AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis

Muhammad Abdul-Mageed[*†] **and Mona Diab**[†]

[*†] Department of Linguistics and School of Library & Information Science; [†]Center for Computational Learning Systems
Indiana University, Bloomington, IN 47405-3907 USA; Columbia University, NY 10115 USA
mageed,mdiab@ccls.columbia.edu

Abstract

We present *AWATIF*, a multi-genre corpus of Modern Standard Arabic (MSA) labeled for subjectivity and sentiment analysis (SSA) at the sentence level. The corpus is labeled using both regular as well as crowd sourcing methods under three different conditions with two types of annotation guidelines. We describe the sub-corpora constituting the corpus and provide examples from the various SSA categories. In the process, we present our linguistically-motivated and genre-nuanced annotation guidelines and provide evidence showing their impact on the labeling task.

Keywords Arabic, sentiment analysis, opinion mining

1. Introduction

The area of *subjectivity and sentiment analysis* (SSA) has been receiving a booming interest in both the academia and the industry. Subjectivity in natural language refers to aspects of language used to express feelings, opinions, evaluations, and speculations (Banfield, 1982; Wiebe, 1994) and, as such, it incorporates sentiment. Subjectivity classification is the task of teasing apart objective (e.g., *The new iPhone is out.*) from subjective (e.g., *I'll finally buy the amazing iPhone!*) text units. Subjective text is further classified with *sentiment* or *polarity*. For sentiment classification, the task refers to identifying whether a subjective text is *positive* (e.g., *NLP rocks!*), *negative* (e.g., *The Syrian dictator has a lot of blood on his hands!*), *neutral* (e.g., *I may be off to Cairo next month.*), or, sometimes, *mixed* (e.g., *I adore this iPad, but it is prohibitively expensive*).

Two issues arise in SSA. First, in spite of the 'rush to the ground' the area of SSA has been witnessing, only a few attempts have been made to build SSA systems for *morphologically-rich languages (MRL)* (i.e., languages in which significant information concerning syntactic units and relations are expressed at the word-level (Tsarfaty et al., 2010). One of the serious issues with building systems for these languages is the unavailability of labeled data sets. Modern Standard Arabic (MSA), a very morphologically rich language e.g., (Diab et al., 2007; Habash et al., 2009), is a significant case in point. In this paper, we report efforts to bridge this gap in research by presenting *AWATIF*, a multi-genre corpus for MSA SSA.

Second, available approaches to labeling data for SSA vary considerably. In this paper, we extend our previous work, i.e., (Abdul-Mageed and Diab, 2011), by seeking to show how annotation studies within SSA can both be inspired by existing linguistic theory and cater for genre nuances. More specifically, by describing our efforts to label AWATIF for SSA and summarizing our linguistically-motivated, genre-nuanced guidelines for the task, we hope to trigger a stronger tie between existing linguistic (and in general rhetorical) theory and efforts to label data for social meaning tasks such as SSA. In order to show the impact of our annotation guidelines, we perform annotations under three different conditions with two types of annotation instructions and report the results we acquire. In the process, we explain sources of difficulty in our data and exemplify the various sentiment categories from our multi-genre corpus.

The rest of the paper is organized as follows: Section 2. provides an overview of the conditions under which we perform the annotation process and introduces the multi-genre corpus. In Section 3., we summarize our linguistically-motivated, genre-nuanced guidelines. In Section 4. we describe the annotation procedure along with relevant annotation conditions and agreement rates. In Section 6., we provide some examples from the labeled data. In Section 5., we discuss and exemplify sources of difficulty in the data. In Section 7., we review related literature and in Section 8. we conclude.

2. Annotation Conditions and Data sets

Setting off with two types of annotation guidelines, i.e., (1) simple (SIMP) and (2) linguistically-motivated and genre-nuanced (LG), we make attempts to measure how annotators would perform the task of labeling sentence-level MSA data under different conditions. For SIMP, we briefly introduce the fact that a sentence can be positive (POS), negative (NEG), or neutral (NEUT) and provide two examples of each of these three categories. SIMP, thus, is designed for a three-way classification task where an annotator is required to pick a tag from the set {*POS,NEG,NEUT*}. For LG, we (a) expose annotators to a linguistics background that we believe is useful for subjectivity and sentiment labeling and (b) explain the nuances of the genre to which each data set belongs. With these two types of guidelines we have annotators label each, or parts of some, of the data sets under three conditions. These conditions are:

1. **GH-LG:** Where annotation is performed in-lab with students of linguistics who have received clear training. We will refer to this group of annotators as Gold Human (GH). The GH cohort in this experimental condition annotates using the LG guidelines.

