A Supervised POS Tagger for Written Arabic Social Networking Corpora

Rania Al-Sabbagh
Department of Linguistics
University of Illinois at Urbana-Champaign
USA
+1-217-848-0020
alsabba1@illinois.edu

Roxana Girju
Department of Linguistics
University of Illinois at Urbana-Champaign
USA
+1-217-244-3104
girju@illinois.edu

Abstract

This paper presents an implementation of Brill's Transformation-Based Part-of-Speech (POS) tagging algorithm trained on a manually-annotated Twitter-based Egyptian Arabic corpus of 423,691 tokens and 70,163 types. Unlike standard POS morpho-syntactic annotation schemes which label each word based on its word-level morpho-syntactic features, we use a function-based annotation scheme in which words are labeled based on their grammatical functions rather than their morpho-syntactic structures given that these two do not necessarily map. While a standard morpho-syntactic scheme makes comparisons with other work easier, the function-based scheme is assumed to be more efficient for building higher-up tools such as base-phrase chunkers, dependency parsers and for NLP applications like subjectivity and sentiment analysis. The function-based scheme also gives new insights about linguistic structural realizations specific to Egyptian Arabic which is currently an under-resourced language.

1 Introduction

Part-of-Speech (POS) tagging is an enabling technology required for higher-up Natural Language Processing (NLP) tools such as chunkers and parsers – syntactic, semantic and discourse; all of which are used for such applications as subjectivity and sentiment analysis, text summarization and machine translation among others. Labeling words for their grammatical categories (i.e. POS tagging) is a non-trivial process given the inherent ambiguity of natural languages at various linguistic levels.

Genre-specific features can also pose extra challenges to POS tagging. The interactive conversational nature of the microblogging service Twitter introduces highly-dialectal input in which new words are coined on frequent basis. This implies that using non-corpus-based approaches, using POS taggers designed for Modern Standard Arabic (MSA) or leveraging Egyptian Arabic (EA) taggers from MSA ones are unlikely to perform well. These implications are empirically proved in prior work. Habash and Rambow (2006) achieve a coverage rate of only 60% for Levantine Arabic using MSA morphological analyzer. Abo Baker et al. (2008) and Salloum and Habash (2011) build linguistically inaccurate morphological analyzers trying to extend MSA tools to Arabic dialects. Duh and Kirchhoff (2005) build a minimally supervised POS tagger for EA of only 70.88% accuracy by using an MSA morphological analyzer and adding the Levantine Arabic TreeBank to their EA training corpus to benefit from the cross-dialect overlap in Arabic.

We, therefore, start our experiments for building a POS tagger for EA tweets with a supervised Transformation-Based Learning (TBL) approach trained and tested on EA tweets only. One advantage of this approach is its non-stochastic mechanism that is unlikely to be affected by frequent new word coinages and the sparsity they introduce to the training corpus.

We propose a function-based POS annotation scheme for this paper. Instead of standard morpho-syntactic annotation schemes which use word-level morpho-syntactic information for POS tagging, our scheme labels words based on their grammatical functions instead of their morpho-syntactic structures. Direct mapping between the word grammatical function and its morpho-syntactic structure is not always granted. This annotation scheme requires a new tagset that adapt tags from standard tagsets like Arabic TreeBank (ATB) and also uses new tags. The main advantage of our function-based scheme and tagset is to enhance developing such NLP tools as chunkers, dependency and discourse parsers and such applications as

subjectivity and sentiment analysis systems which is one main application in mind while building our POS tagger. Evidence on enhancing subjectivity and sentiment analysis systems and dependency parsers as well as comparing our scheme and tagset to more standardized ones are both kept for future work. Our POS tagger is still a contribution to the repository of Dialectal Arabic (DA) NLP tools which is to-date limited.

The rest of this paper is organized as follows: section 2 briefly discusses related work to POS tagging, focusing particularly on dialectal Arabic. Section 3 describes our POS tagset and adaptations made from standardized tagsets. Section 4 explains our function-based tokenization and tagging scheme. Section 5 describes the corpus and preprocessing procedures. Section 6 explains the annotation process and inter-annotator agreement rates. Section 7 discusses Brill's implementation of transformation-based learning to POS tagging and one application of it on MSA. Section 8 elaborates on our evaluation results and error analysis. Finally, section 8 outlooks major conclusions and future plans.

2 Related Work

Of the most recent NLP tools built for EA is Habash et al. (2012). Extending the Egyptian Colloquial Arabic Lexicon (Kilany et al., 2002) and following the POS guidelines by the Linguistic Data Consortium (LDC) for Egyptian Arabic (Maamouri et al., 2012a as cited in Habash et al., 2012), they build the large-scale morphological analyzer – CALIMA. It relies on tabular representation of complex prefixes, complex suffixes, stems and compatibility across them (prefix-stem, prefix-suffix and stem-suffix). Tested against a manually-annotated EA corpus of 3,300 words (Maamouri et al., 2012b as cited in Habash et al., 2012), CALIMA achieves a coverage rate of 92.1% where coverage is defined as the percentage of the test words whose correct analysis in context appears among the analyses returned by the analyzer. It also provides among its results a correct answer for POS tags 84% of the time.

With the goal of utilizing MSA morphological tools to create an EA training corpus and using data from several varieties of Arabic in combination, Duh and Kirchhoff (2005) build a minimally supervised EA tagger without the need to develop dialect-specific tools or resources. For data, they use the CallHome Egyptian Arabic corpus from LDC, the LDC Levantine Arabic corpus and the Penn Arabic Treebank corpus parts 1 to 3. For the morphological analyzer, they use Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2002) designed for MSA. Their approach bootstraps the EA tagger in an unsupervised way using POS information from BAMA and subsequently improves it by integrating additional data from other dialects given the assumption that Arabic dialects do overlap. Tested against Egyptian Colloquial Arabic Lexicon (Kilany et al., 2002), their best accuracy rate is 70.88%. Adding word-level features such as affixes and constrained lexicon first raises accuracy from 62.76% to 69.83% and then adding Levantine data to the training corpus raises accuracy to 70.88%.

