3arif: A Corpus of Modern Standard and Egyptian Arabic Tweets Annotated for Epistemic Modality Using Interactive Crowdsourcing

Rania Al-Sabbagh[†], **Roxana Girju**[†], **Jana Diesner**[‡]
[†]Department of Linguistics and Beckman Institute
[‡]School of Library and Information Science
University of Illinois at Urbana-Champaign, USA
{alsabba1, girju, jdiesner} @illinois.edu

Abstract

We present *3arif*[1], a large-scale corpus of Modern Standard and Egyptian Arabic tweets annotated for epistemic modality[2]. To create *3arif*, we design an interactive crowdsourcing annotation procedure that splits up the annotation process into a series of simplified questions, dispenses with the requirement for expert linguistic knowledge and captures nested modality triggers and their attributes semi-automatically.

1 Introduction

Epistemic modality, according to Palmer (2001), defines the speaker's subjective knowledge, beliefs and judgments about the world's states of affairs. Epistemic modality is used as a linguistic feature for multiple NLP tasks and applications, including sentiment analysis (Abdul-Mageed and Diab 2011), opinion mining (Benamara et al. 2012) and scientific discourse evaluation (Waard and Maat 2012), among others.

To-date, there are no large-scale modality-annotated Arabic corpora compared to English (Baker et al. 2010, 2012; Rubinstein et al. 2013), Chinese (Cui and Chi 2013), Portuguese (Hendrickx et al. 2012) and Japanese (Matsuyoshi et al. 2010). The creation of modality-annotated corpora is non-trivial because there is no consensus definition of modality and its attributes in theoretical linguistics to be rendered into annotation tasks and guidelines. Furthermore, most current modality annotation schemes rely on sophisticated theoretically-grounded guidelines that require annotators from linguistics background; hence, annotation is usually restricted to small-scale in-lab settings.

In this paper, we present *3arif*, a large-scale Arabic corpus annotated for epistemic modality. *3arif* comprises 9822 unique tweets in Modern Standard Arabic (MSA) and Egyptian Arabic (EA), annotated for 9966 tokens that map to 214 unique types of epistemic modality. Each epistemic modality is annotated for sense, polarity, intensification, tense, holder(s) and scope(s). The reason that *3arif* features the tweets' genre with an emphasis on MSA and EA tweets is that it comes as part of a larger project to incorporate linguistic features, such as modality, with network-based features to automatically identify the key players of Twitter's political discourse in counties of political unrest such as Egypt. We harvested *3arif* from a variety of Twitter users including newspapers, TV stations, political campaigns, among others, as well as individuals. As a result *3arif* is diglossic for MSA, the formal Arabic variety, and EA, the native Arabic dialect of Egypt.

For the annotation of *3arif*, we design a simplified procedure that depicts the following ideas: first, it defines each annotation task as a series of open and closed questions that do not require sophisticated linguistics background and, meanwhile, provide annotators with self-explanatory annotation guidelines; second, it is interactive so that questions are displayed/hidden based on annotators' prior answers; and finally, it semi-automatically identifies and merges nested epistemic modality based on annotators' answers to a number of easy-to-administer questions.

[1] Pronounced as *ʕa:rif* in Arabic IPA and as *EArif* in Buckwalter's transliteration scheme. It means *I/he know(s)*.
[2] *3arif* is available at http://www.rania-alsabbagh.com/3arif.html

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1521–1532, Dublin, Ireland, August 23-29 2014.

We evaluate our annotation results using Krippendorff's reliability (Krippendorff 2011) and agreement. Results show high inter-annotator reliability and agreement rates and indicate that our annotation scheme and procedure are efficient. The contribution of this research, therefore, is twofold: first, we create a novel resource for Arabic NLP which is expected to enhance research on modality automatic identification and extraction; second, we present an efficient and easy-to-administer annotation procedure with interactive crowdsourcing potentials for the complex task of modality annotation.

The rest of this paper is organized as follows: Section 2 outlines our annotation scheme including annotation tasks, guidelines and the interactive structure; Section 3 gives examples for the representation of the final annotation outputs; Section 4 describes corpus harvesting and sampling; Section 5 discusses the results and presents a disagreement analysis; Section 6 compares and contrasts our work to related work; and Section 7 highlights the points not covered in this current version of *3arif*.

2 Annotation Scheme

Our annotation scheme consists of six tasks to label sense, polarity, intensification, tense, holders and scopes for each epistemic modality. Prior to the beginning of the interactive annotation procedure, we highlighted all candidate epistemic modalities in each tweet using a string-match algorithm and the lexicons from Al-Sabbagh et al. (2013, 2014). The algorithm finds all potential epistemic modality triggers (i.e. words and phrases that may convey epistemic modality) within each tweet in our corpus and marks them as annotation units. A total of 9966 candidate epistemic modality triggers are highlighted in 9822 tweets.

2.1 Task 1: Sense

Sense annotation is to decide for each highlighted candidate trigger in context whether it actually conveys epistemic modality. The same lexical verb اشعر *A$Er* is used as an epistemic modality trigger anticipating a future possibility in example 1; but as a non-modal lexical verb in example 2.

1. [3] **اشعر** ان[نا سنكسر رقم ال30 مليون متظاهر]
A$Er *An[nA snksr rqm Al30 mlywn mtZAhr]*
I **feel** that [we will get 30+ million protesters].

2. #هيكل: **اشعر** بالفخر والقلق أيضا في ذكرى حرب أكتوبر
#hykl: ***A$Er*** *bAlfxr wAlqlq >yDA fy *krY Hrb >ktwbr*
#Heikl: I **feel** proud but also worried when I remember October's war.