2. **GH-SIMP:** Where GH annotation is performed with SIMP annotation guidelines.

3. **AMT-SIMP:** Where the task is crowd-sourced on Amazon Mechanical Turk (AMT) with SIMP.

By comparing GH-LG to AMT-SIMP, we seek to measure the impact of the guidelines *per se*. Similarly, by comparing GH-SIMP to AMT-SIMP, we hope to measure the effect of varying the nature of annotation process itself (i.e., comparing crowd-sourcing to GH annotation with the same set of guidelines).

2.1. Penn Arabic Treebank (PATB)

This data set belongs to Part 1 V 3.0 (ATB1V3) of the Penn Arabic TreeBank (PATB) (Maamouri et al., 2004). The sentences make up the first 400 documents of that part of PATB amounting to a total of 54.5% of the ATB1V3. ATB1V3 is a collection of news wire stories from various domains (e.g., political, economic, sports).

2.2. Wikipedia Talk Pages

We harvested a corpus of 30 Wikipedia Talk Pages (WTP). The corpus makes up 5342 sentences. Attempts have been made to select Talk Pages (TP) from as varied as possible domains with extended conversations between Wikipedia editors. The 30 selected TPs were chosen from a pool of about 3000 TPs. The rest of the pool TPs were excluded since they were found to have either very short or no conversations. Selected TPs belong mostly to the political and religious/ideological domains. Examples of TPs from the political domains are 'Arab Revolutions,' 'Syrian Protests,' 'Libyan Civil War 2011,' and 'January 25th Revolution.' [1] The religious/ideological domain included TPs like 'Secularism,' 'Salafism,' 'Mormons,' and 'Ali ib Abi Talib.'

2.3. Web Forums

The Web forum (WF) collection comprises 2532 threaded-conversations from 7 WFs. The collection is part of a bigger blog and forum corpus that pertains to different varieties of Arabic, including dialects. For the current study, we have filtered the corpus to exclude non-MSA threads. For filtering, we followed several criteria. First, we manually inspected for majority dialect affiliation, i.e. checking to see if it comes from a specific variety (e.g., MSA) URL to see if the majority of the threads on that URL are indeed in that variety. Second, we ranked the threads based on the percentage of words that are analyzable by an MSA morphological analyzer, BAMA, version 2 (Buckwalter, 2002). The hypothesis is that if a word is found in BAMA then it is MSA. We recognize the caveat here that many words look orthographically similar to MSA while they could be *faux amis*, or even bear the same meaning as an MSA variant but are phonologically Dialectal. Clearly there is significant lexical overlap between the dialect and MSA. Moreover, the second caveat is that BAMA itself may contain dialectal entries, Indeed we found at least 18 dialectal words in BAMA out of a total of 78,839, making up a very tiny percentage, almost negligible, which renders our initial filtering approach feasible. Third, we make use of dialectal words in our in-house dialect dictionaries (COLABA dictionaries) to rank the varieties of threads. We create a single word list of all the dialectal entries in our dictionaries (which might comprise orthographically similar words to MSA) and use it as a word look up table.

[1] We provide the names of the TPs as they stood during the time of data scouting (i.e., early August, 2011). As such, our ideological inclinations are not reflected in these labels.

3. Linguistically-Motivated, Genre-Nuanced Guidelines

As mentioned earlier, our goal was to label a multi-genre corpus (from the newswire, Wikipedia Talk Pages, and Web Forums) with two types of guidelines (SIMP and LG), under three conditions (GH-LG, GH-SIMP, and AMT-SIMP). As we pointed out above, with SIMP, annotators are told that a sentence can be POS, NEG, or NEUT; provided with two examples for each category; and are asked to label sentences accordingly. Unlike SIMP, LG is built with the linguistic background and genre nuances of the three involved genres in mind. In this section, we summarize LG and provide an overview of the linguistics and genre-based concepts driving it, illustrating related and relevant literature.

3.1. Annotator's Background Knowledge

The type of sentiment expressed may vary based on the background knowledge of an annotator/reader (Balahur and Steinberger, 2009). For example, the sentence "Secularists will be defeated", may be positive to a reader who opposes secularism. However, if the primary intention of the author is judged to be communicating negative sentiment, annotators are supposed to assign a NEG tag. In general, annotators have been advised to avoid interpreting the subjectivity of text based on their own economic, social, religious, cultural, etc. background knowledge.