MAGEAD (Habash and Rambow, 2006) is a morphological analyzer and generator for the Arabic language family – MSA and dialect. It uses the root-patter-features representation for online analysis and generation of Arabic words. Tested against the Levantine Arabic Treebank, MAGEAD achieves a context-type recall rate of 95.4% and a context-token recall rate of 94.2%. Context-token/type recall is defined as the number of times MAGEAD gets the contextually correct analysis for that word token/type.

Diab et al. (2010) build a large annotated corpus for multiple Arabic dialects, of which EA is a part. The corpus contains texts from blogs covering the domains of social issues, religion and politics linguistically analyzed at different levels. In addition to morphological analyses, Diab et al. (2010) give information about POS tags, the degree of dialectness of each word and sentence boundaries. Much of the work is being done manually or is the output of MAGEAD – after being tuned for DA. Performance rates for each task are not, however, reported.

3 The POS Tagset

There is a large number of Arabic POS tagsets including: BUCKWALTER (Buckwalter, 2002) used in the Penn Arabic TreeBank (ATB), Khoja tagset (Khoja, 2001), PADT tagset (Hajič et al., 2004), Reduced Tagset (RTS) (Diab, 2007) and CATiB POS tagset (Habash and Roth, 2009). Each of these tagsets represents a different level of granularity: at one end of the continuum is the most fine-grained tagset of Buckwalter with over 500 tags and at the opposite end is the most coarse-grained tagset of CATiB with only six tags. For a full review of these tagsets refer to Habash (2010).

Our tagset mixes and matches tags across fine-grained tagsets – to achieve the following goals:

- Give a fine-grained level of accurate linguistic description for the word POS and its semantic features of gender, number, person, aspect, voice, tense and mode.
- Tag words that can be used – as in future work – as classification features for base-phrase chunkers and parsers. These words include function words such as interrogatives, complementizers, conditionals and the like.
- Tag parts of speech that can be used – as in future work – as subjectivity and sentiment classification features like modals and negation among others.

In appendix (A), we compare and contrast our tagset with that of ATB and RTS to facilitate comparing results with other taggers – if any – that are using different tagsets. Our tagset is a subset of ATB and is a superset of RTS.

We add new tags to label Twitter-specific information and EA-specific grammatical categories like fixed expressions, existentials and aspectual progressives. Twitter-specific information requires tags for mentions(MNT), hashtags (HSH), emoticons (EMO), URLs (URL) and speech effects (LNG; for LeNGthened words) (e.g. اووووووي *Awwwwwy* (very) and خنيييق *xnyyyyyq* (boring)).

Approximately, 1% of our corpus is given our new tag EXP – for fixed expression. We define fixed expressions according to the following criteria:

- They can be either unigram or multiword expressions;
- Multiword fixed expressions are frozen in the sense that their individual words are not substitutable for synonyms. However, some of those expressions might have shorter versions;
- Their meaning is not compositional and are rarely – if not never – used literally;
- Their grammatical behavior does not match that of nouns, verbs, adjectives or adverbs. In other words, they cannot be head nouns or verbs in noun or verb phrases, respectively. They cannot modify nouns like adjectives or modify verbs like adverbs.
- They are used for pragmatic purposes to show, for example, shock as in يالهوي *yAlhwy* (Oh my goodness!), surprise as in ياحلولي *yAHlwly* (how interesting!) and frustration as in الصبر من عندك يارب *AlSbr mn Endk yArb* (lit: patience is from you, Lord; gloss: Oh, Lord! Grant me patience) among other emotions.

The unigram fixed expression يالهوي *yAlhwy* (Oh my goodness) is diachronically composed of the vocative particle يا *yA* (oh), the noun لهو *lhw* (goodness) and the possessive pronoun ي *y* (my). Yet, it cannot be decomposed into its parts and none of its parts can be substituted for a synonym. It functions only to show shock, anger, frustration emotions and the like. Meanwhile, its syntactic behavior does not fit in the paradigms of verbs, nouns, adjectives or adverbs.

The same thing applies to the multiword fixed expression الصبر من عندك يارب *AlSbr mn Endk yArb* (lit: patience is from you, Lord; gloss: Oh, Lord! Grant me patience). It is very rarely used literally as a prayer. As one expression, it does not grammatically behave like nouns, verbs, adjectives or adverbs. It is typically used as an expression of frustration or anger. Yet, it shows some degree of structural flexibility given that a shorter form exists الصبر يارب *AlSbr yArb*.

Two other tags that we use although they do not have equivalents in previous Arabic tagsets are: EX and PG for existentials and aspectual progressives, respectively. Unlike MSA, existentials in EA are not expressed by the deictic هناك *hnAk* (there) or the imperfect verb يوجد *ywjd* (exist). They are expressed by the preposition في *fy* or فيه *fyh*. Should these prepositions be used as existentials, they can syntactically map to a complete sentence such as فيه موافقة *fyh mwAfqp* (there is an agreement). Thus adding the EX tag to our tagset serves the purpose of facilitating phrase-boundary identification in later annotation layers for chunkers and parsers.

Unlike MSA, EA has an aspectual progressive verb prefix بـ *b* found in examples like بيكتب *byktb* (he's writing), بيفكر *byfkr* (he's thinking) and بتقول *btqwl* (she's saying). The aspectual progressive prefix is split off in tokenization and is tagged as PG.

Another tag that we add is MD to tag modals and modal adjuncts – in both verbal and nominal forms. For example, both the modal verb يمكن *ymkn* (may) in يمكن يجتمعو *ymkn yjtmEw* (they may meet) and the modal adjective ضروري *Drwry* (must) in ضروري نروح *Drwry nrwH* (we must go) are both tagged as MD. MD is used to tag all modality types – epistemic, deontic and evidential.

