

Twitter Universal Dependency Parsing for African-American and Mainstream American English

Su Lin Blodgett Johnny Tian-Zheng Wei Brendan O’Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

blodgett@cs.umass.edu jwei@umass.edu brenocon@cs.umass.edu

Abstract

Due to the presence of both Twitter-specific conventions and non-standard and dialectal language, Twitter presents a significant parsing challenge to current dependency parsing tools. We broaden English dependency parsing to handle social media English, particularly social media African-American English (AAE), by developing and annotating a new dataset of 500 tweets, 250 of which are in AAE, within the Universal Dependencies 2.0 framework. We describe our standards for handling Twitter- and AAE-specific features and evaluate a variety of cross-domain strategies for improving parsing with no, or very little, in-domain labeled data, including a new data synthesis approach. We analyze these methods’ impact on *performance disparities* between AAE and Mainstream American English tweets, and assess parsing accuracy for specific AAE lexical and syntactic features. Our annotated data and a parsing model are available at: <http://slanglab.cs.umass.edu/TwitterAAE/>.

1 Introduction

Language on Twitter diverges from well-edited Mainstream American English (MAE, also called Standard American English) in a number of ways, presenting significant challenges to current NLP tools. It contains, among other phenomena, non-standard spelling, punctuation, capitalization, and syntax, as well as Twitter-specific conventions such as hashtags, usernames, and retweet tokens (Eisenstein, 2013). Additionally, it contains an abundance of dialectal language, includ-

ing African-American English (AAE), a dialect of American English spoken by millions of individuals, which contains lexical, phonological, and syntactic features not present in MAE (Green, 2002; Stewart, 2014; Jones, 2015).

Since standard English NLP tools are typically trained on well-edited MAE text, their performance is degraded on Twitter, and even more so for AAE tweets compared to MAE tweets—gaps exist for part-of-speech tagging (Jørgensen et al., 2016), language identification, and dependency parsing (Blodgett et al., 2016; Blodgett and O’Connor, 2017). Expanding the linguistic coverage of NLP tools to include minority and colloquial dialects would help support equitable language analysis across sociolinguistic communities, which could help information retrieval, translation, or opinion analysis applications (Jurgens et al., 2017). For example, sentiment analysis systems ought to count the opinions of all types of people, whether they use standard dialects or not.

In this work, we broaden Universal Dependencies (Nivre et al., 2016) parsing¹ to better handle social media English, in particular social media AAE. First, we develop standards to handle Twitter-specific and AAE-specific features within Universal Dependencies 2.0 (§3), by selecting and annotating a new dataset of 500 tweets, 250 of which are in AAE.

Second, we evaluate several state-of-the-art dependency parsers, finding that, as expected, they perform poorly on our dataset relative to the UD English Treebank (§4). Third, since the UD English Treebank contains substantial amounts of traditional MAE data for training, we investigate cross-domain training methods to improve Twitter AAE dependency parsing with no, or very little,

¹<http://universaldependencies.org/>

in-domain labeled data, by using Twitter-specific taggers, embeddings, and a novel heuristic training data synthesis procedure. This helps close some of the gap between MAE and AAE performance. Finally, we provide an error analysis of the parsers' performance on AAE lexical and syntactic constructions in our dataset (§5.4).²

2 Related Work

2.1 Parsing for Twitter

Parsing for noisy social media data presents interesting and significant challenges. Foster et al. (2011) develop a dataset of 519 constituency-annotated English tweets, which were converted to Stanford dependencies. Their analysis found a substantial drop in performance of an off-the-shelf dependency parser on the new dataset compared to a WSJ test set. Sanguinetti et al. (2017) annotated a dataset of 6,738 Italian tweets according to UD 2.0 and examined the performance of two parsers on the dataset, finding that they lagged considerably relative to performance on the Italian UD Treebank.

Kong et al. (2014) develop an English dependency parser designed for Twitter, annotating a dataset of 929 tweets (TWEEBANK V1) according to the unlabeled FUDG dependency formalism (Schneider et al., 2013). It has substantially different structure than UD (for example, prepositions head PPs, and auxiliaries govern main verbs).

More recently, Liu et al. (2018) developed TWEEBANK V2, fully annotating TWEEBANK V1 according to UD 2.0 and annotating additionally sampled tweets, for a total of 3,550 tweets. They found that creating consistent annotations was challenging, due to frequent ambiguities in interpreting tweets; nevertheless, they were able to train a pipeline for tokenizing, tagging, and parsing the tweets, and develop ensemble and distillation models to improve parsing accuracy. Our work encounters similar challenges; in our approach, we intentionally oversample AAE-heavy messages for annotation, detail specific annotation decisions for AAE-specific phenomena (§3.2), and analyze parser performance between dialects and for particular constructions (§5.3–5.4). Future work may be able to combine these annotations for effective multi-dialect Twitter UD parsers, which

²Our annotated dataset and trained dependency parser are available at <http://slanglab.cs.umass.edu/TwitterAAE/> and annotations are available in the public Universal Dependencies repository.

may allow for the use of pre-existing downstream tools like semantic relation extractors (e.g. White et al. (2016)).

