? Our pilot experiments indicated that such normalization tends to slightly but consistently hurt performance, so we opted to leave the Arabic text as is. The only type of preprocessing we performed was more on the "cleanup" side of things rather than computationally motivated normalization, such as proper conversion of HTML entities (e.g., `&quot;` to `"`) and mapping Eastern Arabic numerals to their European equivalents.

6. Applying DID to a Large-Scale Arabic Web Crawl

We conducted a large-scale Web crawl to gather Arabic text from the on-line versions of newspapers from various Arabic-speaking countries. The first batch contained 319 online Arabic-language newspapers published in 24 countries. This list was compiled from `http://newspapermap.com/` and `http://www.onlinenewspapers.com/`, which are Web sites that show the location and language of newspapers published around the world. The list contained 55 newspapers from Lebanon, 42 from Egypt, 40 from Saudi Arabia, 26 from Yemen, 26 from Iraq, 18 from Kuwait, 17 from Morocco, 15 from Algeria, 12 from Jordan, and 10 from Syria. The data were gathered from July–Sept 2011.

We mirrored the 319 Web sites using `wget`, resulting in 20 million individual files and directories. We identified 3,485,241 files that were likely to contain text by selecting the extensions htm, html, cmff, asp, pdf, rtf, doc, and docx. We converted these files to text using xpdf's pdftotext for PDFs and Apple's textutil for HTML and Doc files. When concatenated together, the text files contained 438,940,861 lines (3,452,404,197 words). We performed de-duplication to remove identical lines, after which 18,219,348 lines (1,393,010,506 words) remained.

We used the dialect-annotated data to train a language model for each of the four Arabic varieties (`MSA`, `LEV`, `GLF`, `EGY`), as described in the previous section. We used these models to classify the crawled data, assigning a given sentence the label corresponding to the language model under which that sentence received the highest score. Table 8 gives the resulting label breakdown. We see that the overwhelming majority of the sentences are classified as MSA, which comes as no surprise, given the prevalence of MSA in the newspaper genre. Figure 16 shows some sentences that were given non-MSA labels by our classifier.

7. Related Work

Habash et al. (2008) presented annotation guidelines for the identification of dialectal content in Arabic content, paying particular attention to cases of code switching. They present pilot annotation results on a small set of around 1,600 Arabic sentences (19k words), with both sentence- and word-level dialectness annotations.

Table 8
Predicted label breakdown for the crawled data, over the four varieties of Arabic. All varieties were given equal priors.

Variety	Sentence Count	Percentage
MSA	13,102,427	71.9%
LEV	3,636,525	20.0%
GLF	630,726	3.5%
EGY	849,670	4.7%
ALL	18,219,348	100.0%

The Cross Lingual Arabic Blog Alerts (COLABA) project (Diab et al. 2010) is another large-scale effort to create dialectal Arabic resources (and tools). They too focus on online sources such as blogs and forums, and use information retrieval tasks to measure their ability to properly process dialectal Arabic content. The COLABA project demonstrates the importance of using dialectal content when training and designing tools that deal with dialectal Arabic, and deal quite extensively with resource creation and data harvesting for dialectal Arabic.

AR (LEV): فيصل القاسم (مقاطعا) : طيب جميل هذا الكلام ، خلينا بموضوع التدويل

TL: *fySl AlqAsm (mqATςA) : Tyb jmyl hðA AlklAm , xlynA bmwDwς Altdwyl*

EN: Faisal Al-Qasem (interrupting) : OK that is very nice , let us stay on the topic of internationalization

AR (LEV): ناس فاضيه بدل ماتحسن من اوضاع المدرسين والخدمات للطلبه

TL: *nAs fADyh bdl matHsn mn AwDAς Almdrsyn wAlxdmAt llTlbh*

EN: Such empty-headed people this is instead of improving the conditions of teachers and services for students

AR (GLF): ليش ماتسون لهم حلبات مثل باقي الدول ؟؟؟

TL: *lyš mA tswn lhm HlbAt mθl bAqy Aldwl ???*

EN: Why not make tracks for them like other countries do ???

AR (GLF): عادي ناس واجد الحين يموتون وينقتلون ويتذبحون يوميا بالالاف

TL: *ςAdy nAs wAjd AlHyn ymwtwn wynqtlwn wytðbHwn ywmyA bAlAlAf*

EN: This is normal I now see people die and are killed and slaughtered daily by the thousands

AR (EGY): أسامة الغزالي حرب : أنت في الحزب الوطني ولا إيه يا عم أحمد ؟

TL: *ÂsAmħ AlγzAly Hrb : Ânt fy AlHzb AlwTny wlA Ǎyh yA ςm ÂHmd ?!*

EN: Osama Al-Ghazali Harb : are you in the National Party or what mister Ahmad ?