3.2. Good & Bad News

In general, news (as expressed in the news genre, but also potentially elsewhere) can be either good or bad. For instance, whereas "Five persons were killed in a car accident" is bad news, "It is sunny and warm today in Chicago" is good news. Our coders were instructed not to consider *good* news *POS* nor *bad* news *NEG* if they think the sentences expressing them are objectively reporting information. Thus, bad news and good news can be NEUT as is the case in both examples. Indeed, this specific nuance makes news-focused SSA a difficult task.

3.3. Politeness Theory

Politeness (e.g., (Brown and Levinson, 1987)) is related to the 'etiquette' of involving in a conversation. Politeness can play an important role that intersects with how subjectivity is expressed in interactive genres like WTPs and WFs, but also in quoted content in the newswire. The concept of politeness is related to that of *face*, the 'prestige,' 'esteem,' 'dignity,' etc. people create for themselves in social interactions and strive to maintain ((Goffman, 1955)). Although 'face' is claimed to be universal (Agassi and Jarvie, 1969), politeness varies cross-culturally e.g., (Matsumoto, 1989;

Gu, 1990). (Brown and Levinson, 1987) propose a *politeness theory* where they maintain a distinction between *positive face* (i.e., a person's desire to be respected, appreciated, approved of, etc.) and *negative face* (i.e., a person's desire not to be 'bothered,' constrained, imposed on, etc.). To illustrate, while politely asking someone to do something for us (e.g., "could you kindly open the window for me?") saves their positive face, it threatens their negative face (since it is some sort of imposition on them). We use insights from politeness theory to educate our annotators about interpersonal communication as expressed in written texts. We first define some variables for users. For example, for the WTP and WF genres, we assume that all interactants are equal in status and have similar age, unless these are recoverable from the interactants' participations. Similarly, we advise annotators to take the gender of interactants into account if they can identify it from texts. Although some of this contextual information would not be recoverable, attempts could be made to mine the profiles of users for such variables and provide them to annotators. However, such efforts are beyond our immediate focus.

Second, we introduce some heuristics so that insights from politeness theory work straightforwardly for our current purpose of data labeling with sentiment tags. For example, we advise annotators to label indirect and/or softened requests (e.g., "could you kindly open the window for me?") as POS, these being positive face saving. Since the same proposition can simultaneously be saving positive face and threatening negative face, we had to choose between the two options of assigning a POS or a NEG tag and we decided that the POS tag is more likely since the writer employs lexical items (such as "could you" and "kindly") to maintain politeness. Requests impose on others and hence threaten their negative face. We advised annotators to assign direct request like "Do not delete the first paragraph in section 2" a NEG tag because these threaten negative face and do not save positive face. We also instructed coders to label softened and indirect requests (e.g., "May I ask that we keep this section," "I personally believe we should move the discussion beyond this specific point") with POS tags since these save positive face.

3.4. (Dis-)Agreement/(Dis-)Approval

In WTPs, the task of Wikipedia editors is to work together to improve the quality of Wikipedia articles. Thus, discussions on WTPs are focused at the content of the articles, rather than just being *Wikichat* (e.g., a discussion about the behavior of individuals mentioned in the article or other Wikipedia editors). Wikipedia editors are advised to avoid incivility and personal attacks, again on the grounds that the WTPs are for discussing 'content' not 'contributors.' As such, the sentiment expressed in WTPs is, more often than not, subtle and embedded in attempts to *approve* or *disapprove* certain parts of associated articles. In their endeavors to 'improve' the quality of Wikipedia articles, editors also *agree* and *disagree* with one another and make attempts to refute others' arguments and provide their own pieces of evidence. Such discussions, although at times very direct, are at times wrapped in skillful usages of politeness strategies as ones mentioned above. WFs are also loci of agreement/disagreement. We instructed our annotators to label direct, unsoftened disagreements (e.g., "Well, you are definitely missing the point here.") with NEG tags and indirect, softened disagreement with NEUT tags (e.g., "I see your point, but I think it could be the other way round, couldn't it?"). Agreements (e.g., "Yes, your take on this seems to be hitting the point") and approvals (e.g., "The changes you made to the first section are useful and the sources you cite are also authoritative") were treated as carrying POS sentiment.