One line of work for parsing noisy social media data, including Khan et al. (2013) and Nasr et al. (2016), examines the effects of the domain mismatches between traditional sources of training data and social media data, finding that matching the data as closely as possible aids performance. Other work focuses on normalization, including Daiber and van der Goot (2016) and van der Goot and van Noord (2017), which develop a dataset of 500 manually normalized and annotated tweets, and uses normalization within a parser. Separately, Zhang et al. (2013) created a domain-adaptable, parser-focused system by directly linking parser performance to normalization performance.

2.2 Parsing for Dialects

For Arabic dialects, Chiang et al. (2006) parse Levantine Arabic by projecting parses from Modern Standard Arabic translations, while Green and Manning (2010) conduct extensive error analysis of Arabic constituency parsers and the Penn Arabic Treebank. Scherrer (2011) parse Swiss German dialects by transforming Standard German phrase structures. We continue in this line of work in our examination of AAE-specific syntactic structures and generation of synthetic data with such structures (§4.2.1).

Less work has examined parsing dialectal language on social media. Recently, Wang et al. (2017) annotate 1,200 Singlish (Singaporean English) sentences from a Singaporean talk forum, selecting sentences containing uniquely Singaporean vocabulary items. Like other work, they observe a drop in performance on dialectal Singlish text, but increase performance through a stacking-based domain adaptation method.

3 Dataset and Annotation

3.1 Dataset

Our dataset contains 500 tweets, with a total of 5,951 non-punctuation edges, sampled from the publicly available TwitterAAE corpus.³ Each tweet in that corpus is accompanied by a model's demographically-aligned topic model probabilities jointly inferred from Census demographics and word likelihood by Blodgett et al. (2016), including the African-American and White topics.

³<http://slanglab.cs.umass.edu/TwitterAAE/>

We create a balanced sample to get a range of dialectal language, sampling 250 tweets from those where the African-American topic has at least 80% probability, and 250 from those where the White topic has at least 80% probability. We refer to these two subcorpora as AA and WH; [Blodgett et al. \(2016\)](#) showed the former exhibits linguistic features typical of AAE.

The 250 AA tweets include many alternate spellings of common words that correspond to well-known phonological phenomena—including *da*, *tha* (the), *dat*, *dhat* (that), *dis*, *dhis* (this), *ion*, *iont* (I don't), *ova* (over), *yo* (your), *dere*, *der* (there), *den*, *dhen* (then), *ova* (over), and *nall*, *null* (no, nah)—where each of the mentioned italicized AAE terms appears in the AAE data, but never in the MAE data. We examine these lexical variants more closely in §5.4. Across the AA tweets, 18.0% of tokens were not in a standard English dictionary, while the WH tweets' OOV rate was 10.7%.⁴ We further observe a variety of AAE syntactic phenomena in our AA tweets, several of which are described in §3.2 and §5.4.

3.2 Annotation

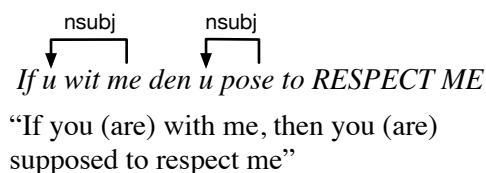
To effectively measure parsing quality and develop better future models, we first focus on developing high-quality annotations for our dataset, for which we faced a variety of challenges. We detail our annotation principles using Universal Dependency 2.0 relations ([Nivre et al., 2016](#)).

All tweets were initially annotated by two annotators, and disagreements resolved by one of the annotators. Annotation decisions for several dozen tweets were discussed in a group of three annotators early in the annotation process.

Our annotation principles are in alignment with those proposed by [Liu et al. \(2018\)](#), with the exception of contraction handling, which we discuss briefly in §3.2.2.

3.2.1 Null Copulas

The AAE dialect is prominently characterized by the drop of copulas, which can occur when the copula is present tense, not first person, not accented, not negative, and expressing neither the habitual nor the remote present perfect tenses ([Green, 2002](#)). We frequently observed null copulas, as in:

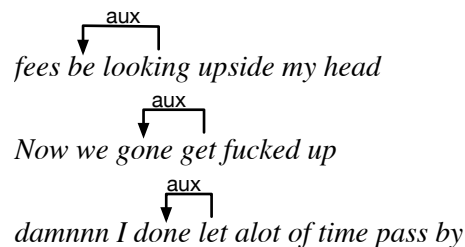


If u wit me den u pose to RESPECT ME
 “If you (are) with me, then you (are) supposed to respect me”

The first dropped *are* is a null copula; UD2.0 would analyze the MAE version as *you* \xleftarrow{nsubj} *me* \xrightarrow{cop} *are*, which we naturally extend to analyze the null copula by simply omitting *cop* (which is now over a null element, so cannot exist in a dependency graph). The second *are* is a null auxiliary (in MAE, *you* \xleftarrow{nsubj} *supposed* \xrightarrow{aux} *are*), a tightly related phenomenon (for example, [Green et al. \(2007\)](#) studies both null copulas and null auxiliary *be* in infant AAE), which we analyze similarly by simply omitting the *aux* edge.