AR (EGY): انا عن نفسي منعرفهاش

TL: *AnA ςn nfsy mnςrfhaš*

EN: I myself do not know her

Figure 16
Example sentences from the crawled data set that were predicted to be dialectal, two in each of the three Arabic dialects.

Chiang et al. (2006) investigate building a parser for Levantine Arabic, *without* using any significant amount of dialectal data. They utilize an available Levantine–MSA lexicon, but no parses of Levantine sentences. Their work illustrates the difficulty of adapting MSA resources for use in a dialectal domain.

Zbib et al. (2012) show that incorporating dialect training data into a statistical machine translation system vastly improves the quality of the translation of dialect sentences when compared to a system trained solely on an MSA-English parallel corpus. When translating Egyptian and Levantine test sets, a dialect Arabic MT system outperforms a Modern Standard Arabic MT system trained on a 150 million word Arabic–English parallel corpus—over 100 times the amount of data as their dialect parallel corpus.

As far as we can tell, no prior dialect identification work exists that is applied to Arabic text. However, Lei and Hansen (2011) and Biadsy, Hirschberg, and Habash (2009) investigate Arabic dialect identification in the *speech* domain. Lei and Hansen (2011) build Gaussian mixture models to identify the same three dialects we consider, and are able to achieve an accuracy rate of 71.7% using about 10 hours of speech data for training.

Biadsy, Hirschberg, and Habash (2009) utilize a much larger data set (170 hours of speech data) and take a phone recognition and language modeling approach (Zissman 1996). In a four-way classification task (with Iraqi as a fourth dialect), they achieve a 78.5% accuracy rate. It must be noted that both works use *speech* data, and that dialect identification is done on the *speaker* level, not the sentence level as we do.

8. Conclusion

Social media, like reader commentary on on-line newspapers, is a rich source of dialectal Arabic that has previously not been studied in detail. We have harvested this type of resource to create a large data set of informal Arabic that is rich in dialectal content. We selected a large subset of this data set, and had the sentences in it manually annotated for dialect. We used the collected labels to train and evaluate automatic classifiers for dialect identification, and observed interesting linguistic aspects about the task and annotators' behavior. Using an approach based on language model scoring, we develop classifiers that significantly outperform baselines that use large amounts of MSA data, and we approach the accuracy rates exhibited by human annotators.

In addition to *n*-gram features, one could imagine benefiting from morphological features of the Arabic text, by incorporating analyses given by automatic analyzers such as BAMA (Buckwalter 2004), MAGEAD (Habash and Rambow 2006), ADAM (Salloum and Habash 2011), or CALIMA (Habash, Eskander, and Hawwari 2012). Although the difference between our presented approach and human annotators was found to be relatively small, incorporating additional linguistically motivated features might be pivotal in bridging that final gap.

In future annotation efforts, we hope to solicit more detailed labels about dialectal content, such as specific annotation for *why* a certain sentence is dialectal and not MSA: Is it due to structural differences, dialectal terms, and so forth? We also hope to expand beyond the three dialects discussed in this article, by including sources from a larger number of countries.

Given the recent political unrest in the Middle East (2011), another rich source of dialectal Arabic are Twitter posts (e.g., with the `#Egypt` tag) and discussions on various political Facebook groups. Here again, given the topic at hand and the individualistic nature of the posts, they are very likely to contain a high degree of dialectal data.

Appendix A

The Arabic transliteration scheme used in the article is the Habash-Soudi-Buckwalter transliteration (HSBT) mapping (Habash, Soudi, and Buckwalter 2007), which extends the scheme designed by Buckwalter in the 1990s (Buckwalter 2002). Buckwalter's original scheme represents Arabic orthography by designating a single, distinct ASCII character for each Arabic letter. HSBT uses some non-ASCII characters for better readibility, but maintains the distinct 1-to-1 mapping.

Figure 17 lists the character mapping used in HSBT. We divide the list into four sections: vowels, forms of the *hamzah* (glottal stop), consonants, and pharyngealized