3.5. Epistemic Modality

Epistemic modality serves to reveal how confident writers are about the truth of the ideational material they convey (Palmer, 1986). Epistemic modality is classified into *hedges* and *boosters*. *Hedges* are devices like *perhaps* and *I guess* that speakers employ to reduce the degree of liability or responsibility they might face in expressing the ideational material. *Boosters*[2] are elements like *definitely, I assure that*, and *of course* that writers or speakers use to emphasize what they really believe. Both hedges and boosters can (1) turn a given unit of analysis from objective into subjective and (2) modify polarity (i.e., either strengthen or weaken it). Consider, for example, the sentences (1) "Gaddafi has murdered hundreds of people", (2) "Gaddafi may have murdered hundreds of people", and (3) "Unfortunately, Gaddafi has definitely murdered hundreds of people". While (1) is NEUT, since it lacks any subjectivity cues, (2) is NEUT because the proposition is not presented as a fact but rather is softened and hence offered as subject to counter-argument, (3) is a strong NEG (i.e., it is NEG as a result of the use of "unfortunately", and *strong* due to the use of the booster *definitely*). Our annotators were explicitly alerted to the ways epistemic modality markers interact with subjectivity.

3.6. Role of Perspective

Sentences can be written from different *perspectives* (Lin et al., 2006) or points of view. Consider the two sentences (1) "Israeli soldiers, our heroes, are keen on protecting settlers" and (2) "Palestinian freedom fighters are willing to attack these Israeli targets". While sentence (1) is written from an Israeli perspective, sentence (2) is written from a Palestinian perspective. The perspective from which a sentence is written interplays with how sentiment is assigned. Sentence (1) can usually be considered positive from an Israeli perspective, yet the act of protecting settlers is, more often than not, viewed as negative from a Palestinian perspective. Similarly, attacking Israeli targets may be positive from a Palestinian vantage point, but will perhaps be negative from an Israeli perspective. Coders were instructed to assign a tag based on their understanding of the type of sentiment, if any, the author of a sentence is trying to communicate. Thus, we have tagged the sentences from the perspective of their authors. As it is easy for a human to identify the perspective of an author (Lin et al., 2006), this measure facilitated the annotation task. Thus, knowing that the sentence (1) is written from an Israeli perspective, the annotator assigns it a POS tag.

[2] (Polanyi and Zaenen, 2006) call these *intensifiers*.

	OBJ	POS	NEG	NEUT	Total
OBJ	1192	21	57	11	1281
POS	47	439	2	3	491
NEG	69	0	614	6	689
NEUT	115	2	9	268	394
Total	1423	462	682	288	2855

Table 1: Agreement for SSA sentences for ATB1V3

	POS	NEG	NEUT	MIXED	Total
POS	379	26	26	20	451
NEG	20	509	29	29	587
NEUT	5	12	216	4	237
MIXED	16	29	7	181	233
Total	420	576	278	234	1508

Table 2: Agreement for SSA sentences for WTP

	POS	NEG	NEUT	MIXED	Total
POS	243	5	30	4	282
NEG	6	261	56	3	326
NEUT	10	25	316	5	356
MIXED	1	1	3	50	55
Total	260	292	405	62	1019

Table 3: Agreement for SSA sentences for WF

3.7. Illocutionary Speech Acts

Occurrences of language expressing *apologies, congratulations, praise*, etc. are referred to as *illocutionary speech acts* (ISA) (Searle, 1975). We strongly believe that ISAs are relevant to the expression of sentiment in natural language. For example, the two categories *expressives* (e.g., congratulating, thanking, apologizing) and *commiss ives* (e.g., promising) of (Searle, 1975)'s taxonomy of ISAs are specially relevant to SSA. In addition, (Bach and Harnish, 1979) define an ISA as a medium of communicating attitude and discuss ISAs like *banning, bidding, indicting, penalizing, assessing* and *convicting*. For example, the sentence "The army should never do that again" is a *banning* act and hence is NEG. Although our coders were not required to assign ISA tags to the sentences, we have brought the the concept of ISAs to their attention as we believe a good understanding of the concept does facilitate annotating data for SSA.

4. Annotation Procedure

In this section, we describe the annotation procedure followed for tagging each data set, relevant annotation conditions, and agreement rates.

4.1. GH-LG

We label (parts of) the three data sets under this condition. For ATB1V3, each of the two trained annotators assigned one of 4 possible labels: (1) Objective (OBJ), (2) Subjective-Positive (POS), (3) Subjective-Negative (NEG), and (4) Subjective-Neutral (NEUT). [3] We followed (Wiebe et al., 1999) in operationalizing the subjective vs. the objective categories. In other words, if the primary goal of a sentence is perceived to be the objective reporting of information, it was labeled OBJ. Otherwise, a sentence would be a candidate for one of the three subjective classes. Table 1 shows the contingency table for the two annotators judgments.