Some tagsets like RTS give simple, comparative and superlative adjectives one tag – JJ. ATB labels only simple and comparative adjectives, given that superlative adjectives are not morphologically marked. In our tagset, simple, comparative and

superlative adjectives are tagged as JJ, JJR and JJS, respectively.

We collapse some tag subsets in ATB into one. Instead of distinguishing connective particles from coordinating conjunctions, both are given the CC tag because both coordinate base and complex phrases. Thus و *w* (and), أو *>w* (or), وبعدين *wbEdyn* (and then) and وكمان *wkmAn* (and also) are all tagged as CC.

Another collapsed tag subset from ATB is the interrogative subset. Instead of three different tags for interrogative adverbs, particles and pronouns, one tag is used, namely INT for interrogative. For instance, the interrogative adverbs إزاي *<zAy* (how) and فين *fyn* (where) as well as the interrogative pronouns ايه *Ayh* (what) and مين *myn* (who) are all tagged as INT. The Arabic interrogative particles – هل *hl* and أ > used in MSA to form yes/no questions are not used in EA. When encountered in MSA tweets, they are also labeled as INT.

Relative adverbs and pronouns are also collapsed into one tag RL. EA has one relative pronoun – اللي *Ally* (who, which, that). When MSA relative pronouns or adverbs are encountered, they are tagged as RL.

Our verb tag subset does not define the voice feature (active vs. passive) which is given a separate tag – P for passive and the absence of such a tag indicates an active voice. Our noun tag subset does not indicate the number feature – singular, dual or plural – of the noun since these are given their separate tag subset. Therefore, NN is a common noun and NNP is a proper noun whether singular, dual, plural or a collective noun.

Based on our 49 base tagset, each content word and some function words are given complex tag vectors of the form <person>_<number>_<gender>_<voice>_<grammatical category>. Currently, our corpus has a total of 4,272 unique vectors. Some examples are in table (1).

Input	**Tokenized**	**POS Tagged**
بيقول *byqwl* (he's saying)	*b- yqwl*	*b*/PG *yqwl*/3_SG_M_VBP
شافهم *$Afhm* (he saw them)	*$Af -hm*	*$Af*/3_SG_M_VBD *hm*/3_PL_OBJP
بتتكتب *bttktb* (it's being written)	*b-ttktb*	*b*/PG *ttktb*/3_SG_F_P_VBP
الحكومة *AlHkwmp* (the government)	*Al-Hkwmp*	*Al*/DT *Hkwmp*/SG_F_NN
كويسين *kwysyn* (good; plural)	*kwysyn*	*kwysyn*/PL_JJ
هي *hy* (she)	*Hy*	*hy*/3_SG_F_SBJP
ليكي *lyky* (for you)	*ly-ky*	*ly*/IN *ky*/2_SG_F_OBJP

Table 1: Tokenization and POS tagging examples

4 Function-Based Tokenization and POS Tagging

Almost all tokenization and POS tagging approaches for Arabic – MSA or dialectal Arabic – rely on word-level morpho-syntactic structures for tokenization and POS tagging. In this paper, we present a function-based tokenization and POS tagging scheme in which words are tokenized and POS tagged based on their grammatical functions rather than their morpho-syntactic structures given that these two do not necessarily map. For example, زمانها *zmAnhA* in المسيرة زمانها خلصت *Almsyrp zmAnhA xlSt* (the march must have finished) is labeled as MD (i.e. modal) because it functions as a modal (*must have*) despite being morpho-structurally a noun زمان *zmAn* (time; era) and a possessive pronoun ها *hA* (her; hers). The same word in another context – as in مصر بتبتدي زمانها الجديد *mSr btbtdy zmAnhA Aljdyd* (Egypt is starting its new era) – is tokenized as *zmAn-hA* and tagged as *zmAn*/SG_M_NN *hA*/3_SG_F_PP$.

Using this function-based scheme for our tokenization and POS tagging provides a gold standard corpus for training and testing lexico-syntactic disambiguators, base-phrase chunkers and parsers. We leave it for future work to compare the performance of those tasks when trained on our function-based scheme and when trained on morpho-syntactic schemes.

The grammatical categories affected by our function-based scheme are: existentials, prepositional phrases, active participles, modals, superlative adjectives, multiword connective particles and fixed expressions. These are the grammatical categories which are typically ambiguous in terms of the mapping between their morpho-syntactic structures and their grammatical functions.

The existential فيه *fyh* (there is/are) in فيه مؤتمر صحفي الساعة 9 *fyh m&tmr SHfy AlsAEp 9* (there is a press conference at 9 o'clock) consists morphologically from the preposition في *fy* (in) and the enclitic pronoun ـه *h* (him). However, in this context, this morphological structure is irrelevant

because the enclitic pronoun is an impersonal pronoun without a referent. The entire word functions as an existential and is thus tagged as *fyh*/EX without tokenizing the final pronoun *h*. The same word in حنتجمع فيه *HntjmE fyh* (we'll gather in it) is treated differently because it literally means *in it*; thus it is first tokenized as *fy-h* and then tagged as *fy*/IN and *h*/3_SG_M_OBJP.

Prepositional Phrases (PPs) are not always literally used in EA. For example, بسرعة *bsrEp* - which morphologically consists of the preposition بـ *b* (with) and the noun سرعة *srEp* (speed) – functions as an adverb in انزلوا بسرعة ع الميدان *Anzlw bsrEp E AlmydAn* (come quickly to the square). Therefore, in this sentence, it is tokenized as one word and tagged as *bsrEp*/RB. In مهتم بسرعة حل الأزمة *mhtm bsrEp Hl Al>zmp* (he's concerned with a quick crisis solution), the same PP is literally used and thus it is first tokenized as *b-srEp* and then tagged as *b*/IN *srEp*/SG_F_NN. The same procedure is used with more complex PPs like بشكل مظبوط *b$kl mZbwT* (in a perfect way). In عايزين نحلها بشكل مظبوط *EAyzyn nHlhA b$kl mZbwT* (we want to sort it out in a perfect way), the PP functions as an adverb modifying the verb نحلها *nHlhA* (sort it out). Therefore, in this context it is tagged as one word – *b$kl_mZbwT*/RB.