	ASCII	Arabic	Pronunciation Guide
	A	ا	The vowel '**a**' (e.g. f**a**ther or c**a**t)
→	*ħ*	ة	The vowel '**a**' (only appears at word's end, e.g. Al-Manam**ah**)
→	*ý*	ى	The vowel '**a**' (only appears at word's end, e.g. Mon**a**)
	w	و	The vowel '**o**' (e.g. h**o**me, s**oo**n), or the consonant '**w**' (e.g. **w**ait)
	y	ي	The vowel '**e**' (e.g. t**ee**n, r**ai**n), or the consonant '**y**' (e.g. **y**es)
	'	ء	Various forms of the Arabic letter *hamzah* , which is the glottal stop (the consonantal sound in '**uh**-**oh**', and the allophone of '**t**' in some pronunciations of bu**tt**on). Determining which form is appropriate depends on the location of the *hamzah* within the word, and the vowels immediately before and after it.
	Ā	آ	
	Â	أ	
	ŵ	ؤ	
	Ǎ	إ	
	ŷ	ئ	
→	*š*	ش	**sh**oe
→	*ð*	ذ	**th**e
	b	ب	**b**a**b**y
	d	د	**d**a**d**
	f	ف	**f**ather
→	*γ*	غ	French Pa**r**is (guttural)
	H	ح	a raspier version of '**h**' (IPA: voiceless pharyngeal fricative)
	h	ه	**h**ouse
	j	ج	**j**ump or bei**g**e
	k	ك	**k**iss
	l	ل	**l**eaf
	m	م	**m**o**m**
	n	ن	**n**u**n**
	q	ق	like a '**k**' further back in the throat (IPA: voiceless uvular stop)
	r	ر	Scottish bo**rr**ow (rolled)
	s	س	**s**un
	t	ت	**t**en
→	*θ*	ث	**th**ink
→	*x*	خ	German Ba**ch**, Spanish o**j**o
	z	ز	**z**ebra
	D	ض	Pharyngealized '**d**'
→	*ς*	ع	Pharyngealized glottal stop (IPA: voiced pharyngeal fricative)
	S	ص	Pharyngealized '**s**'
	T	ط	Pharyngealized '**t**'
→	*Ď*	ظ	Pharyngealized '**th**' (of **th**e)

Figure 17
The character mapping used in the HBST scheme. Most mappings are straightforward; a few non-obvious mappings are highlighted with an arrow (→) next to them. For brevity, the mappings for short vowels and other diacritics are omitted. Note that we take the view that ς is a pharyngealized glottal stop, which is supported by Gairdner (1925), Al-Ani (1970), Kästner (1981), Thelwall and Sa'Adeddin (1990), and Newman (2002). For completeness, we indicate its IPA name as well.

consonants. Pharyngealized consonants are "thickened" versions of other, more familiar consonants, voiced such that the pharynx or epiglottis is constricted during the articulation of the sound. Those consonants are present in very few languages and are therefore likely to be unfamiliar to most readers, which is why we place them in a separate section—there is no real distinction in Arabic between them and other consonants.

HSBT also allows for the expression of short vowels and other Arabic diacritics, but because those diacritics are only rarely expressed in written (and typed) form, we omit them for brevity.

Acknowledgments

This research was supported in part by the DARPA GALE program under contract no. HR0011-06-2-0001, the DARPA BOLT program contract no. HR0011-12-C-0014, the EuroMatrixPlus project funded by the European Commission (7th Framework Programme), the Human Language Technology Center of Excellence, and by gifts from Google and Microsoft. The views and findings are the authors' alone. They do not reflect the official policies or positions of the Department of Defense or the U.S. Government. The authors would like to thank the anonymous reviewers for their extremely valuable comments on earlier drafts of this article, and for suggesting future work ideas.

References

Abdel-Massih, Ernest T., Zaki N. Abdel-Malek, and El-Said M. Badawi. 1979. *A Reference Grammar of Egyptian Arabic*. Georgetown University Press.

Al-Ani, Salman H. 1970. *Arabic Phonology: An Acoustical and Physiological Investigation*. Mouton.

Aoun, Joseph, Elabbas Benmamoun, and Dominique Sportiche. 1994. Agreement, word order, and conjunction in some varieties of Arabic. *Linguistic Inquiry*, 25(2):195–220.

Badawi, El-Said and Martin Hinds. 1986. *A Dictionary of Egyptian Arabic*. Librairie du Liban.

Bassiouney, Reem. 2009. *Arabic Sociolinguistics*. Edinburgh University Press.

Biadsy, Fadi, Julia Hirschberg, and Nizar Habash. 2009. Spoken Arabic dialect identification using phonotactic modeling. In *Proceedings of the EACL Workshop on Computational Approaches to Semitic Languages*, pages 53–61, Athens.

Buckwalter, Tim. 2002. Buckwalter Arabic transliteration. `http://www.qamus.org/transliteration.htm`.

Buckwalter, Tim. 2004. Buckwalter Arabic morphological analyzer version 2.0. Linguistic Data Consortium, Philadelphia, PA.

Callison-Burch, Chris and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12, Los Angeles, CA.

Cavnar, William B. and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94*, pages 161–175, Vilnius.