We define sense annotation as a synonymy judgment task, following Al-Sabbagh et al. (2013). Epistemic modality is represented by an exemplar set manually selected so that: (1) each exemplar is an unambiguous epistemic trigger, (2) exemplars are in both MSA and EA, (3) exemplars comprise both simple words and multiword expressions, (4) exemplars are both affirmative and negative, and (5) exemplars are of different lexical intensities. Furthermore, we create multiple versions of the same set so that we cover the inflections for gender, number, person, tense, mood, and aspect in Arabic. We then use the set that morphologically matches the candidate trigger to be annotated. Presented with a pre-highlighted candidate trigger in context and the exemplar set, annotators are to decide whether the given candidate trigger is synonymous to the exemplar set, and is hence an epistemic modality trigger, or not.

If an annotator decides that a given candidate trigger does not convey epistemic modality, no further questions about polarity, intensification, tense, holders or scopes are displayed. To guarantee that annotators do not select the non-synonymous option as an easy escape, they are not allowed to move forward without submitting at least one synonym of their own to the candidate trigger.

Designing the interactive procedure as such results in disagreement propagation. If one annotator decides that a given candidate trigger is not epistemic, but another annotator decides that it is, the former will not have to answer any further questions about polarity, intensification, tense, holders or scopes; whereas the latter will have to provide answers for each of those annotation tasks.

[3] Throughout the examples, epistemic modality triggers are represented in boldface and scopes are in-between square brackets.

2.2 Task 2: Polarity

Task 2 uses as input the candidates labeled as valid epistemic modality triggers in Task 1 and labels each as either affirmative or negative. An affirmative trigger indicates that the speaker holds the given state of affairs (i.e. propositions) as TRUE; whereas a negative trigger indicates that the given propositions are held as FALSE by the speaker.

To decide on whether the polarity is affirmative or negative, annotators are instructed to look for the absence/presence of such negation markers as:

- **Negation particles** such as مش *m$* (not), لا *lA* (not) and غير *gyr* (not), among others.
- **Negation affixes** like the circumfix *m...$* in مظنش *mZn$* (I do not think).
- **Negative polarity items** like عمري *Emry* (never) and لم يعد *lm yEd* (no longer).
- **Negative auxiliaries** where negation is placed on the past tense auxiliary as in مكنتش واثق *mknt$ wAvq* (I was not sure).
- **Inherently-negative triggers** that encode negation in their lexical meanings such as مستحيل *mstHyl* (impossible).

Annotators are instructed that using multiple negation markers results in an affirmative sense. Thus, ليس من المستحيل *lys mn AlmstHyl* (it is not impossible) means that the proposition is actually possible according to the speaker. Put differently, it means that the speaker holds the proposition as TRUE. Annotators are required to give the reason for negation if they decide that a given trigger is negative.

2.3 Task 3: Intensification

Epistemic modality triggers can have different lexical intensities (i.e. intensities encoded in the lexical meaning of the word/phrase regardless of the context). For instance, even without a context, Arabic speakers know that متأكد *mt>kd* (I am/he is sure) expresses higher possibility than متهيألي *mthy>ly* (I imagine). When used in context, the trigger's lexical intensity can be maintained as is. Yet, it can also be amplified or mitigated by various linguistic means such as:

- **Modification:** adverbs like تماما *tmAmA* (absolutely) and بالفعل *bAlfEl* (indeed), among others, amplify lexical intensity; whereas mitigation can be caused by such adverbs as تقريبا *tqrybA* (almost) and غالبا *gAlbA* (most probably), among others.
- **Categorical negation** typically amplifies lexical intensity as in مش ممكن أبدا *m$ mmkn >bdA* (it is not possible at all).
- **Emphatic expressions** such as قد *qd* (indeed) and والله *wAllh* (I swear), among others, lead to lexical intensity amplification.
- **Coordination** of two or more triggers usually results in intensity amplification as in عارف ومتأكد *EArf wmt>kd* (I know and I am sure).

The annotators' task for intensification is to decide for each candidate labeled as a valid epistemic modality trigger in Task 1 whether its lexical intensity is amplified (AMP), mitigated (MTG), or maintained (AS IS). During interactive annotation, annotators are asked to provide the reason for their selection; that is, whether the lexical intensity is affected by an adverb, categorical negation, an emphatic expression, coordination, or any other reason.

2.4 Task 4: Tense

In this version of *3arif*, we work on the present and past tenses only. Thus, Task 4 is to decide for each valid epistemic trigger from Task 1 whether it is present (PRS) or past (PST). Tense can be marked either morphologically by inflections and affixes or contextually by auxiliary verbs such as كان *kAn* (was), among others. Annotators are also required to give their reasons for selecting either PRS or PST.

2.5 Task 5: Holder

Holder annotation is to identify the holder of the epistemic modality which is the ±RATIONAL entity that expresses its knowledge, beliefs or judgments about the world's states of affairs.

Holders can be −RATIONAL entities as in example 3. The entity that is making the assumption that the former Palestinian president - Yasser Arafat - may have died of natural causes is the report issued by the French government.

3. [تقرير فرنسي: [وفاة #عرفات **ربما** تعود لاسباب طبيعية]
*tqryr frnsy: [wfAp #ErfAt **rbmA** tEwd lAsbAb TbyEyp]*
A French report: [natural causes **might** be behind the death of #Arafat].

The holder is not necessarily the same as the trigger's grammatical subject. In example 4, the grammatical subject of يبدو *ybdw* (seems) is الاعلان الدستوري *AlAElAn Aldstwry* (the constitutional declaration). However, the entity that is making the judgment about this declaration is the French government, which is then the real holder of *ybdw*.

4. [فرنسا: [الاعلان الدستوري الجديد لمرسي لا **يبدو** انه يسلك الاتجاه الصحيح]
*frnsA: [AlAElAn Aldstwry Aljdyd lmrsy lA **ybdw** Anh yslk AlAtjAh AlSHyH]*
France: [Morsi's new constitutional declaration does not **seem** to be a correct move].