Active participles in ATB are tagged as nouns or adjectives in POS annotation level and then a verbal noun (VN) tag is added in the treebank annotation level. In EA, active participles are not only used as nouns and adjectives but also as imperfect verbs. Thus they are tagged according to their grammatical function in context as NN, JJ or VBP. In عايزين نفهم *EAyzyn nfhm* (we want to understand), the active participle عايزين *EAyzyn* is used as a verb meaning *we want*; thus it is tagged as *Eyzyn*/1_PL_VBP. The same applies to فاهماهم *fAhmAhm* in هي مش فاهماهم *hy m$ fAhmAhm* (she does not understand them) in which *fAhmAhm* functions as a verb meaning *understand* attached to the object pronoun *them*. Thus it is tokenized as *fAhmA-hm* and tagged as *fAhmA*/3_SG_F_VBP and *hm*/3_PL_OBJP.

When active participles are used as nouns or adjectives, they are tagged as such. In هو شاهد اثبات *hw $Ahd AvbAt* (he's an prosecution witness), the active participle شاهد *$Ahd* (witness) is tagged as *$Ahd*/SG_M_NN being used as a noun. In المیه كافيه *Almyh kAfyh* (the water is enough), the active participle كافيه *kAfyh* is used as an adjective and is thus tagged as *kAfyh*/SG_F_JJ.

Modals – including verbal, nominal and adjunct modals – are also tokenized and POS tagged functionally. In يمكن نيجي *ymkn nyjy* (we may come), the modal verb يمكن *ymkn* (may) is tagged as *ymkn*/MD. The same modal function can be realized using an adjective form ممكن *mmkn* (may) as in ممكن مايكونش مفهوم *mmkn mAykwn$ mfhwm* (I may not be understood); which is thus tagged as *mmkn*/MD. The same adjective in a different context like كل شيء ممكن *kl $y' mmkn* (everything is possible) is not used modally and is tagged as *mmkn*/SG_M_JJ.

Modal adjuncts are typically multiword, yet should they be modals, they are tokenized and tagged as one unit. In احتمال انه ماينفعش في *fy AHtmAl Anh mAynfE$* (there is a possibility that it won't work), the modal adjunct احتمال في *fy AHtmAl* is tokenized and tagged as one word – *fy_AHtmAl*/MD. The same word احتمال *AHtmAl* can be tagged as a noun in a different context like احتمال فوزه ضعيف *AHtmAl fwzh DEyf* (his possibility/chance of winning is weak) – *AHtmAl*/SG_M_NN.

Multiword connective particles like وكمان *wkmAn* (and also) and وبعدين *wbEdyn* (and then) are also tokenized and tagged as one word. Each of these particles morphologically consists of the coordinating conjunction و *w* (and) and a connective particle. These are tokenized and tagged as *wkmAn*/CC and *wbEdyn*/ CC.

Multiword fixed expressions – that match our definition of multiword fixed expressions in section 3 – function as one whole unit to serve a certain pragmatic meaning. These are tokenized and tagged, thus, as one unit. The multiword fixed expression لا مؤاخذة *lA m&Axzp* (lit: no offense; gloss: excuse me) consists of the negative particle لا *lA* (no) and the noun مؤاخذة *m&Axzp* (offense); yet being a fixed frozen expression it is tagged as *lA_m&Axzp*/EXP. Similarly, موتوا بغيظكم *mwtwA bgyZkm* (lit: get lost with your anger; gloss: go to hell!) is not decomposed into a verb, a preposition, a noun and a possessive pronoun but is tokenized and tagged as *mwtwA_bgyZkm*/ EXP. Tagging the pragmatic values of these fixed expressions – both unigram and multiword – is kept for future work.

5 Corpus Description

Our corpus is 22,834 tweets complied over the period from May 2011 to December 2011. It is a subset of the microblog portion of YADAC (Al-Sabbagh and Girju, 2012). The corpus contains 423,691 tokens and 70,163 types preprocessed according to the procedures in the following lines.

Only tweets scored above the degree of dialectness threshold set by Al-Sabbagh and Girju (2012) are selected. This guarantees a highly dialectal corpus; yet MSA is still likely to be found. Arabic tweets written in Latin Script (AiLS) are already excluded as well as Foreign tweets written in Arabic Script (FiAS). Two normalization steps are used for spelling variation and speech effects.

To reduce the effect of spelling variation – given the lack of standard spelling conventions in EA and most Arabic dialects – we used a normalization rule-based module based on the conventions set by the Conventional Orthography for Dialectal Arabic (CODA) (Habash et al., 2011)[1]. To augment the performance of the spelling normalization module and deal with cases which CODA does not currently handle, we use the vowel-based spelling variation model and the 138-entry lexicon of unpredictable spelling variations both built by Al-Sabbagh and Girju (2 012).

Using a regular-expression module, we normalize speech effects like ياااااا سلاااام *yAAAAAAA slAAAAAm* (oh, wow!) that are used for pragmatic reasons, typically showing strong emotions. In addition to be tagged with their regular POS tag, the LNG tag is added to the tag vector of these words. Thus ياااااا سلاااام *yAAAAAAA slAAAAAm* is normalized as يا سلام *yA slAm* and being an EXP, it is tagged as *yA_slAm*/EXP_LNG. Similarly, جمييييل *jmyyyyl* (beautiful) is normalized as جميل *jmyl* and is tagged as *jmyl*/SG_M_JJ_LNG.