3.2.2 AAE Verbal Auxiliaries

We observed AAE verbal auxiliaries, e.g.,



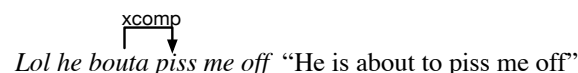
fees be looking upside my head
Now we gone get fucked up
damnnn I done let alot of time pass by

including habitual *be* (“Continually, over and over, fees are looking at me...”), future *gone* (“we are going to get...”), and completive *done* (“I did let time pass by,” emphasizing the speaker completed a time-wasting action).

We attach the auxiliary to the main verb with the *aux* relation, as UD2.0 analyzes other English auxiliaries (e.g. *would* or *will*).

3.2.3 Verbs: Auxiliaries vs. Main Verbs

We observed many instances of quasi-auxiliary, “-to” shortened verbs such as *wanna*, *gotta*, *finna*, *bouta*, *tryna*, *gonna*, which can be glossed as *want to*, *got to*, *fixing to*, *about to*, etc. They control modality, mood and tense—for example, *finna* and *bouta* denote an immediate future tense; [Green \(2002, ch. 2\)](#) describes *finna* as a preverbal marker. From UD’s perspective, it is difficult to decide if they should be subordinate auxiliaries or main verbs. In accordance with the UD Treebank’s handling of MAE *want to X* and *going to X* as main verbs (*want* \xrightarrow{xcomp} *X*), we analyzed them similarly, e.g.



Lol he bouta piss me off “He is about to piss me off”

⁴The dictionary of 123,377 words with American spellings was generated using <http://wordlist.aspell.net/>.

This is an instance of a general principle that, if there is a shortening of an MAE multiword phrase into a single word, the annotations on that word should mirror the edges in and out of the original phrase’s subgraph (as in [Schneider et al. \(2013\)](#)’s fudge expressions).

However, in contrast to the UD Treebank, we did not attempt to split up these words into their component words (e.g. *wanna* → *want to*), since to do this well, it would require a more involved segmentation model over the dozens or even hundreds of alternate spellings each of the above can take;⁵ we instead rely on [Owoputi et al. \(2013\)](#); [O’Connor et al. \(2010\)](#)’s rule-based tokenizer that never attempts to segment within such shortenings. This annotation principle is in contrast to that of [Liu et al. \(2018\)](#), which follows UD tokenization for contractions.

3.2.4 Non-AAE Twitter issues

We also encountered many issues general to Twitter but not AAE; these are still important to deal with since AAE tweets include more non-standard linguistic phenomena overall. When possible, we adapted [Kong et al. \(2014\)](#)’s annotation conventions into the Universal Dependencies context, which are the only published conventions we know of for Twitter dependencies (for the FUDG dependency formalism). Issues include:

- @-mentions, which require different treatment when they are terms of address, versus nominal elements within a sentence.
- Hashtags, which in their tag-like usage are utterances by themselves (*#tweetliketheoppositegender Oh damn .*) or sometimes can be words with standard syntactic relations within the sentence (*#She’s A Savage*, having *#She’s* $\xleftarrow{\text{nsubj}}$ *Savage*). Both hashtag and @-mention ambiguities are handled by [Owoputi et al. \(2013\)](#)’s POS tagger.
- Multiple utterances, since we do not attempt sentence segmentation, and in many cases sentential utterances are not separated by explicit punctuation. FUDG allows for multiple roots for a text, but UD does not; instead we follow UD’s convention of the *parataxis* relation for what they describe as “side-by-side run-on sentences.”

⁵For example, [Owoputi et al. \(2013\)](#)’s Twitter word cluster 0011000 has 36 forms of *gonna* alone: http://www.cs.cmu.edu/~ark/TweetNLP/cluster_viewer.html

- Emoticons and emoji, which we attach as *discourse* relations to the utterance root, following UD’s treatment of interjections.
- Collapsed phrases, like *omw* for “on my way.” When possible, we used the principle of annotating according to the root of the subtree of the original phrase. For example, UD 2.0 prescribes *way* $\xrightarrow{\text{xcomp}}$ *get* for the sentence *On my way to get...*; therefore we use *omw* $\xrightarrow{\text{xcomp}}$ *get* for *omw to get*.
- Separated words, like *uh round* for “around,” which we analyze as multiword phrases (*flat* or *compound*).