Twitter users do not only post their own knowledge, beliefs and judgments about the world's states of affairs, but also they (1) directly and indirectly quote others and (2) make assumptions about others' knowledge, beliefs and judgments. This means that we can have nested holders, according to Wiebe et al. (2005) and Saurí and Pustejovsky (2009), where we know about others' knowledge, beliefs and judgments only though the writer or the Twitter user in our case.

In example 5, the Twitter user quotes Elbaradei stating that he may run for presidency if the people want him to. That is, the holder of the epistemic modality is actually Elbaradei not the Twitter user.

5. البرادعي: **قد** [أترشح في انتخابات الرئاسة] إذا طلب الشعب
*AlbrAdEy: **qd** [>tr$H fy AntxAbAt Alr}Asp] <*A Tlb Al$Eb*
Elbaradei: I **may** [run for presidency] if the people want me to.

The holder of the epistemic modality in example 6 is not the Twitter user, either. However, the Twitter user is not quoting anyone here, but is rather making an assumption about what the Egyptian National Party holds as TRUE.

6. [الحزب الوطني **مقتنع** ان[ه ممكن يرجع] #Jan25
*AlHzb AlwTny **mqtnE** An[h mmkn yrjE] #Jan25*
The National Party is **convinced** that [it may get back to authority]. #Jan25

We can have two or more nested holders. In example 5, we have two: the first is ElBaradei and the second is the Twitter user who is quoting ElBaradei. Similarly, in example 6, we have two nested holders: the first is the Egyptian National Party and the second is the Twitter user who makes the assumptions about the party's beliefs.

In example 7, however, we have three nested holders. The first is الاخوان *AlAxwAn* (the Muslim Brotherhood) that holds as TRUE the proposition that the Military Council is conspiring against them. That belief of the Muslim Brotherhood is communicated to us through the politician ابو الفتوح *Abw AlftwH* (Abulfotoh) who is then the second holder. Yet, Abulfotoh has not posted his assumption about the Muslim Brotherhood's belief on his personal account. Instead, he has been quoted by another Twitter user, who is the third holder.

7. [ابو الفتوح: الاخوان **تصوروا** ان [هناك مؤامرة من العسكري]
*Abw AlftwH: AlAxwAn **tSwrwA** An [hnAk m&Amrp mn AlEskry]*
Abulfotoh: The Muslim Brotherhood members **thought** that [there was a conspiracy by the Military Council].

During the interactive procedure, annotators are first asked whether the holder is the same as the Twitter user. If not, more questions are displayed to determine: (1) who the real holder is; (2) whether the tweet is a(n) (in)direct quote (e.g. there are direct quotation markers or such words as قال *qAl* (he said) and صرح *SrH* (he declared), among others), or the tweet conveys the Twitter user's assumptions about others.

When the holder is not the same as the Twitter user, annotators are asked to mark the boundaries of the linguistic unit that corresponds to the holder in the tweet's text, following the maximal length principle from Szarvas et al. (2008), so that they mark the largest possible, meaningful linguistic unit. Hence, in example 8 the holder is *the Islamist opponents in #KSA* not only *the Islamist opponents*.

8. [الإسلاميون المعارضون في #السعودية **موقنون** أن[ها تسعى لقتل الثورة في #مصر]
*Al<slAmywn AlmEArDwn fy #AlsEwdyp **mwqnwn** >n[hA tsEY lqtl Alvwrp fy #mSr]*
Islamist opponents in #KSA **know for sure** that [it tries to put an end to #Egypt's revolution].

2.6 Task 6: Scope

Scopes are the states of affairs modified by the epistemic modality triggers. Modality scopes in Arabic are most likely realized as clauses, deverbal nouns or to-infinitives, according to Al-Sabbagh et al. (2013). We use the same maximal length guideline from Task 5 so that the scope segment marked by the annotators is the largest possible segment typically delimited by: (1) punctuation markers and (2) subordinate conjunctions such as لان *lAn* (because) and لو *lw* (if), among others.

In the case of nested triggers as in example 9, where a trigger and its scope are both embedded in another trigger's scope, the interactive procedure prompts the annotators to label each trigger and its scope separately at first. Afterwards, we automatically merge them as we further explain in Section 3.

9. [[الحزب الوطني **مقتنع** أن[ه **ممكن** [يرجع]] #Jan25
*AlHzb AlwTny **mqtnE** >n[h **mmkn** [yrjE]] #Jan25*
The National Party is **convinced** that [it **may** [get back to power]] #Jan25

Annotators are instructed that a single trigger may have one or more scopes. In example 10, the trigger بيتهيألهم *bythy>lhm* (they imagine) scopes over two complement clauses, which annotators are required to identify. Furthermore, annotators are given the guideline that two or more triggers - typically conjoined by a coordinating conjunction - can share the same scope as in example 11. In the cases like example 11, each trigger and its attributes are first annotated separately and then once our system finds out that they share the same polarity, intensification, tense, holder, and scope, they are merged together as we show in Section 3.

10. [أولادنا **بيتهيألهم** ان [دم اخواتهم راح هدر] وان[هم عندهم ثأر مع السلطة بكل أشكالها]
*>wlAdnA **bythy>lhm** An [dm AxwAthm rAH hdr] wAn[hm Endhm v>r mE AlslTp bkl >$kAlhA]*
Our children **imagine** that [their friends were killed for no reason] and that [they now have to take revenge from the authorities].

11. البرادعي **عارف ومتاكد** ان [نسبة 12 % بس هتنتخبه] وعلشان كدة مش هيرشح نفسه
*AlbrAdEy **EArf wmtAkd** An [nsbp 12% bs htntxbh] wEl$An kdp m$ hyr$H nfsh*
Elbaradei **knows and is sure** that [only 12% will vote for him]. So, he will not run for presidency.

Annotators are instructed that scopes are not necessarily adjacent to their triggers. In example 12, the scope starts three words to the right of its trigger باقتنع *bAqtnE* (get convinced) given that the adverbial phrase اكتر واكتر *Aktr wAktr* (more and more) falls in between it and its scope.