6 Gold-Standard Annotation

Two annotators of intermediate-level linguistic training (i.e. undergraduate linguistics students) – who are native EA speakers – are used to annotate the corpus over a period of 4 months. Two other annotators – graduate linguistics students – are then used to review the annotations for consistency and correctness over a period of one month. The first two annotators achieve an inter-annotator Kappa coefficient rate of 97.3% for tokenization and 88.5% for POS tagging. The review annotators achieve a rate of 99.6% for tokenization and 98.2% for POS tagging. Main differences between the two groups of annotators are about the consistency in applying our function-based annotation scheme.

The first annotation phase – tokenization – is a light stemming process to split off the following clitics and affixes:

- definite article الـ *Al* (the)
- prepositions بـ *b* (with) and لـ *l* (for)
- connective conjunctions و *w* (and) and فـ *f* (then);
- vocative particle يا *yA* (oh, hey)
- object pronouns ني *ny* (me), نا *nA* (us), ك *k* (you; singular), كم *km* (you; plural), ـه *h* (him), ها *hA* (her) and هم *hm* (them)
- possessive pronouns ي *y* (my), نا *nA* (our), ك *k* (your; singular), كم *km* (your; plural), ـه *h* (his), ها *hA* (her) and هم *hm* (their)
- aspectual progressive prefix بـ *b*
- future tense prefixes هـ *h* and حـ *H* (will)

Gender and number affixes are not split off, however, given that they affect semantic feature tagging: جميل *jmyl* (beautiful; masculine) is tagged as *jmyl*/SG_M_JJ; whereas جميلة *jmylp* (beautiful; feminine) is tagged as *jmylp*/SG_F_JJ.

Light stemming also involves reversing morphotactic changes resulted from clitic/affix attachment. Attaching a possessive pronoun to a noun ending in *ta' marbuta* – ـة p – changes it into *ta' maftouha* – ت *t* – as in قضية *qDyp* (issue) and قضيتي *qDyty* (my issue). Similarly, attaching an object pronoun to a plural verb ending in وا *wA* – the verb plural marker – leads to removing the *alef* as in شافوا *$AfwA* (they saw) and شافوني *$Afwny* (they saw me). When splitting off clitics and affixes, these morphotactic changes are reversed.

The second annotation phase – POS tagging – involves tagging words according to our function-based scheme and lexicon lookups. For fixed expressions that match our definition, we have a lexicon of 539 expressions that are 0.5% of our corpus. We also have a lexicon of unambiguous function words that contains 2,193 words. A lexicon from (Elghamry et al., 2008) is used for tagging the semantic features of gender and number. Words not found in the lexicon are manually labeled.

7 Transformation-Based Learning

Brill (1994) introduces Transformation-Based Learning (TBL) to POS tagging as an error-driven, corpus-based approach to induce tagging rules out of a gold-standard training corpus**.** It captures linguistic information in a small number of simple non-

[1] Nizar Habash, Mona Diab and Owen Rambow. 2012. Conventional Orthography for Dialectal Arabic (CODA) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May 2012, Istanbul, Turkey, 711-718

stochastic rules as opposed to large numbers of lexical and contextual probabilities.

AlGahtani et al. (2009) apply TBL to MSA tagging with a main modification of applying the algorithm to lexeme-affix combinations. Affixes are used as cues for POS tags, while affix-free words are looked up in a lexicon. For Out-Of-Vocabulary (OOV) words, they first use the Buckwalter Arabic Morphological Analyzer (BAMA) (Buckwalter, 2002) with a disambiguation module to pick the contextually correct word analysis; and second, words found in neither the lexicon nor BAMA output are tagged as proper nouns. Trained on 90% of ATB and tested on 10%, the algorithm achieves an accuracy rate of 96.5% which is comparable to the state-of-the-art results achieved for MSA using other algorithms like Hidden Markov Models (AlShamsi and Guessoum, 2006), Support Vector Machines (Diab, 2009) and memory-based learning approaches (Marsi and Bosch, 2005). That is why AlGahtani et al. (2009) argue that TBL is simple, non-complex, language-independent and of comparable results to other POS tagging approaches.

In this paper, we use Brill's TBL implementation for POS tagging and a tokenizer built on the same algorithm.

8 Evaluation and Discussion

We perform 10-fold cross validation and use standard precision, recall and F-measure as evaluation matrices. Our output is evaluated in three modes: tokenization (TOK) only, POS tagging without semantic feature labeling (POS) and POS tagging with tokenization and semantic features labeling (ALL). According to the results in table (2), the tokenization module achieves comparable results to the stat-of-the-art systems built for MSA. The performance of POS module and the semantic-feature module – that decreases performance dramatically – still need improvements.

	Precision	**Recall**	**F-Measure**
TOK	95%	94%	94.5%
POS	86.5%	88.8%	87.6%
ALL	81%	83.6%	82.3%

Table 2: Precision, recall and F-measure rates for the TOK, POS and ALL modules

About 6% of our corpus is three-letter words like أكل *>kl*. These words are highly ambiguous as they can have multiple readings based on the short vowel pattern with which they are produced. Given that EA text – like most MSA text – does not use diacritics to mark the short vowel patterns in text, these word are highly ambiguous. Our example *>kl* can read as *>kl* (food), *>kal* (he ate), *>akul* (I eat). The word مصر *mSr* can be *maSr* (Egypt) or *muSir* (persistent). Ambiguity increases when the word has more than one reading of the same grammatical category – nominal or verbal. If a word is ambiguous as a verb or a noun, the different contextual distribution of the verbs and nouns can resolve the ambiguity. Yet, if the word is ambiguous as noun or adjective – both are nominal categories – contextual distribution might not be as efficient for disambiguation. The same applies when the word is ambiguous between an imperfect verb and a perfect verb. Therefore, the adjective *muSir* (persistent) is always tagged as noun for *maSr* (Egypt) in our output. Similarly, *>akul* (I eat) is erroneously tagged as the perfect verb *>kal* (he ate).