We discuss details for these and other cases in the online appendix.

4 Experiments

4.1 Models

Our experiments use the following two parsers.

UDPipe ([Straka et al., 2016](#)) is a neural pipeline containing a tokenizer, morphological analyzer, tagger, and transition-based parser intended to be easily retrainable. The parser attains 80.2% LAS (labeled attachment score) on the UD English treebank with automatically generated POS tags, and was a baseline system used in the CoNLL 2017 Shared Task ([Zeman et al., 2017](#)).⁶

Deep Biaffine ([Dozat et al., 2017](#); [Dozat and Manning, 2016](#)) is a graph-based parser incorporating neural attention and biaffine classifiers for arcs and labels. We used the version of the parser in the Stanford CoNLL 2017 Shared Task submission, which attained 82.2% LAS on the UD English treebank with automatically generated tags, and achieved the best performance in the task. The model requires pre-trained word embeddings.⁷

4.2 Experimental Setup

We considered a series of experiments within both a cross-domain scenario (§4.2.1), where we trained only on UD Treebank data, and an in-domain scenario (§4.2.2) using small amounts of our labeled data. We use the parsing systems’ default hyperparameters (e.g. minibatch size and learning rate) and the default training/development split of the treebank (both systems perform early stopping based on development set performance).

⁶<https://github.com/ufal/udpipe>

⁷<https://github.com/tdozat/UnstableParser/>

4.2.1 Cross-Domain Settings

Morpho-Tagger vs. ARK POS tags: The UD Treebank contains extensive fine-grained POS and morphological information, on which UDPipe’s morphological analyzer and tagging system is originally trained. This rich information should be useful for parsing, but the analyzers may be highly error-prone on out-of-domain, dialectal Twitter data, and contribute to poor parsing performance. We hypothesize that higher quality, even if coarser, POS information should improve parsing.

To test this, we retrain UDPipe in two different settings. We first retrain the parser component with fine-grained PTB-style POS tags and morphological information provided by the tagger component;⁸ we call this the *Morpho-Tagger* setting. Second, we retrain the parser with morphological information stripped and its tags predicted from the ARK Twitter POS tagger (Owoputi et al., 2013), which is both tailored for Twitter and displays a smaller AAE vs MAE performance gap than traditional taggers (Jørgensen et al., 2016); we call this the *ARK Tagger* setting.⁹ The ARK Tagger’s linguistic representation is impoverished compared to Morpho-Tagger: its coarse-grained POS tag system does not include tense or number information, for example.¹⁰

Synthetic Data: Given our knowledge of Twitter- and AAE-specific phenomena that do not occur in the UD Treebank, we implemented a rule-based method to help teach the machine-learned parser these phenomena; we generated synthetic data for three Internet-specific conventions and one set of AAE syntactic features. (This is inspired by Scherrer (2011)’s rule transforms between Standard and Swiss German.) We performed each of the following transformations separately on a copy of the UD Treebank data and concatenated the transformed files together for the final training and development files, so that each final file contained several transformed copies of the original UD Treebank data.

1. *@-mentions, emojis, emoticons, expressions, and hashtags:* For each sentence in the UD Treebank we inserted at least one @-mention, emoji, emoticon, expression (Internet-specific words and

⁸We also retrained this component, to maintain consistency of training and development split. We also remove the universal (coarse) POS tags it produces, replacing them with the same PTB tags.

⁹We strip lemmas from training and development files for both settings.

¹⁰Derczynski et al. (2013)’s English Twitter tagger, which outputs PTB-style tags, may be of interest for future work.

abbreviations such as *lol*, *kmsl*, and *xoxo*), or hashtag, annotated with the correct relation, at the beginning of the sentence. An item of the same type was repeated with 50% probability, and a second item was inserted with 50% probability. @-mentions were inserted using the *ATTENTION* token and emojis using the *EMOJI* token. Emoticons were inserted from a list of 20 common emoticons, expressions were inserted from a list of 16 common expressions, and hashtags were sampled for insertion according to their frequency in a list of all hashtags observed in the TwitterAAE corpus.

2. *Syntactically participating @-mentions:* To replicate occurrences of syntactically participating @-mentions, for each sentence in the UD Treebank with at least one token annotated with an *nsubj* or *obj* relation and an *NNP* POS tag, we replaced one at random with the *ATTENTION* token.

3. *Multiple utterances:* To replicate occurrences of multiple utterances, we randomly collapsed pairs of two short sentences (< 15 tokens) together, attaching the root of the second to the root of the first with the *parataxis* relation.