For both WTP and WF, two college-educated native speakers of Arabic labeled the respective data. For each sentence, each annotator was asked to assign a tag from the set {*POS, NEG, NEUT, MIXED*}. Unlike the ATB1V3 data set where the MIXED category occurs in a negligible number of cases, these two social media data sets were expected to have more MIXED sentences. From WF, 1508 sentences were labeled. Table 2 shows the confusion matrix for the agreement between the two annotators who worked on the WF corpus.

Table 3 shows the confusion matrix for the 1019 WTP sentences that were labeled under the GH-LG condition.

Table 4 shows the sizes and Kappa (k) agreement values for all data labeled under the GH-LG condition. As the table shows, Kappa values for both the WTP and WF are lower than the ATB1V3. This difference in Kappa values suggest that social media, interactive, genres like WTP and WF may be slightly difficult to label for subjectivity and sentiment than newswire data.

4.2. AMT-SIMP vs. GH-SIMP

For the AMT-SIMP condition, we put 10500 sentences from the WF corpus and 5341 sentences on AMT and had three turkers label each sentence with a tag from the set {*POS, NEG, NEUT*}. A total of 387 turkers worked on our data. In order to sort out spammers from 'faithful' workers, we used several criteria. First, we allowed only workers with $>$ 95% life-time approval rate to work on our data. Second, we blocked all workers who spent an average of $<$ 5 seconds on a sentence or failed to agree at least 95% of the time with a small (N=15 sentences) set of gold data we prepared as a final manual procedure before allowing a worker to continue working on the task. In this way, we excluded the work of 49 turkers (%= 13). When we impose a stricter measure that the three turkers agree, we retrieve 944 sentences (%= 17.67) from the WTP corpus and 2372 sentences (%= 22.59) from the WF corpus. When we impose the less strict procedure that at least two turkers agree for us to keep a labeled sentence in our database, we retrieve 4399 sentences (%= 82.36%) from the WTP data set and 9063 sentences (%= 86.31%) from the WF set. We be-

Data set	Sentences Labeled	Kappa (k)
ATB1V3	2855	0.820
WTP	1019	0.790
WF	1508	0.793

Table 4: Size and agreement for GH-LG data

[3] We only saw sentences with a MIXED (i.e., both POS and NEG sentiment) attested in a negligible percent of the sentences and hence decided to tag this very few number of subjective MIXED cases with a NEUT category.

lieve these are reasonable percentages of the data and plan to re-run the rest of respective sentences where at least no annotators agreed on AMT.

As mentioned earlier, we wanted to identify the difference resulting from varying the annotation real-world variables via comparing regular (GH-SIMP) and crowd-sourcing (AMT-SIMP) labeling results. To that goal, we instructed one college-educated native speaker of Arabic with the same simple guidelines we used with AMT and maintained the same tag set (i.e., {*POS, NEG, NEUT*}) and had her label a subset (N= 500 sentences) of the WTP corpus (WTP-GH-SUB). We refer to the corresponding subset of the WTP corpus labeled under AMT-SIMP condition as WTP-AMT-SUB.

In order to have a gold standard to compare WTP-GH-SUB and WTP-AMT-SUB to, we labeled 500 sentences (WTP-GOLD) from the WTP data with the same {*POS, NEG, NEUT*} tag set. We found that both WTP-GH-SUB and WTP-AMT-SUB agree only slightly with WTP-GOLD (i.e., with a kappa k= 0.19 in the case of WTP-GH-SUB and a kappa k= 0.065 in the case of WTP-AMT-SUB). This shows only slight agreement in the case of WTP-GH-SUB and very slight agreement in the case of WTP-AMT-SUB (Landis and Koch, 1977).

These low agreement values suggest that (1) the detailed and nuanced guidelines have a positive effect on annotation, and (2) the regular annotation process is slightly preferable to the AMT process. In addition, the low agreement values reflect the difficulty of the task, an issue that we turn to in the next section.

5. Sources of Difficulty in the Data

Both the WF and the WTP data sets included some content that we believe is difficult even for a college-educated native speaker of Arabic. First, some of the content comes from a highly elevated literary register and employs language that some of the annotators may not be familiar with. In example 1 below, the author is presenting a critical analysis of a line of verse and makes use of syntactic (i.e., "pronoun", "noun") and literary (i.e., "at the level of the case") terminology with which not all annotators may be familiar.

(١) اقتران الضمير بالاسم يدل على حالة تشبه المعانقة والتي يصعب معها الفصل على صعيد الحالة.