12. [كل يوم بيعدي **باقتنع** اكتر واكتر ان[نا كنا محتاجيين دكتاتور وطني عادل]
*kl ywm byEdy **bAqtnE** Aktr wAktr An[nA knA mHtAjyn dktAtwr wTny EAdl]*
Every day, I **get more and more convinced** that [we needed a patriotic and fair dictator].

Annotators are also instructed that scopes can (1) precede, (2) follow or (3) surround their triggers. Many of the aforementioned examples have the scopes following their triggers. Yet, in example 13 the scope surrounds its trigger and in example 14 it precedes its trigger.

13. [وعود مرسي ليست **فيما يبدو** دين عليه]
*[wEwd mrsy lyst **fymA ybdw** dyn Elyh]*
[Morsi's promises are not **seemingly** doable].

14. **فيما يبدو** [حملة تشويه ثورة يناير وإعادة عقارب الساعة تماما إلى الوراء بدأت]
*[Hmlp t$wyh vwrp ynAyr w<EAdp EqArb AlsAEp tmAmA <lY AlwrA' bd>t] **fymA ybdw***

[A campaign to distort the image of January's revolution and to restore everything back to its original state has started], **seemingly**.

3 Final Output Representation

All elicited answers during annotation are automatically organized into the representations illustrated in the examples below. The representation of example 15 reads as follows: the USER (i.e. the Twitter user) used to moderately hold as TRUE the proposition that the revolutionist candidates were unable to compete for presidency. We know that this is a past belief that the USER used to have because annotators have labeled the trigger تصورت *tSwrt* (I thought) as past (PST). There are no nested holders given that the USER is the same as the holder. The intensity value of MODerate comes from the fact that تصورت *tSwrt* (I thought) is of a moderate lexical intensity being weaker than such epistemic triggers as متاكد *mtAkd* (I am sure) and عارف *EArf* (I know) but stronger than such epistemic triggers as اظن *AZn* (I guess) and متهيالي *mthyAly* (I imagine). Meanwhile, the lexical intensity of *tSwrt* is neither amplified nor mitigated; hence annotators have given it an AS IS intensification label in Task 3. Consequently, in the final annotation output the original lexical intensity value has been used to represent how far the holder used to consider his/her belief as TRUE.

15. [في البداية **تصورت** ان [مرشحي الثورة اضعف من المنافسة للرئاسة]
*fy AlbdAyp **tSwrt** An [mr$Hy Alvwrp ADEf mn AlmnAfsp llr}Asp]*
At first, I **thought** that [the revolutionist candidates are too weak to compete for presidency].

rep. USER, MOD PST TRUE, (*mr$Hy Alvwrp ADEf mn AlmnAfsp llr}Asp*)

Example 16 shows how two epistemic modality triggers in the same tweet are given two separate representations because they share the same holder but neither the same intensity nor the same scopes. The first representation illustrates the epistemic trigger ارى *ArY* (I think) and reads as follows: the USER currently holds as TRUE the proposition that the media is misleading the people; s/he is MODerately confident about that. The second representation is for the epistemic trigger واضح *wADH* (obviously). It indicates that the same USER strongly holds as TRUE the proposition that the media is trying to stop the change that the people are longing for. Both triggers are labeled as present (PRS) tense. Furthermore, both triggers are labeled as maintaining their lexical intensity AS IS. The trigger ارى *ArY* (I think) is then labeled in the final representation as being of MODerate intensity because it is weaker than متاكد *mtAkd* (I am sure), for instance, but stronger than متهيالي *mthyAly* (I imagine); whereas the trigger واضح *wADH* (obviously) is labeled as indicating a strong (STRG) belief being synonymous to متاكد *mtAkd* (I am sure) and اعرف *AErf* (I know) among other triggers that express speakers' high confidence about their knowledge, beliefs and judgments.

16. [ارى ان [الاعلام يقدم شباب يخدرون الشعب] **واضح** ان[هم يقاومون التغيير الذى نطمح له]
***ArY** An [AlAElAm yqdm $bAb yxdrwn Al$Eb] **wADH** An[hm yqAwmwn Altgyyr Al*y nTmH lh]*
I **think** [the media presents young speakers who mislead the people]. **Obviously**, [they are resisting the change we are longing for].

rep1. USER, MOD PRS TRUE, (*AlAElAm yqdm $bAb yxdrwn Al$Eb*)
rep2. USER, STRG PRS TRUE, (*hm yqAwmwn Altgyyr Al*y nTmH lh*)

Example 17 illustrates how two coordinating epistemic triggers sharing the same polarity, tense, intensification, holder and scope are represented. They are simply merged in one representation. The same example shows how assumptions made by Twitter users about others' knowledge, beliefs and judgments are represented. The representation reads as follows: the USER MODerately holds as TRUE the proposition that Elbaradei strongly (STRG) holds as TRUE that only 12% of the Egyptians will vote for him for presidency. The values of TRUE, MODerate and present (PRS) assigned to the USER's assumption about Elbaradei are default values used to mark Twitter users' assumptions about others' knowledge, beliefs and judgments.

17. البرادعى **عارف ومتاكد** ان [نسبة 12 % بس هتنتخبه] وعلشان كدة مش هيرشح نفسه
*AlbrAdEy **EArf wmtAkd** An [nsbp 12% bs htntxbh] wEl$An kdp m$ hyr$H nfsh*
Elbaradei **knows and is sure** that [only 12% will vote for him]. So, he will not run for presidency.

rep. USER, MOD PRS TRUE, (*AlbrAdEy*, STRG PRS TRUE, (*nsbp 12% bs htntxbh*))

Example 18 represents an epistemic trigger with multiple scopes. The example also represents Twitter users making assumptions about others' knowledge, beliefs and judgments. As we mentioned in example 17, the values of TRUE, MODerate and present (PRS) assigned to the USER's assumption are assigned by default. The trigger بيتهيألهم *bythy>lhm* (they imagine) is labeled as a present (PRS) tense affirmative trigger. Its original lexical intensity - which is weak (WK) - is labeled as being maintained AS IS. The trigger بيتهيألهم *bythy>lhm* (they imagine) is of a weak lexical intensity because it is weaker than متاكد *mtAkd* (I am sure) and even اظن *AZn* (I think).