4. *AAE preverbal markers and auxiliaries:* We introduced instances of verbal constructions present in AAE that are infrequent or non-existent in the UD Treebank data. First, constructions such as *going to*, *about to*, and *want to* are frequently collapsed to *gonna*, *bouta*, and *wanna*, respectively (see §3.2.2); for each sentence with at least one of these constructions, we randomly chose one to collapse. Second, we randomly replaced instances of *going to* with *finna*, a preverbal marker occurring in AAE and in the American South (Green, 2002). Third, we introduced the auxiliaries *gone* and *done*, which denote future tense and past tense, respectively; for the former, for each sentence containing at least one auxiliary *will*, we replace it with *gone*, and for the latter, for each sentence containing at least one non-auxiliary, non-passive, past-tense verb, we choose one and insert *done* before it. Finally, for each sentence containing at least one copula, we delete one at random.

Word Embeddings: Finally, since a tremendous variety of Twitter lexical items are not present in the UD Treebank, we use 200-dimensional word embeddings that we trained with *word2vec*¹¹ (Mikolov et al., 2013) on the

¹¹<https://github.com/dav/word2vec>

TwitterAAE corpus, which contains 60.8 million tweets. Before training, we processed the corpus by replacing @-mentions with ATMENTION, replacing emojis with EMOJI, and replacing sequences of more than two repeated letters with two repeated letters (e.g. *partyyyyy* → *partyy*). This resulted in embeddings for 487,450 words.

We retrain and compare UDPipe on each of the *Morpho-Tagger* and *ARK Tagger* settings with synthetic data and pre-trained embeddings, and without. We additionally retrain Deep Biaffine with and without synthetic data and embeddings.¹²

4.2.2 In-domain Training

We additionally investigate the effects of small amounts of in-domain training data from our dataset. We perform 2-fold cross-validation, randomly partitioning our dataset into two sets of 250 tweets. We compare two different settings (all using the UDPipe *ARK Tagger* setting):

Twitter-only: To explore the effect of training with Twitter data alone, for each set of 250 we trained on that set alone, along with our Twitter embeddings, and tested on the remaining 250.

UDT+Twitter: To explore the additional signal provided by the UD Treebank, for each set of 250 we trained on the UD Treebank concatenated with that set (with the tweets upweighted to approximately match the size of the UD Treebank, in order to use similar hyperparameters) and tested on the remaining 250.

5 Results and Analysis

In our evaluation, we ignored punctuation tokens (labeled with *punct*) in our LAS calculation.

5.1 Effects of Cross-Domain Settings

***Morpho-Tagger* vs. *ARK Tagger*:** As hypothesized, UDPipe’s *ARK Tagger* setting outperformed the *Morpho-Tagger* across all settings, ranging from a 2.8% LAS improvement when trained only on the UD Treebank with no pre-trained word embeddings, to 4.7% and 5.4% improvements when trained with Twitter embeddings and both Twitter embeddings and synthetic data, respectively. The latter improvements suggest that the *ARK Tagger* setup is able to take better advantage of Twitter-specific lexical information from the embeddings

¹²As the existing implementation of Deep Biaffine requires pre-trained word embeddings, for the Deep Biaffine baseline experiments we use the CoNLL 2017 Shared Task 100-dimensional embeddings that were pretrained on the English UD Treebank.

Model	LAS
(1) UDPipe, Morpho-Tagger, UDT	50.5
(2) + Twitter embeddings	53.9
(3) + synthetic, Twitter embeddings	58.9
(4) UDPipe, ARK Tagger, UDT	53.3
(5) + Twitter embeddings	58.6
(6) + synthetic, Twitter embeddings	64.3
Deep Biaffine, UDT	
(7) + CoNLL MAE embeddings	62.3
(8) + Twitter embeddings	63.7
(9) + synthetic, Twitter embeddings	65.0

Table 1: Results from cross-domain training settings (see §4.2.1).

Model	LAS
(10) UDPipe, Twitter embeddings	62.2
(11) + UDT	70.3

Table 2: Results from in-domain training settings (with the *ARK Tagger* setting, see §4.2.2).

and syntactic patterns from the synthetic data. Table 1 shows the LAS for our various settings.

After observing the better performance of the *ARK Tagger* setting, we opted not to retrain the Deep Biaffine parser in any *Morpho-Tagger* settings due to the model’s significantly longer training time; all our Deep Biaffine results are reported for models trained with an *ARK Tagger* setting.

Synthetic data and embeddings: We observed that synthetic data and Twitter-trained embeddings were independently helpful; embeddings provided a 1.4–5.3% boost across the UDPipe and Deep Biaffine models, while synthetic data provided a 1.3–5.7% additional boost (Table 1).

UDPipe vs. Deep Biaffine: While the baseline models for UDPipe and Deep Biaffine are not directly comparable (since the latter required pre-trained embeddings), in the Twitter embeddings setting Deep Biaffine outperformed UDPipe by 5.1%. However, given access to both synthetic data and Twitter embeddings, UDPipe’s performance approached that of Deep Biaffine.