Transliteration: AqtrAn AlDmyr bAlAsm ydl ElY HAlp t$bh AlmEAnqp wAlty ySEb mEhA AlfSl ElY SEyd Alklmp wElY SEyd AlHAlp.

English: The coupling of the pronoun and the noun indicates a state that is similar to cuddling with which separation at the level of a case becomes difficult. [4]

The way already unfamiliar lexica combine with other lexica makes the text even harder. For instance, example 2 using the word (أبدية "Obdyp" "eternal") to qualify (طلاسم "TlAsm" "talismans") adds ambiguity to the sentence. While sch an ambiguity is part of the attractiveness of a literary work, it makes the task of sentiment labeling harder for annotators.

(٢) فبين الحين والحين أجدني مكبلا بطلاسمك الأبدية.

Transliteration: fbyn AlHyn wAlHyn Ojdny mkblA bTlAsmk AlObdyp.

English: Every now and then I find myself cuffed with your eternal talismans.

The degree of difficulty embodied in employing of very specialized lexica and high abstraction is very clear in Example (3) below.

(٣) الذي بين يدينا هو نص يتسم بالتركيز والتكثيف الشديدين ويتسم بمسافة أبعد من الابهام والذهنية وهذا يجعل من النص أحيانا أداة لّتباعد بين النص والمتلقي كما أنه بالنسبة للمتلقي المتخصّ أو غير العادي ينشط آليات التفكير والخروج من القراءة الحرفية للنص إلى قراءة أبعد.

Transliteration: Al*y byn ydynA hw nS ytsm bAltrkyz wAltkvyf Al$dydyn wytsm bmsAfp ObEd mn AlIbhAm wAl*hnyp wh*A yjEl mn AlnS OHyAnFA OdAp lltbAEd byn AlnS wAlmtlqy kmA Onh bAlnsbp llmtlqy AlmtxSS Ow gyr AlEAdy yn$T —lyAt Altfkyr wAlxrwj mn AlqrA'p AlHrfyp llnS IlY qrA'p ObEd.

English: What we have is a very concentrated and condensed text, characterized with a further distance of ambiguity and abstraction, which sometimes makes the text a tool for distancing audience and the text itself. In addition, for a specialized audience, it activates thinking mechanisms and induces a non-literal reading.

In love poetry, an author may be portraying a positive picture of himself by meticulously describing how it is that he/she intentionally excruciates himself/herself for the sake of love. While such a technique may be readily clear to a person with a literary background, this may not be the case for an annotator. For example, it takes careful thinking to identify example (4) below as a POS instance.

(٤)... لكنّي عبثا أحاول حينما أستحضر امتلاك غيري لك.

Transliteration: ... lknny EbvA OHAwl HynmA OstHDr AmtlAk gyry lk.

English: ... but in vain I try when it haunts me that someone else's you ARE.

Another source of difficulty comes from the use of classical Arabic (CA) lexica, expressions, and idioms typical of a religious register. For instance, the idiomatic expression in Example (3) below is a positive one, but it is unfamiliar to many (even highly educated) Arabic speakers.

[4] We would like to remind the reader that the translation of some of the examples does not necessarily render sentences of the same level of difficulty as the original.

(٥) لله درك.

Transliteration: llh drk.

English: Impressive is what you are doing!

Similarly, the word (ديدنه "dydnh" "his typical method") in Example (4) is unfamiliar and so is the word (m_dhbyT "m*hbyp" "belief-related").

(٦) ووجدت أنّ هذا ديدنه دائما خاصة بالأمور المذهبية.

Transliteration: wwjdt On h*A dydnh dAmAF xASp bAlOmwr Alm*hbyp.

English: I found that this is his typical method, especially with relation to issues of belief.

Within the religious register, authors sometimes employ very specialized terminology that comes from certain religious sciences. For instance, the term متواتر ("*mutawatir*"; trustworthy because it has been narrated by frequently enough, trustworthy people) in Example (5) comes from (علم الحديث "Elm AlHdyv" "science of Prophetic Tradition").

(٧)... فهي متواترة تواترا معنويا بحد أقل.

Transliteration:... fhy mtwAtrp twAtrAF mEnwyAF bHd Oql.

English:... so it is less trustworthy, in a moral sense.

6. Examples

6.1. POS Examples

Examples 8 (WTP) and 9 (WF) illustrate sentences labeled with the POS tag in the data.

(٨) فتحطيم خط بارليف لا شك أنه ينم عن براعة قل نظيرها.

Transliteration: ftHTym xT bArlyf lA $k Onh ynm En brAEp ql nZyrhA.