18. [أولادنا بيتهيألهم ان [دم اخواتهم راح هدر] وان[هم عندهم ثأر مع السلطة بكل أشكالها]
*>wlAdnA **bythy>lhm** An [dm AxwAthm rAH hdr] wAn[hm Endhm v>r mE AlslTp bkl >$kAlhA]*
Our children **imagine** that [their friends were killed for no reason] and that [they now have to take revenge from the authorities].

rep. USER, MOD PRS TRUE, (*>wlAdnA,* WK PRS TRUE ,(*dm AxwAthm rAH hdr; hm Endhm v>r mE AlslTp bkl >$kAlhA*))

Example 19 illustrates embedded triggers. Its representation reads as: the USER MODerately holds as TRUE that the Egyptian National Party strongly (STRG) holds as TRUE that it (i.e. the Egyptian National Party) may get back to ruling. It is important to notice that both the matrix trigger مقتنع *mqtnE* (is convinced) and the embedded trigger (i.e. ممكن *mmkn* (may)) share the same holder which is the Egyptian National Party.

19. [[الحزب الوطني **مقتنع** ان[ه **ممكن** [يرجع]] #Jan25
*AlHzb AlwTny **mqtnE** An[h mmkn [yrjE]] #Jan25*
The National Party is **convinced** that [it **may** [get back to power]].

rep. USER, MOD PRS TRUE, (*AlHzb AlwTny*, STRG PRS TRUE,(MOD PRS TRUE, (*yrjE*)))

Example 20 shows how reported knowledge, beliefs and judgments are represented. The USER in this example has no other role but to report Darrag's strong belief that the army will interfere to stop the chaos.

20. دراج: [#الجيش **حتما** سيتدخل في حالة الفوضى] #مصر #مرسي #الاخوان
*drAj: [#Aljy$ **HtmA** sytdxl fy HAlp AlfwDY] #mSr #mrsy #AlAxwAn*
Darrag: [the #army will **definitely** interfere in the case of chaos] #Egypt #Morsi #Ikhwan

rep. USER, report, (*drAj*, STRG PRS TRUE (*#Aljy$ sytdxl fy HAlp AlfwDY*))

4 Corpus Harvesting

In order to restrict our corpus to political discourse and ensure that we compile a representative corpus of epistemic modality, we harvested our corpus so that each tweet (1) has at least one trendy political English or Arabic hashtag such as #Egypt and #مرسي *mrsy* (Morsi)[4], and (2) has at least one epistemic modality trigger from the Arabic Modality Lexicons of Al-Sabbagh et al. (2013, 2014). Table 1 gives statistics for the sampled corpus that comprises 9822 unique tweets, with 9966 candidate epistemic modality triggers that map to 214 unique types.

	Tokens	**Types**
Epistemic candidates	9966	214
All words	175964	47696

Table 1: Statistics for the sampled corpus

5 Annotation Results

5.1 Evaluation Methodology and Metrics

Our annotation tasks are of two types: (1) Tasks 1-4 are label-based where there is a pre-defined set of labels from which annotators choose; and (2) Tasks 5-6 are segmentation-based where the output of the annotation is a text segment. For the segmentation-based tasks, we use an all-or-nothing method to

[4] A total of 304 unique English and Arabic hashtags are found in the sampled corpus.

measure reliability and agreement: for segments to be considered as agreement, they must share both the beginning and end boundaries. We use Krippendorff's alpha α (Krippendorff 2011) as our inter-annotator reliability measure, following the most recent work on modality annotation for other languages including English (Rubinstein et al. 2013) and Chinese (Cui and Chi 2013). For more details on Krippendorff's alpha and a comparison of inter-annotator agreement measures, we refer the reader to Artstein and Poesio (2008).

5.2 Results

We use the surveygizmo services to implement our interactive annotation procedure given that their survey structure is one that allows for using conditional branching and skip logic[5]. We distributed the survey on Twitter and we had three annotators participating. According to the short qualifying quiz given at the beginning of the survey, all three participants are native Egyptian Arabic (EA) speakers who have at least two-year experience with using Twitter. They are also university graduates who, therefore, master Modern Standard Arabic. None of the participants has a linguistics background.

Table 2 shows alpha and agreement rates for each annotation task. We measure the rates in four different scenarios so that we can (1) estimate the effect of the inclusion of the NON-EPISTEMIC category agreement, (2) estimate the effect of disagreement propagation from Task 1, and (3) evaluate the guidelines and procedures for each annotation task separately. The four scenarios are:

- **w/NONE w/DP:** candidates agreed upon as non-epistemic and disagreement propagating from Task 1 are both included.
- **w/NONE w/o DP:** candidates agreed upon as non-epistemic are included, but disagreement propagating from Task 1 is excluded.
- **w/o NONE w/DP:** candidates agreed upon as non-epistemic are excluded, but disagreement propagating from Task 1 is included.
- **w/o NONE w/o DP:** candidates agreed upon as non-epistemic and disagreement propagating from Task 1 are both excluded. This scenario focuses on each annotation task separately without any distractions.