Chen, Stanley F. and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

Chiang, David, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of EACL*, pages 369–376, Trento.

Cowell, Mark W. 1964. *A Reference Grammar of Syrian Arabic*. Georgetown University Press.

Diab, Mona, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. COLABA: Arabic dialect annotation and processing. In *Proceedings of the LREC Workshop on Semitic Language Processing*, pages 66–74.

Dunning, T. 1994. Statistical identification of language. Technical Report MCCS 94-273, New Mexico State University.

Erwin, Wallace. 1963. *A Short Reference Grammar of Iraqi Arabic*. Georgetown University Press.

Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gairdner, William Henry Temple. 1925. *The Phonetics of Arabic*. Oxford University Press.

Garera, Nikesh and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of ACL*, pages 710–718, Singapore.

Habash, Nizar. 2008. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In *Proceedings of ACL, Short Papers*, pages 57–60, Columbus, OH.

Habash, Nizar, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 711–718, Istanbul.

Habash, Nizar, Ramy Eskander, and Abdelati Hawwari. 2012. A morphological analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal.

Habash, Nizar and Owen Rambow. 2006. MAGEAD: A morphological analyzer and generator for the Arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney.

Habash, Nizar, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of Arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic World*, pages 49–53, Marrakech.

Habash, Nizar, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic transliteration. In Antal van den Bosch, Abdelhadi Soudi, and Günter Neumann, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Kluwer/Springer Publications, chapter 2.

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

Haeri, Niloofar. 2003. *Sacred Language, Ordinary People: Dilemmas of Culture and Politics in Egypt*. Palgrave Macmillan.

Holes, Clive. 2004. *Modern Arabic: Structures, Functions, and Varieties*. Georgetown Classics in Arabic Language and Linguistics. Georgetown University Press.

Ingham, Bruce. 1994. *Najdi Arabic: Central Arabian*. John Benjamins.

Kästner, Hartmut. 1981. *Phonetik und Phonologie des modernen Hocharabisch*. Verlag Enzyklopädie.

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Lei, Yun and John H. L. Hansen. 2011. Dialect classification via text-independent training and testing for Arabic, Spanish, and Chinese. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):85–96.

Mitchell, Terence Frederick. 1990. *Pronouncing Arabic*. Clarendon Press.

Mohand, Tilmatine. 1999. Substrat et convergences: Le berbére et l'arabe nord-africain. *Estudios de Dialectologiá Norteaafricana y andalusí*, 4:99–119.

Newman, Daniel L. 2002. The phonetic status of Arabic within the world's languages. *Antwerp Papers in Linguistics*, 100:63–75.

Novotney, Scott, Rich Schwartz, and Sanjeev Khudanpur. 2011. Unsupervised Arabic dialect adaptation with self-training. In *Interspeech*, pages 541–544, Florence.

Řehůřek, Radim and Milan Kolkus. 2009. *Language Identification on the Web: Extending the Dictionary Method*, volume 5449 of *Lecture Notes in Computer Science*, pages 357–368. SpringerLink.

Salloum, Wael and Nizar Habash. 2011. Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the EMNLP Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh.

Shlonsky, Ur. 1997. *Clause Structure and Word Order in Hebrew and Arabic: An Essay in Comparative Semitic Syntax*. Oxford University Press.

Souter, Clive, Gavin Churcher, Judith Hayes, John Hughes, and Stephen Johnson. 1994. Natural language identification using corpus-based models. *Hermes Journal of Linguistics*, 13:183–203.

Suleiman, Yasir. 1994. Nationalism and the Arabic language: A historical overview. In Yasir Suleiman, editor, *Arabic Sociolinguistics*. Curzon Press.

Thelwall, Robin and M. Akram Sa'Adeddin. 1990. Arabic. *Journal of the International Phonetic Association*, 20(2):37–39.

Verma, Brijesh, Hong Lee, and John Zakos. 2009. *An Automatic Intelligent Language Classifier*, volume 5507 of *Lecture Notes in Computer Science*, pages 639–646. SpringerLink.

Versteegh, Kees. 2001. *The Arabic Language*. Edinburgh University Press.

Zaidan, Omar, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the*

North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 260–267, Rochester, NY.

Zaidan, Omar F. 2012. *Crowdsourcing Annotation for Machine Learning in Natural Language Processing Tasks*. Ph.D. thesis, Johns Hopkins University, Baltimore, MD.

Zaidan, Omar F. and Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: An annotated dataset of informal Arabic with high dialectal content. In *Proceedings of ACL*, pages 37–41, Portland, OR.

Zbib, Rabih, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In the *2012 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 49–59, Montreal.

Zissman, Marc A. 1996. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1):31–44.