The same thing applies to two-letter words. The word جد *gd* can be a noun meaning *grandfather*, a noun meaning *seriousness* or an adverb meaning *seriously*. Similarly, حب *Hb* can be a noun meaning *love* or *grains*, or a perfect verb meaning *he loved.*

Intra-grammatical-category ambiguity is also evident in longer words like المصري *AlmSry* (the Egyptian) which has two possible nominal tags – noun vs. adjective. Being both nominal, nouns and adjective occur in similar contexts and also share a considerable number of clitics and affixes, which might not be useful POS features in this case.

Ambiguous function words with lexical meanings also lead to output errors. For example, طيب *Tyb* is both an interjection meaning *then* as in *then what*? and an adjective meaning *kind*. In الشعب المصري طيب *Al$Eb AlmSry Tyb*, *Tyb* can mean both depending on how the sentence is read. With a rising final intonation, it means *then* (*the Egyptian people, then?*). With a falling final intonation, it means *kind* (*the Egyptian people is kind*). There is no way to represent intonation in written text and a wider context across multiple tweets is required to decide whether this tweet is a part of a conversation: if it is, then both intonations are possible and a deeper analysis of the conversation is required to know which intonation is intended; if not, then the falling intonation and thus the *kind* meaning is more likely.

The word بقى *bqY* can be a perfect verb meaning *remained* as in هو ده اللي بقى *hw dh Ally bqY* (this is what remained) or a discourse particle meaning *so* as in بقى هو ده الحل العبقري؟ *bqY hw dh AlHl AlEbqry* (so is this the genius solution?). Similarly, خلاص *xlAS*

can be a fixed unigram expression meaning *"that's it"* as in خلاص مش نافع *xlAS m$ nAfE* (that's it. It's not working) or a noun meaning *salvation* خلاص الناس دي على ايدهxlAS *AlnAs dy ElY Aydh* (the salvation of those people is through him).

Typos contribute less than 0.5% of errors. This might indicate that the corpus is not as noisy as it might have been assumed.

Our function-based scheme might have introduced ambiguities at this level of annotation because for example instead of tagging all instances of بسرعة *bsrEp* as *b*/IN and *srEp*/SG_F_NN, the algorithm has to learn when each instance is used as an adverb and when it is used as a PP. the same thing applies to all grammatical categories affected by our scheme. This, however, does not mean that these are *new* ambiguity types caused by our scheme; eventually these ambiguities will come up in other higher annotation layers.

Results in table (2) show that tagging the semantic features of gender, number, person, aspect, tense and mode decrease performance by about 5%. EA normalizes the morphological distinctions of many of these features and only through long dependencies – which are beyond our tagger – these features can be recovered. For example, كتبت *ktbt* can refer to a perfect verb in 1st person (I wrote), 2nd (you wrote) or 3rd feminine person (she wrote) based on how it is pronounced. With the absence of disambiguating diacritics in written EA, verbs of this class are highly ambiguous in terms of person.

The morphological distinction between duals and plurals is waived in the morphology of EA nouns, verbs, adjectives and pronouns. A plural form of any of these grammatical categories can refer to either duals or plurals. Long dependencies and sometimes metalinguistic information are required to recover the number feature. Alkuhlani and Habash (2012) conduct a series of experiments regarding recovering such latent semantic features; some of which are tried for EA in our future work.

9 Conclusion and Future Work

This paper presented a transformation-based POS tokenizer and tagger for Egyptian Arabic tweets. It proposed a function-based scheme in which words are tokenized and tagged based on their function rather than their morpho-syntactic structure. Among the grammatical categories in which morpho-syntactic structures and grammatical functions do not always map are existentials, prepositional phrases, active participles, modals, superlative adjectives, multiword connective particles and fixed expressions. The function-based algorithm is expected to enhance performance for higher-order NLP tools such as chunkers and parsers. Despite the promising results, which introduce a new NLP tool to the repository of the resource-poor EA language, much improvement is required.

Short-term improvement plans include: (1) using a different algorithm known for high performance on text processing tasks like Support Vector Machine and defining both tokenization and POS tagging as classification problems; (2) comparing the function-based scheme to ours to know how much ambiguity is resolved or introduced by our function-based scheme; and (3) comparing the two scheme in terms of their performance and enhancement for higher-order NLP tools.

Long-term improvement plans include: (1) building working on word sense disambiguation modules to improve performance on highly ambiguous words and (2) building modules to accommodate for such features as intonation that are unrecoverable from text, yet can affect performance.

References

Eric Brill. 1994. Some Advances in Rule-Based Part of Speech Tagging. *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, Washington, 722-727

Erwin Marsi and Antal van den Bosch. 2005 Memory-based Morphological Analysis Generation and Part-Of-Speech Tagging of Arabic. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages* 1-8. Ann Arbor: Association for Computational Linguistics

Fatma AlShamsi and Ahmed Guessoum. 2006. A Hidden Markov Model-Based POS Tagger for Arabic. *JADT 2006: 8th International Conference of Text Statistical Analysis.*

Hanaa Kilany H. Gadalla A. Arram, Yacoub, A. El-Habashi and C. McLemore. 2002. Egyptian Colloquial Arabic Lexicon. LDC catalog number LDC99L22

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. *The 6th International Conference on Informatics and Systems*, *INFOS2008*. Cairo University.

Jan Hajič, Otakar Smrž, Peter Zemánek, Jan Šnaidauf, Emanuel Beška. 2004. Prague Arabic Dependency Treebank: Development in Data and Tools. Proceedings of NEMLAR-2004, 110–117.