		Alpha				**Agreement**			
		w/NONE		**w/o NONE**		**w/NONE**		**w/o NONE**	
Annotation Task		**w/ DP**	**w/o DP**	**w/ DP**	**w/o DP**	**w/ DP**	**w/o DP**	**w/ DP**	**w/o DP**
1	Sense	--	0.899	--	--	--	0.949	--	--
2	Polarity	0.904	0.974	0.798	0.949	0.939	0.983	0.895	0.976
3	Intensification	0.880	0.942	0.658	0.768	0.926	0.966	0.844	0.939
4	Tense	0.911	0.995	0.772	0.983	0.947	0.997	0.909	0.994
5	Holder	0.878	0.930	0.672	0.727	0.933	0.956	0.884	0.969
6	Scope	0.825	0.916	0.620	0.618	0.899	0.955	0.819	0.911

Table 2: Inter-annotator alpha reliability and agreement rates

In the case of Task 1 (i.e. sense annotation), only the second scenario is applicable: we cannot exclude the candidates agreed upon as non-epistemic because the target is to know how reliable the annotation is with regards to distinguishing between epistemic and non-epistemic candidates. It is the first annotation task, thus there is no prior disagreement propagation. From Table 2, we derive the following observations:

- Disagreement in Task 1 propagates ~ 0.05 to 0.1 disagreement for the other annotation tasks.
- Adding the agreed upon non-epistemic candidates yields up to ~ 0.2 gain for both alpha reliability and agreement rates.
- For an end-to-end automatic system that first identifies triggers and then their attributes, the benchmark rates are those from the w/NONE w/DP scenario.

[5] http://www.surveygizmo.com/

5.3 Discussion and Disagreement Analysis

Among the factors that lead to high inter-annotator alpha reliability and agreement rates are that: (1) the vast majority of negation is explicitly marked by negation particles that are easy to detect by human annotators; (2) the vast majority of triggers are used without any amplification or mitigation markers; and (3) punctuation markers are surprisingly informative for marking scope boundaries and direct quotations and, hence, holders.

Sense-related disagreement is attributed to: (1) nominal triggers with main grammatical functions, (2) stative triggers, (3) opinionated-evidential triggers and (4) highly-polysemous triggers.

The majority of epistemic triggers are adjunct constituents that add an extra-layer of meaning and can be removed without disturbing the syntactic structure of their propositions. Yet, in example 21, احتمال *AHtmAl* (a possibility) is the grammatical subject of the proposition it modifies. Most of the exemplars from Section 2.1 are adjuncts and, thus, none can be both a lexical and a grammatical substitute for احتمال *AHtmAl* (a possibility) in such a context.

21. **احتمال** ان [رئيس منتخب يحل المجلس اثناء صياغة دستور جديد] **احتمال** وهمي
AHtmAl *An [r}ys mntxb yHl Almjls AvnA' SyAgp dstwr jdyd]* ***AHtmAl*** *whmy*
The **possibility** that [an elected president dissolves the parliament during the constitution's write-up] is an unrealistic **possibility**.

Stative triggers such as يعرف *yErf* (he knows) and يدرك *ydrk* (he realizes) invoke disagreement as to whether they indicate the acquisition of new information; that is, they literally mean *perceive*, or they mark confirmed beliefs as in *be sure that*. For example 22, the annotators have two interpretations: (1) a non-modal interpretation that *whoever says so does not perceive that the Supreme Guide cannot make resolutions without the Brotherhood*, and (2) a modal interpretation that *whoever says so does not believe that the Supreme Guide cannot make resolutions without the Brotherhood.*

22. [الذي يقول هذا الكلام لا **يعرف** ان [المرشد لا يستطيع اخذ قرار دون الرجوع الى الجماعة]
*Al*y yqwl h*A AlklAm lA* ***yErf*** *An [Almr$d lA ystTyE Ax* qrAr dwn AlrjwE AlY AljmAEp].*
Whoever says so does not **perceive/believe** that [the Supreme Guide cannot make resolutions without the Brotherhood].

Opinionated-evidential triggers like يزعم *yzEm* (he claims) do not only mark reported speech, but also they communicate the reporter's own opinion about the truth value of the reported proposition. They entail that from the reporter's perspective the proposition is FALSE. Hence, annotators disagree as to whether *yzEm* and similar triggers should be labeled as epistemic or not. We have eventually excluded such triggers as epistemic and have included them as evidential triggers for another corpus that is left for a future publication.

Highly-polysemous triggers like يمكن *ymkn* (can/possible) lead to disagreement because in many cases even the context is ambiguous. In example 23, both interpretations of *it is not possible that* (epistemic) and *it is not doable that* (abilitive) seem to be acceptable.

23. [لا **يمكن** [فهم كتاب مرسي "ثائر من الشرق" الا بتامل الكتابين المجاورين: "سرقات صغيرة" و"جنون الحكم"]
lA ***ymkn*** *[fhm ktAb mHmd mrsy "vA}r mn Al$rq" AlA btAml AlktAbyn AlmjAwryn: "srqAt Sgyrp" w "jnwn AlHkm"*]
It is **not possible/doable** [to understand Morsi's book - *A Revolutionist from the East* - without reading the other two books of *Small Robberies* and *Ruling Mania*].

Intensity-related disagreement is attributed to (1) intensity on the holder that propagates to the trigger and (2) negation with moderate-intensity triggers. In example 24, the USER uses categorical negation on the holder لا يوجد اي انسان عاقل *lA ywjd Ay AnsAn EAql* (there is no one sane person). For some annotators, the power of categorical negation spreads to the trigger, moving its intensity up the scale. As for negation with moderate-intensity triggers, some annotators think that لا يمكن *lA ymkn* (not possible) is synonymous to *impossible*. Hence, they consider the negation as an amplification marker.

24. لا يوجد أي انسان عاقل **يعتقد** بأن [الارهاب يعالج بالسياسة]
lA ywjd >y AnsAn EAql ***yEtqd*** *b>n [AlArhAb yEAlj bAlsyAsp]*
There is no one sane person who **thinks** that [terrorism can be defeated through politics].