English: Destroying the Bar-Lev line reflects rare versatility.

(٩) جميل للغاية ... هذا التعبير تشبيه البشر بالمصابيح المطفأة وحقيقة هذا التعبير يدل على أنه كان الشمس التي تمدهم بالضياء.

Transliteration: jmyl llgAyp...h*A AltEbyr t$byh Alb$r bAlmSAbyH AlmTfp wHqyqp h*A AltEbyr ydl ElY Onh kAn Al$ms Alty tmdhm bAlDyA'.

English: Extremely beautiful...This expression likens humans to extinguished lamps, which means in essence that he was the sun that provides them with the light.

6.2. NEG Examples

Examples 10 (WTP) and 11 (WF) illustrate negative sentences.

(١٠) انها آراء منسوبة لأصحابها وليست ذات أهمية أو قيمة أو مغزى أو جدوى أو بال.

Transliteration: swY OnhA —rA' mnswbp lOSHAbhA wlyst *At Ohmyp Ow qymp Ow mgzY Ow jdwY Ow bAl.

English: These views are only ascribed to those who held them, rather then being important, valuable, significant, useful, nor worthy views.

(١١) واشمئزاز نفوسنا من طلعته البهية ما زال يلعب أنفسنا، مات مايكل الساكت قلبه كمدا.

Transliteration: wA$mzAz nfwsnA mn TlEth Albhyp mA zAl yulEbu OnfsnA, mAt mAykl AlsAkt qlbh kmdA.

English: Our souls are still dancing of disgust from seeing his beautiful face; Michael, the grieve-hearted, is dead.

6.3. NEUT Examples

Examples 12 (WTP) and 13 (WF) illustrate negative sentences.

(١٢) من الممكن أن نضع عنوان فرعي لهذا المقطع نقول فيه (ردود فعل الاخوان المسلمين) بدلا من حذفه.

Transliteration: mn Almmkn On nDE EnwAn frEy lh*A AlmqTE nqwl fyh (rdwd fEl AlIxwAn Almslmyn) bdlA H*fh.

English: We can keep this part as a sub-section titled "Reactions of the Muslim Brotherhood", instead of deleting it.

(١٣) في قنينة حبري الاسود ذوبتها.

Transliteration: fy qnynp Hbry AlAswd *wbthA.

English: It melted it down in my bottle of black ink.

6.4. OBJ Examples

Examples 14 and 15, both from ATB1V3 illustrate sentences labeled with the OBJ tag.

(١٤) ويبلغ عدد المشردين في كنتية لوس انجلس نحو ٨٤ الف شخص.

Transliteration: wyblg Edd Alm$rdyn fy kwntyp lws Onjlys nHw 84 Olf $xS.

English: The number of homeless in Los Angeles County is about 48 thousand.

(١٥) طهران ٧-١٥ (أف ب) - وقع ١٦ انفجارا
مساء اليوم السبت في وزارة الاستخبارات حيث
استدعيت العديد من سيارات الاسعاف كما أكد
شاهد عيان لوكالة فرانس برس.

Transliteration: ThrAn 15-7 (A f b) - wqE 16 AnfjArA msA' Alywm Alsbt fy wzArp AlAstxbArAt Hyv. AstdEyt AlEdyd mn syArAt AlIsEAf kmA Okd $Ahd EyAn lwkAlp frAns brs.

English: Tehran 15-7 (AFP) - An eye witness affirmed to AFP that 16 explosions occurred late Saturday at the Ministry of Intelligence where many ambulances were summoned.

7. Related Work

There are a number of datasets annotated for SSA. Most relevant to us is work on the newswire and Web forum genres. (Wiebe et al., 2005) describe a fine-grained news corpus manually labeled for SSA at the word and phrase levels. (Balahur et al., 2009) report work on labeling quotations from the news involving one person mentioning another entity and maintain that quotations typically contain more sentiment expressions than other parts of news articles. Our work is different from that of (Balahur et al., 2009) in that we label all sentences regardless whether they include quotations or not. (Abbasi et al., 2008) briefly describe labeling a collection of documents from Arabic Web forums. (Abbasi et al., 2008)'s dataset, however, is not publicly available and detailed information as to how the data was annotated is lacking. Our work is different from (Abbasi et al., 2008)'s in that we label instances at the sentence level, and hence our corpora are more fine-grained. We do not know of any corpus from the WTP genre tagged for sentiment, and hence our WTP corpus is expected to trigger interest in that direction. In our own previous efforts, i.e., (Abdul-Mageed and Diab, 2011), to label an MSA corpus for SSA, we only focused on the newswire genre. In addition, we did not incorporate crowd-sourcing nor did we seek to identify how the annotation process can be affected by varying annotation conditions as we do in the current work. As such, this work extends our previous work in various ways.