5.2 Effects of In-Domain Training

Perhaps surprisingly, training with even limited amounts of in-domain training data aided in parsing performance; training with just in-domain data produced an LAS comparable to that of the baseline Deep Biaffine model, and adding UD Treebank data further increased LAS by 8.1%, indicat-

Model	AA LAS	WH LAS	Gap
(1) UDPipe, Morpho-Tagger	43.0	57.0	14.0
(2) + Twitter embeddings	45.5	61.2	15.7
(3) + synthetic, Twitter embeddings	50.7	66.2	15.5
(4) UDPipe, ARK Tagger	50.2	56.1	5.9
(5) + Twitter embeddings	54.1	62.5	8.4
(6) + synthetic, Twitter embeddings	59.9	68.1	8.2
Deep Biaffine, ARK Tagger			
(7) + CoNLL MAE embeddings	56.1	67.7	11.6
(8) + Twitter embeddings	58.7	66.7	8.0
(9) + synthetic, Twitter embeddings	59.9	70.8	10.9

Table 3: AA and WH tweets’ labeled attachment scores for UD Treebank-trained models (see §5.3 for discussion); *Gap* is the WH – AA difference in LAS.

ing that they independently provide critical signal.

5.3 AAE/MAE Performance Disparity

For each model in each of the cross-domain settings, we calculated the LAS on the 250 tweets drawn from highly African-American tweets and the 250 from highly White tweets (see §3 for details); we will refer to these as the AA and WH tweets, respectively. We observed clear disparities in performance between the two sets of tweets, ranging from 5.9% to 15.7% (Table 3). Additionally, across settings, we observed several patterns.

First, the UDPipe *ARK Tagger* settings produced significantly smaller gaps (5.9–8.4%) than the corresponding *Morpho-Tagger* settings (14.0–15.7%). Indeed, most of the performance improvement of the *ARK Tagger* setting comes from the AA tweets; the LAS on the AA tweets jumps 7.2–9.2% from each *Morpho-Tagger* setting to the corresponding *ARK Tagger* setting, compared to differences of –0.9–1.9% for the WH tweets.

Second, the Deep Biaffine *ARK Tagger* settings produced larger gaps (8.0–11.6%) than the UDPipe *ARK Tagger* settings, with the exception of the embeddings-only setting.

Finally, we observed the surprising result that adding Twitter-trained embeddings and synthetic data, which contains both Twitter-specific and AAE-specific features, increases the performance gap across both UDPipe settings. We hypothesize that while UDPipe is able to effectively make use of both Twitter-specific lexical items and annotation conventions within MAE-like syntactic structures, it continues to be stymied by AAE-like syntactic structures, and is therefore unable to make use of the additional information.

We further calculated recall for each relation

type across the AA tweets and WH tweets, and the resulting performance gap, under the UDPipe *Morpho-Tagger* and *ARK Tagger* models trained with synthetic data and embeddings. Table 4 shows these calculations for the 15 relation types for which the performance gap was highest and which had at least 15 instances in each of the AA and WH tweet sets, along with the corresponding calculation under the *ARK Tagger* model. The amount by which the performance gap is reduced from the first setting to the second setting is also reported. Of the 15 relations shown, the gap was reduced for 14, and 7 saw a reduction of at least 10%.

5.4 Lexical and Syntactic Analysis of AAE

In this section, we discuss AAE lexical and syntactic variations observed in our dataset, with the aim of providing insight into decreased AA parsing accuracy, and the impact of various parser settings on their parsing accuracy.

AAE contains a variety of phonological features which present themselves on Twitter through a number of lexical variations (Green, 2002; Jones, 2015), many of which are listed in §3.1, instances of which occur a total of 80 times in the AA tweets; notably, none occur in the WH tweets.

We investigated the accuracy of various cross-domain parser settings on these lexical variants; for each of the baseline *Morpho-Tagger*, baseline *ARK Tagger*, *ARK Tagger* with embeddings, and *ARK Tagger* with synthetic data and embeddings models, we counted the number of instances of lexical variants from §3.1 for which the model gave the correct head with the correct label.