Polarity-related disagreement is mainly caused by negation due to (1) negated holders and (2) contextual negation. Negated holders as in example 24 perplex the annotators as to whether the negation scopes over the holder only or both the holder and the trigger. Thus, for some annotators, يعتقد *yEtqd* (he thinks) is affirmative; and for others it is negative. By contextual negation we mean using words such as المشكلة *Alm$klp* (the problem) to describe triggers as in example 25. The USER says that *the problem is to think that it is a small-scale conflict.* To describe this as a *problem* means that the USER thinks of the proposition as FALSE; that is, according to the USER it is actually a large-scale conflict.

25. المشكلة إننا **نتصور** إن [الصراع محصور فى الدائرة الضيقة اللى بنتحرك فيها]
*Alm$klp <nnA **ntSwr** <n [AlSrAE mHSwr fY AldA}rp AlDyqp AllY bntHrk fyhA]*
The problem is to **think** that [the conflict is only happening at this small-scale we are working on].

Holder-related disagreement is attributed mainly to generic nouns and impersonal pronouns such as الشعب *Al$Eb* (the people) and الواحد *AlwAHd* (one). Some annotators interpret them as implicitly referring to the USER. Therefore, they select the USER as the only holder with zero nesting in example 26. Other annotators interpret them as referring to people in general but not necessarily with the USER included; and thus, they select two-level nested holders.

26. الشعب **يعرف** ان [الممارسة الديمقراطية هى التي ستأتي باعضاء مجلس الشعب والرئيس القادم]
*Al$Eb **yErf** An [AlmmArsp AldymwqrATyp hy Alty st>ty bAEDA' mjls Al$Eb wAlr}ys AlqAdm]*
People **know** that [democracy will result in real parliamentary and presidential elections].

Scope-related disagreement is attributed to (1) ambiguous subordinate conjunctions, (2) triggers modifiers, (3) absent punctuation markers, and (4) embedding within the scope boundaries. For instance, in example 27, the adverbial clause starting with بعد *bEd* (after) confuses the annotators as to whether it is part of the scope or it describes the verb epistemic trigger اتوقع *AtwqE* (I expect).

27. **اتوقع** جدا ان [اعتصام التحرير يتفض بنفس طريقة فض الاعتصام الاخير بعد ظهور اشكال غريبة فلجان الامن]
***AtwqE** jdA An [AEtSAm AltHryr ytfD bnfs Tryqp fD AlAEtSAm AlAxyr bEd Zhwr A$kAl grybp fljAn AlAmn]*
I very much **expect** that [the sit-in in Tahrir will be broken up in the same way as the last sit-in after seeing some strange faces at the security checkpoints].

Tense yields almost perfect inter-annotator alpha reliability and agreement rates. The one main disagreement factor, however, is such contexts as ابتديت اصدق *Abtdyt ASdq* (I started to believe). While the majority of annotators agree that such contexts mark present tense knowledge, beliefs and judgments, some annotators consider them as past tense.

5.4 Majority Statistics for *3arif*

Based on majority annotations, Table 3 gives statistics for *3arif* in terms of sense, polarity, intensification and tense. Furthermore, approximately 62% of the triggers have zero-nested holders (i.e. the Twitter user is the same as the holder). As for scope syntactic structures, they are distributed as 86% clauses, 9% deverbal nouns and the rest are to-infinitives.

	Sense		**Polarity**		**Intensification**			**Tense**	
	Epistemic	Non-epistemic	True	False	Amplified	Mitigated	As is	Present	Past
Tokens	5591	4375	3425	2166	1083	330	4178	4399	1192
Types	209	175	176	134	133	50	150	175	104

Table 3: Majority statistics for *3arif*

6 Related Work

Epistemic modality has been the focus of many annotation projects for multiple languages. Diab et al. (2009) annotate three belief categories for English: (1) committed belief is when writers indicate that they hold propositions as TRUE, (2) non-committed belief is when writers hold propositions as FALSE, and (3) not applicable is when propositions are not denoting beliefs at all. Interest is given to writers' beliefs only. Thus, a default value for the modality holder is the writer, and nested holders are not an-

notated. Their corpus contains 10k words of running text from different domains and genres, including newswire, blog data, email and letter correspondence and transcribed dialogue data. Inter-annotator agreement rate is 0.95 including the NONE category where no belief markers exist.

Baker et al. (2010, 2012) simultaneously annotate modality and modality-based negation to build modality taggers to enhance Urdu-English machine translation systems. Their annotation scheme distinguishes eight modality types: requirements, permissions, success, effort, intention, ability, desires and beliefs. Originally, their annotation scheme labels three attributes for each modality type: triggers, holders and targets (i.e. scopes). Yet, holders have not been eventually labeled. A unique feature of their annotation scheme is using a simplified operational procedure to label modality semantic meanings. The procedure relies on a list of thirteen choices of the form of H (modal) [P true/false] where H is a holder and P is a proposition or an event. The annotators' task is then to select the best form to represent the modality meaning of a given trigger. Reported kappa κ inter-annotator agreement rates are 0.82 for triggers and 0.76 for targets.

Rubinstein et al. (2013) propose a linguistically-motivated scheme for modality annotation in the MPQA English corpus. They attain macro alpha inter-annotator reliability rates of 0.89 and 0.65 for sense and scope, respectively. Cui and Chi (2013) apply the same scheme from Rubinstein et al. (2013) to the Chinese Penn Treebank and get alpha inter-annotator reliability rates of 0.81and 0.39 for sense and scope annotation, respectively.

Al-Sabbagh et al. (2013) annotate epistemic modality in MSA and EA tweets. We attain kappa inter-annotator agreement rates of 0.90 and 0.93 for sense and scope annotation, respectively, for only 548 epistemic tokens.

Our annotation results, therefore, are comparable to the results in the literature. Furthermore, our annotation scheme is orthogonal to most of the aforementioned schemes. However, the key differences between our work and related work are:

- We annotate nested modality, unlike Diab et al. (2009) and Baker et al. (2010, 2012).
- We use a wider range of negation and intensification markers compared to prior work, especially Al-Sabbagh et al. (2013)
- We use interactive crowdsourcing with simplified guidelines, unlike in-lab annotations including Rubinstein et al. (2013) and Cui and Chi (2013), among others.