8. Conclusion

The concepts of subjectivity and sentiment are fuzzy (Gamon et al., 2005; Wiebe et al., 2005), and hence annotators should be well-trained on the task. The improvement we achieve when we use LG, which incorporates linguistically-motivated and genre-nuanced guidelines, proves that without such training it is difficult to acquire dependable annotations. AWATIF is expected to help bridge a gap in research that it can be exploited for building genre-nuanced SSA systems for Arabic. Our corpus is also expected to help uncover how a MRL like Arabic can be handled in the context of social meaning extraction tasks like that of SSA, as there are currently only few attempts (e.g., (Abdul-Mageed et al., 2011; Abdul-Mageed and Diab, 2012)) to build SSA systems and resources for Arabic. AWATIF is unique in the sense that it is a multi-genre corpus.

9. References

A. Abbasi, H. Chen, and A. Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26:1–34.

M. Abdul-Mageed and M. Diab. 2011. Subjectivity and sentiment annotation of modern standard Arabic newswire. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 110–118, Portland, Oregon, USA, June. Association for Computational Linguistics.

M. Abdul-Mageed and M. Diab. 2012. Toward building a large-scale arabic sentiment lexicon. In *Proceedings of the 6th International Global WordNet Conference*, Matsue, Japan, January.

M. Abdul-Mageed, M. Diab, and M. Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 587–591, Portland, Oregon, USA, June. Association for Computational Linguistics.

J. Agassi and I.C. Jarvie. 1969. A study in westernization. *Hong Kong: A society in transition*, pages 129–163.

K. Bach and R.M. Harnish. 1979. Linguistic communication and speech acts.

A. Balahur and R. Steinberger. 2009. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. *Proceeding of WOMSA*.

A. Balahur, R. Steinberger, E. van der Goot, B. Pouliquen, and M. Kabadjov. 2009. Opinion mining on newspaper quotations. In *2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 523–526. IEEE.

A. Banfield. 1982. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge & Kegan Paul, Boston.

P. Brown and S.C. Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge Univ Pr.

T. Buckwalter. 2002. Arabic morphological analyzer version 1.0. Linguistic Data Consortium, 2002.

M. Diab, K. Hacioglu, and D. Jurafsky. 2007. Automatic processing of Modern Standard Arabic text. *Arabic Computational Morphology*, pages 159–179.

M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. 2005. Pulse: Mining customer opinions from free text. *Advances in Intelligent Data Analysis VI*, pages 121–132.

E. Goffman. 1955. On face-work: an analysis of ritual elements in social interaction. *Psychiatry: Journal for the Study of Interpersonal Processes*.

Y. Gu. 1990. Politeness phenomena in modern chinese. *Journal of pragmatics*, 14(2):237–257.

N. Habash, O. Rambow, and R. Roth. 2009. Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In *Proceedings of the 2nd International*

Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

W.H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. 2006. Which side are you on?: identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 109–116. Association for Computational Linguistics.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The penn arabic treebank: Building a large-scale annotated arabic corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109.

Y. Matsumoto. 1989. Politeness and conversational universals–observations from japanese. *Multilingua-Journal of Cross-Cultural and Interlanguage Communication*, 8(2-3):207–222.

F. Palmer. 1986. *Mood and Modality. 1986*. Cambridge: Cambridge University Press.

L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. *Computing attitude and affect in text: Theory and applications*, pages 1–10.

J.R. Searle. 1975. A taxonomy of speech acts. In K. Gunderson, editor, *Language, mind, and knowledge*, pages 344–369. Minneapolis: University of Minnesota Press.

R. Tsarfaty, D. Seddah, Y. Goldberg, S. Kuebler, Y. Versley, M. Candito, J. Foster, I. Rehbein, and L. Tounsi. 2010. Statistical parsing of morphologically rich languages (spmrl) what, how and whither. In *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, Los Angeles, CA.

J. Wiebe, R. Bruce, and T. O'Hara. 1999. Development and use of a gold standard data set for subjectivity classifications. In *Proc. 37th Annual Meeting of the Assoc. for Computational Linguistics (ACL-99)*, pages 246–253, University of Maryland: ACL.

J. Wiebe, T. Wilson, and C. Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210.

J. Wiebe. 1994. Tracking point of view in narrative. *Computional Linguistics*, 20(2):233–287.