While the lexical variants challenged all four models, switching from the *Morpho-Tagger* set-

Relation	Morpho-Tagger			ARK Tagger			Reduction
	AA Recall	WH Recall	Gap (WH - AA)	AA Recall	WH Recall	Gap (WH - AA)	
<i>compound</i>	36.4	71.2	34.8	42.4	72.9	30.5	4.4
<i>obl:tmob</i>	25.0	51.7	26.7	43.8	55.2	11.4	15.3
<i>nmod</i>	28.6	54.4	25.8	45.7	51.5	5.8	20.1
<i>cop</i>	56.5	82.1	25.6	65.2	79.1	13.9	11.7
<i>obl</i>	41.4	65.4	24.0	56.8	62.5	5.7	18.3
<i>cc</i>	56.9	79.0	22.1	78.5	82.7	4.3	17.8
<i>ccomp</i>	33.3	54.2	20.8	40.5	54.2	13.7	7.1
<i>obj</i>	61.3	81.5	20.2	72.8	83.5	10.7	9.5
<i>case</i>	60.5	79.8	19.3	75.2	83.4	8.2	11.1
<i>det</i>	73.1	90.7	17.5	83.4	92.2	8.8	8.7
<i>advmod</i>	53.8	71.2	17.3	62.9	72.1	9.1	8.2
<i>advcl</i>	31.5	46.8	15.3	25.9	46.8	20.9	-5.6
<i>root</i>	56.4	71.6	15.2	62.8	74.0	11.2	4.0
<i>xcomp</i>	40.0	54.9	14.9	51.2	50.0	1.2	13.7
<i>discourse</i>	30.7	44.9	14.2	46.0	51.4	5.4	8.8

Table 4: Recall by relation type under UDPipe’s *Morpho-Tagger* and *ARK Tagger* settings (+synthetic+embeddings; (3) and (6) from Table 3; §5.3). *Reduction* is the reduction in performance gap from the *Morpho-Tagger* setting to the *ARK Tagger* setting; bolded numbers indicate a gap reduction of ≥ 10.0 .

Feature	AA Count	WH Count	Example
Dropped copula	44	0	<i>MY bestfriendddd mad at me tho</i>
Habitual <i>be</i> , describing repeated actions	10	0	<i>fees be looking upside my head likee ion know wat be goingg on . I know that clown, u don’t be around tho</i>
Dropped possessive marker	5	0	<i>ATTENTION on Tv...tawkn bout dat man gf Twink rude lol can’t be calling ppl ugly that’s somebody child lol...</i>
Dropped 3rd person singular	5	0	<i>When a female owe you sex you don’t even wanna have a conversation with her</i>
Future <i>gone</i>	4	0	<i>she gone dance without da bands lol</i>
<i>it is</i> instead of <i>there is</i>	2	1	<i>It was too much goin on in dat mofo .</i>
Completive <i>done</i>	1	0	<i>damnnn I done let alot of time pass by . .</i>

Table 5: Examples of AAE syntactic phenomena and occurrence counts in the 250 AA and 250 WH tweet sets.

ting to the *ARK Tagger* settings produced significant accuracy increases (Table 6). We observed that the greatest improvement came from using the *ARK Tagger* setting with Twitter-trained embeddings; the Twitter-specific lexical information provided by the embeddings was critical to recognizing the variants. Surprisingly, adding synthetic data decreased the model’s ability to parse the variants.

We next investigated the presence of AAE syntactic phenomena in our dataset. Table 5 shows examples of seven well-documented AAE morphological and syntactic features and counts of their occurrences in our AA and WH tweet sets; again, while several of the phenomena, such as dropped

copulas and habitual *be*, occur frequently in our AA tweets, there is only one instance of any of these features occurring in the WH tweet set.

We measured the parsing accuracy for the two most frequent syntactic features, dropped copulas and habitual *be*, across the four models; accuracies are given in Table 6. For dropped copulas, we measured parsing correctness by checking if the parser correctly attached the subject to the correct predicate word via the *nsubj* relation; for the first example in Table 5, for example, we considered the parser correct if it attached *bestfriendddd* to *mad* via the *nsubj* relation. For habitual *be*, we checked for correct attachment via the *aux* or *cop* relations as in the first and second examples in Ta-

AAE Feature	Morpho-Tagger Baseline	ARK Tagger Baseline	ARK Tagger with Embeddings	ARK Tagger with Synthetic, Embeddings
Lexical Variants (§3.1)	16.3 (13/80)	61.3 (49/80)	63.8 (51/80)	57.5 (46/80)
Dropped copula	54.5 (24/44)	70.5 (31/44)	61.4 (27/44)	68.2 (30/44)
Habitual <i>be</i>	50.0 (5/10)	80.0 (8/10)	90.0 (9/10)	90.0 (9/10)

Table 6: Parsing accuracies of syntactic and lexical variations across four UDPipe models (see §5.4).

ble 5, respectively.

As before, we observed significant increases in accuracy moving from the *Morpho-Tagger* to the *ARK Tagger* settings. However, neither adding embeddings nor synthetic data appeared to significantly increase accuracy for these features. From manual inspection, most of the dropped copulas errors appear to arise either from challenging questions (e.g. *ATTENTION what yo number ?*) or from mis-identification of the word to which to attach the subject (e.g. *He claim he in love llh*, where *he* was attached to *llh* rather than to *love*).