Kevin Duh, and Katrin Kirchhoff. 2005. POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, June 2005, 55-62

Khaled Elghamry, Rania Al-Sabbagh and Nagwa Elzeiny. 2008. Cue-Based Bootstrapping of Arabic Semantic Features. *Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data*, March 2008, Lyon, France, 85-95

Mona Diab. 2007. Improved Arabic Base Phrase Chunking with a New Enriched POS Tag Set. *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, Czech Republic, June 200, page 89-96

Mona Diab. 2009. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, Tagging and Base Phrase Chunking. Proceedings of the Second International Conference on Arabic Language Resources and Tools, 22-23 April 2009, Cairo, Egypt, 285-288

Mona Diab, Hacioglu, K., Jurafsky, D. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base-Phrase Chunks. In: S. Dumais, D.M., Roukos, S. (Eds.) *HLT-NAACL 2004: Short Papers*, pp. 149–152. Association for Computational Linguistics, Boston (2004)

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

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

Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. *Proceedings of COLING-ACL*, Sydney, Australia, 2006

Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic Treebank. Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, Suntec, Singapore, 4 August 2009, 221-224

Nizar Habash, Mona Diab and Owen Rambow. 2011. Conventional Orthography for Dialectal Arabic (CODA) Version 0.1 – July 2011. A Technical Report, Center for Computational Learning Systems – CCLS, Columbia University.

Nizar Habash, Ramy Eskander and Abdelatti Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON 2012)*, pages 1-9, Montreal, Canada, June 7, 2012

Rania Al-Sabbagh and Roxana Girju. 2012. YADAC: Yet Another Dialectal Arabic Corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, 23-25 May 2012, Istanbul, Turkey, 2882-2889

Sarah Alkuhlani and Nizar Habash. 2012. Identifying Broken Plurals, Irregular Gender, and Rationality in Arabic Text. *EACL 2012*, 675-685

Shabib AlGahtani, William Black and John McNaught. 2009. Arabic Part-Of-Speech Tagging Using Transformation-Based Learning. *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 2009

Shereen Khoja. 2001. APT: Arabic Part-of-Speech Tagger. *Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*, Carnegie Mellon University, Pittsburgh, Pennsylvania

Tim Buckwalter. 2002. Arabic Morphological Analyzer (AraMorph). Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49 & ISBN 1-58563-257-0

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic English Statistical Machine Translation. *Proceedings of the First Workshop on Algorithms and Resources for Modeling of Dialects and Language Varieties*, Edinburgh, Scotland, 10–21

Appendix A: A Detailed Description and Comparison of Our Tagset with ATB and RTS Tagsets

POS	ATB	RTS	Ours	Tag	Comments
Abbreviation	✓	✓	✓	ABR	Examples include titles like د. /*d.*/ (Dr.), ا. /*A.*/ (Mr.) and ج.م.ع /*j.m.E*/ (Arabic Republic of Egypt). It is noteworthy that RTS includes abbreviations with the NN tag that is also used for singular common nouns.
Accusative	✓	—	—	—	No case marking neither for EA – which does not consider it – nor for MSA tweets – when found – is given.

Adjective	✓	✓	✓	JJ	simple adjectives like كويس /*kwys*/ (good), شغال /*$gAl*/ (fine)
Adverb	✓	✓	✓	RB	
Case	✓	—	—	—	Case marking is not a feature of EA. Even for MSA tweets in our corpus that can result from users quoting press news and the like, case marking is ignored.
Command Verb	✓	✓	✓	VB	
Cardinal Number	✓	✓	✓	CD	
Comparative Adjective	✓	—	✓	JJR	
Connective Particle	✓	—	✓	CC	Both connective particles and coordinating conjunctions are collapsed into the CC tag. Both coordinate base and complex phrases. Examples of this tag include: و /*w*/ (and), وبعدين /*wbEdyn*/ (and then) and وكمان /*wkmAn*/ (and also).
Coordinating Conjunction	✓	✓			
Definite	✓	—	—	—	Definite particle is tokenized and tagged as DT. The information about whether a noun or an adjective is definite is thus structurally defined: in the tokenized corpus if a noun/adjective is preceded by the definite article, it is definite; otherwise it is not.
Demonstrative Pronoun	✓	✓	✓	DM	Demonstratives are phrase boundary markers and thus they are distinguished from determiners which are not.
Determiner	✓	✓	✓	DT	
Dialect	✓	—	—	—	Although the corpus contains MSA tweets coming mostly from users copying from press agencies, the dialect-standard distinction is not marked in the corpus.
Direct Object	✓	—	—	—	Object nouns are not tagged, but object pronouns are. They are split-off during tokenization and tagged as OBJP. Distinction between direct and indirect pronoun objects is not marked in our tagset given that it is structurally – rather than – morphologically defined: indirect object pronouns are preceded by a preposition but direct ones are not.
Dual	✓	—	✓	DU	The number features – singular, dual and plural – are given separate tags that can be combined with any relevant content word tag such as nouns, verbs, adjectives and personal pronouns.
Emphatic Particle	✓	—	—	—	
Existential	—	—	✓	EX	Neither ATB nor RTS label existentials probably because they are expressed in MSA via the imperfect verb يوجد /*ywjd*/ or the demonstrative هناك /*hnAk*/ both meaning *there is/are*. In EA, there are two possible forms of existentials: في /*fy*/ and فيه /*fyh*/ (lit: in; gloss: there is/are). Given that both existentials are ambiguous with the preposition *in* which has the same