7 Uncovered Points in *3arif*

The current version of *3arif* does not cover modality entailment that example 28 illustrates. The USER criticizes whoever holds as TRUE the proposition that Egypt can blackmail UAE using the Iranian threat. This criticism entails that the USER holds the same proposition as FALSE.

28. [يخطئ من **يظن** ان [#مصر يمكن ان تساوم الامارات بورقة #ايران]
*yxTY' mn **yZn** An [#mSr ymkn An tsAwm #Al<mArAt bwrqp #<yrAn]*
Whoever **thinks** that [Egypt can blackmail #UAE using #Iran] is wrong.

We do not also cover the future tense, the interrogative, the imperative or the hypothetical moods. This is because they have different interpretations when it comes to intensification and polarity that we do not cover in this version of *3arif* but we will in future work.

8 Conclusion

We presented *3arif*, a large-scale corpus annotated for epistemic modality in MSA and EA tweets. We used a simplified approach that defines each annotation task as a series of questions, implemented interactively. Our scheme covers a wide range of the most common annotation units mentioned in the literature, including modality sense, polarity, intensification, tense, holders and scopes. We deal with nested holders that are crucial in a highly interactive genre such as tweets where users frequently quote others and make assumptions about them. We also automatically merge triggers with shared holders and scopes based on elicited annotators' answers. The annotation procedure yields reliable results and creates a novel resource for Arabic NLP. For future versions of the corpus, we plan to cover the points

from Section 7. *3arif* will also be used to train and test an automatic machine learning system to identify epistemic modality and its attributes in MSA and EA tweets.

References

Muhammad Abdul-Mageed and Mona Diab. 2011. Subjectivity and Sentiment Annotation of Modern Standard Arabic Newswire. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*, pages 110-118, Portland, Oregon, June 23-24, 2011.

Rania Al-Sabbagh, Jana Diesner and Roxana Girju. 2013. Using the Semantic-Syntactic Interface for Reliable Arabic Modality Annotation. In *Proceedings of IJCNLP'13*, pages 410-418, Nagoya, Japan, October 14-18, 2013.

Rania Al-Sabbagh, Roxana Girju and Jana Diesner. 2014. Unsupervised Construction of a Lexicon and a Pattern Repository of Arabic Modal Multiword Expressions. In *Proceedings of the 10th Workshop of Multiword Expressions at EACL 2014*, pages 114-123, Gothenburg, Sweden, April 26-27, 2014.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, volume 34, issue 4, pages 555-596.

Kathrin Baker, Michael Bloodgood, Mona Diab, Bonnie Dorr, Nathaniel W. Filardo, Lori Levin and Christine Piatko. 2010. A Modality Lexicon and its Use in Automatic Tagging. In *Proceedings of LREC'10*, pages 1402-1407, Valetta, Malta, May 19-21, 2010.

Kathryn Baker, Michael Bloodgood, Bonnie J. Dorr, Chris Callison-Burch, Nathaniel W. Filardo, Christine Piatko, Lori Levin and Scott Miller. 2012. Modality and Negation in SIMT. *Computational Linguistics*. volume 38, issue 2, pages 411-438.

Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu and Nicholas Asher. 2012. How do Negation and Modality Impact on Opinions. In *Proceedings of the ACL-2012 Workshop on ExProM-2012,* pages 10-18, Jeju, Republic of Korea, July 13, 2012.

Yanyan Cui and Ting Chi. 2013. Annotating Modal Expressions in the Chinese Treebank. In *Proceedings of the IWC 2013 Workshop on Annotation of Modal Meaning in Natural Language (WAMM)*, pages 24-32, Potsdam, Germany, March 19, 2013.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed Belief Annotation and Tagging. In *Proceedings of the 3rd Linguistic Annotation Workshop, ACL-IJCNLP'09*, pages 68-73, Suntec, Singapore, August 6-7, 2009.

Iris Hendrickx, Amàlia Mendes and Silvia Mencarelti. 2012. Modality in Text: A Proposal for Corpus Annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12),* pages 1805-1812, Istanbul, Turkey, May 21-27, 2012.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha Reliability. Annenberg School of Communication, Departmental Papers: University of Pennsylvania.

Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui and Yuji Matsumoto. 2010. Annotating Event Mentions in Text with Modality Focus and Source Information. In *Proceedings of LREC'10*, pages 1456-1463, Valletta, Malta, May 19-21, 2010.

Frank R. Palmer. 2001. *Mood and Modality*. 2nd Edition. Cambridge University Press, Cambridge, UK.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simoson, Graham Katz and Paul Portner. 2013. Toward Fine-Grained Annotation of Modality in Text. In *Proceedings of the IWC 2013 Workshop on Annotation of Modal Meaning in Natural Language (WAMM)*, pages 38-46, Potsdam, Germany, March 19, 2013.

Roser Saurí and James Pustejovsky. 2009. FactBank: A Corpus Annotated with Event Factuality. *Language Resources and Evaluation*, volume 43, pages 227-268.

György Szarvas, Veronika Vincze, Richárd Farkas and János Csirik. 2008. The BioScope Corpus: Annotation for Negation, Uncertainty and their Scope in Biomedical Texts. In *Proceedings of BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, pages 38-45, Columbus, Ohio, USA, June 2008.

Anita de Waard and Henk Pander Maat. 2012. Epistemic Modality and Knowledge Attribution in Scientific Discourse: a Taxonomy of Types and Overview of Features. In *Proceedings of the 50th ACL,* pages 47-55, Jeju, Republic of Korea, July 12, 2012.

Janyce Wiebe, Theresa Wilson and Claire Cardie. 2005. Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, volume 39, issue 2-3, pages 163-210.