6 Conclusion

While current neural dependency parsers are highly accurate on MAE, our analyses suggest that AAE text presents considerable challenges due to lexical and syntactic features which diverge systematically from MAE. While the cross-domain strategies we presented can greatly increase accurate parsing of these features, narrowing the performance gap between AAE- and MAE-like tweets, much work remains to be done for accurate parsing of even linguistically well-documented features.

It remains an open question whether it is better to use a model with a smaller accuracy disparity (e.g. UDPipe), or a model with higher average accuracy, but a worse disparity (e.g. Deep Biaffine). The emerging literature on fairness in algorithms suggests interesting further challenges; for example, Kleinberg et al. (2017) and Corbett-Davies et al. (2017) argue that as various commonly applied notions of fairness are mutually incompatible, algorithm designers must grapple with such trade-offs. Regardless, the modeling decision should be made in light of the application of interest; for example, applications like opinion analysis and information retrieval may benefit from equal (and possibly weaker) performance between groups, so that concepts or opinions in-

ferred from groups of authors (e.g. AAE speakers) are not under-counted or under-represented in results returned to a user or analyst.

Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was supported by a Google Faculty Research Award, and a National Science Foundation Graduate Research Fellowship (No. 1451512).

References

- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *Proceedings of EMNLP*.
- Su Lin Blodgett and Brendan O’Connor. 2017. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv preprint arXiv:1707.00061*; presented at *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) workshop at KDD 2017*.
- David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. 2006. Parsing Arabic dialects. In *Proceedings of EACL*.
- Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the KDD*. ACM.
- Joachim Daiber and Rob van der Goot. 2016. The de-noised web treebank: Evaluating dependency parsing under noisy input conditions. In *Proceedings of LREC*.
- Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Recent Advances in Natural Language Processing*, pages 198–206.
- Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *Proceedings of ICLR*.

- Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford’s graph-based neural dependency parser at the CoNLL 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.
- Jacob Eisenstein. 2013. What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369.
- Jennifer Foster, Özlem Çetinoglu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan, and Josef Van Genabith. 2011. #hardtoparse: Pos tagging and parsing the twitterverse. In *AAAI 2011 Workshop on Analyzing Microtext*.
- Rob van der Goot and Gertjan van Noord. 2017. Parser adaptation for social media by integrating normalization. In *Proceedings of ACL*.
- Lisa Green, Toya A Wyatt, and Qiuana Lopez. 2007. Event arguments and ‘be’ in child African American English. *University of Pennsylvania Working Papers in Linguistics*, 13(2):8.
- Lisa J Green. 2002. *African American English: A linguistic introduction*. Cambridge University Press.
- Spence Green and Christopher D Manning. 2010. Better arabic parsing: Baselines, evaluations, and analysis. In *Proceedings of COLING. ACL*.
- Taylor Jones. 2015. Toward a description of African American Vernacular English dialect regions using “Black Twitter”. *American Speech*, 90(4).
- Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2016. Learning a pos tagger for aave-like language. In *Proceedings of NAACL*. Association for Computational Linguistics.
- David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of ACL*.
- Mohammad Khan, Markus Dickinson, and Sandra Kübler. 2013. Towards domain adaptation for parsing web data. In *RANLP*.
- Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. *Proceedings of ITCS*.
- Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archana Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.
- Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A Smith. 2018. Parsing tweets into universal dependencies. *Proceedings of NAACL*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Alexis Nasr, Geraldine Damnati, Aleksandra Guerraz, and Frederic Bechet. 2016. Syntactic parsing of chat language in contact center conversation corpus. In *Annual SIGdial Meeting on Discourse and Dialogue*.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
- Brendan O’Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory search and topic summarization for Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.
- Olutobi Owoputi, Brendan O’Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*.
- Manuela Sanguinetti, Cristina Bosco, Alessandro Mazzei, Alberto Lavelli, and Fabio Tamburini. 2017. Annotating Italian social media texts in universal dependencies. In *Proceedings of Depling 2017*.
- Yves Scherrer. 2011. Syntactic transformations for Swiss German dialects. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties. ACL*.
- Nathan Schneider, Brendan O’Connor, Naomi Saphra, David Bamman, Manaal Faruqui, Noah A. Smith, Chris Dyer, and Jason Baldridge. 2013. A framework for (under)specifying dependency syntax without overloading annotators. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 51–60, Sofia, Bulgaria. Association for Computational Linguistics.
- Ian Stewart. 2014. Now we stronger than ever: African-american english syntax in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37, Gothenburg, Sweden. Association for Computational Linguistics.
- Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipeline: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of LREC*.

- Hongmin Wang, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu. 2017. Universal dependencies parsing for colloquial Singaporean english. *Proceedings of ACL*.
- Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. [Universal decompositional semantics on universal dependencies](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.
- Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.
- Congle Zhang, Tyler Baldwin, Howard Ho, Benny Kimelfeld, and Yunyao Li. 2013. Adaptive parser-centric text normalization. In *Proceedings of ACL*.