					forms, and that existentials can be phrase boundary anchors, we use the EX tag to label them.
Feminine	✓	—	✓	F	
Focus Particle	✓	—	—	—	
Foreign Word	✓	✓	✓	FW	Tweets in Arabic script that contain one or more foreign words have these words labeled as FW.
Foreign Script	✓	—	—	—	Tweets entirely in foreign script whether they contain words from a foreign language – like English words – or contain Arabic words are filtered out in corpus preprocessing.
Future	✓	—	—	FT	Future tense is marked in EA by the verb prefixes هـ /*h*/ or حـ /*H*/ (will). These two are split out in tokenization and are tagged as FT for future. The same tag is used then for the MSA separate future particle سوف /*swf*/. Although it does not mark phrase boundaries, it is important for verb tense identification.
Future Particle	✓	—	—	—	
Genitive	✓	—	—	—	
Imperfect Verb	✓	✓	✓	VBP	The tense of an imperfect verb can be present, future or progressive. This fine-grained tense classification is not represented by the tagset; yet this information is structurally predictable given that the split-off affixes indicating each tense are POS tagged.
Indefinite	✓	—	—	—	
Indicative	✓	—	—	—	
Interjection	✓	✓	✓	UH	
Interrogative Adverb	✓	✓	✓	INT	Interrogative adverbs like إزاي /*<zAy*/ (how) and فين /*fyn*/ (where) and interrogative pronouns like ايه /*Ayh*/ (what) and مين /*myn*/ (who) are all collapsed into the tag INT. Interrogative particles like هل /*hl*/ and أ /*>*/ used in MSA to form yes/no questions are not used in EA, yet if they exist in MSA tweets, they are also labeled as INT.
Interrogative Particle	✓	—			
Interrogative Pronoun	✓	—			
Jussive Particle	✓	—	—	—	
Masculine	✓	—	✓	M	The gender features – masculine and feminine – are given separate tags that can be combined with any relevant content word tag such as nouns, verbs, adjectives and personal pronouns.
Mood	✓	—	—	—	
Negative Particle	✓	—	✓	NG	Negative particles come in two forms: a circumfix *mA...$* as in ماراحوش /*mArAHw$*/ (they didn't go), and a number of free morphemes including مش /*m$*/ and لا /*lA*/ both meaning *no* among others. Bound and free negative

					particles are both labeled as NG; although the free-morpheme form does not mark phrase boundaries, it is an important marker for sentiment analysis – one goal for building this tagger.
Noun	✓	—	✓	NN	NN is used for common nouns whether singular, dual, plural or collective nouns. The gender and number features are given their own tags.
Noun Quantifiers	✓	—	✓	QNT	Quantifiers like شوية /*$wyp*/ (a little), كتير /*ktyr*/ (a lot) among others are given their own tag – QNT.
Noun Suffix	✓	—	—	—	This is the same as the possessive pronoun – given the PP$ tag. This is the only suffix that is split-off; whereas gender and number suffixes are not.
Nominative	✓	—	—	—	
Ordinal Number	✓	—	✓	OD	
Passive	✓	✓	✓	P	The passive feature is not reflected in the verb tagset for simplicity, yet a P tag combined with a verb tag indicates a passive voice and the absence of the P tag indicates an active voice – the default.
Particle	✓	✓	✓	PRT	All particles that (1) do not mark phrase boundaries and (2) do not mark verb tense or negation are collapsed in one tag – PRT.
Partial Word	✓	—	✓	PW	Given that each tweet is limited to 140 characters, users who go over the limit produce incomplete erroneous words. These count about 0.3% of our corpus.
Perfect Verb	✓	—	✓	VBD	The VBD tag in our tagset does not define voice (active vs. passive); it only defines the perfect aspect of the verb. Voice is given a separate tag – P.
Person	✓	—	✓	1|2|3	1st, 2nd and 3rd person
Plural	✓	—	✓	PL	
Possessive Pronoun	✓	✓	✓	PP$	
Preposition	✓	✓	✓	IN	
Progressive	—	—	✓	PG	One verb prefix that is specific to EA – in comparison to MSA – is the progressive prefix بـ /*b*/ as in بيقول /*byqwl*/ (he's saying) and بنعافر /*bnEAfr*/ (we're struggling). The progressive prefix is split-off in tokenization and is tagged as PG.
Pronoun	✓	✓	✓	OBJP SBJP	Pronouns are split based on their grammatical functions into: object pronouns – typically tokenized from the verb endings – and subject pronouns – which are the free-morpheme subject pronouns here. Possessive pronouns are also given their own tag – PP$.
Proper Noun	✓	✓	✓	NNP	NNP refers to proper nouns whether they are singular, dual or plural. Number features are tagged with a separate set of tags.

Pseudo Verb	✓	—	—	—	
Punctuation	✓	✓	✓	PNC	
Relative Adverb	✓	—	✓	RL	There is only one relative pronoun in EA – اللي /*Ally*/ (who, which, that). Even when relative pronouns/adverbs from MSA are encounter they are given the RL tag.
Relative Pronoun	✓	✓			
Response Conditional Particle	✓	—	✓	CN	Conditional particles are phrase boundary markers.
Restrictive Particle	✓	—	—	—	
Singular	✓	—	✓	SG	
Subjunctive	✓	—	—	—	
Subordinate Conjunction	✓	✓	✓	SC	Like CN, SC are phrase boundary markers.
Suffix	✓	—	—	—	Noun suffixes are the possessive pronouns only, given that our light stemming approach does not split off gender and number suffixes. Possessive pronouns are given the PP$ tag and gender and number features are represented by their own tags.
Superlative Adjective	—	—	✓	JJS	Although superlative adjectives are not morphologically marked, they have implications for sentiment analysis and thus they are tagged as JJS.
Transcription Error	✓	—	—	—	
Typo	✓	—	—	—	
Verb	✓	—	✓	—	
Verbal Noun	✓	—	—	—	Verbal nouns or deverbal nouns are tagged as common nouns
Verbal Adjective	✓	—	—	—	Verbal adjectives are tagged as adjectives.
Vocative Particle	✓	—	✓	VC	
Not found words	—	✓	✓	NF	
Fixed Expressions	—	—	✓	EXP	
Modals	—	—	✓	MD	Fine-grained distinctions between different modality types – epistemic, evidential, deontic and volitive – are not marked in this tag. All linguistic modality types and their verbal or nominal realizations are labeled as MD.

Twitter Mentions	—	—	✓	MNT	
Twitter Hashtags	—	—	✓	HSH	
Twitter Emoticons	—	—	✓	EMO	
Twitter URLs	—	—	✓	URL	
Lengthened Words	—	—	✓	LNG	