A Test Suite for Evaluating POS Taggers across Varieties of English*

Anna Jørgensen
Institute for Logic, Language and Computation
University of Amsterdam
jorgensen@uva.nl

Anders Søgaard
Center for Language Technology
University of Copenhagen
soegaard@hum.ku.dk

ABSTRACT

We present a suite of 12 datasets for evaluating POS taggers across varieties of English to enable researchers to evaluate the robustness of their models. The suite includes three new datasets, sampled from lyrics from black American hip-hop artists, southeastern American Twitter, and the subtitles from the TV series *The Wire*. We present an example evaluation of an off-the-shelf POS tagger across these datasets.

Keywords

Datasets; Performance; Evaluation; POS tagging

1. INTRODUCTION

Most off-the-shelf POS taggers for English have been induced from and evaluated on manually annotated newspaper corpora such as the English Penn Treebank. This has led to community-wide overfitting to a very specific instantiation of newspaper English, and it is well-established that, unsurprisingly, these off-the-shelf taggers perform significantly worse on other domains and varieties of English. Many researchers have tried to obtain better performance across domains or varieties by using heavier model regularization or by learning from mixtures of labeled and unlabeled data. This also means that researchers have gone from evaluating their models on newspaper English, specifically the Wall Street Journal sections 22–24 in the English Penn Treebank, to using other datasets. Typically, however, researchers have focused on one or a few specific domains, again running the risk of over-fitting to these datasets [7]. In this paper, we present a suite of 12 datasets for evaluating POS taggers across varieties of English, including three new datasets of subtitles, tweets, and hip-hop lyrics. In the machine learning community, it is an important rule of thumb to evaluate new classification algorithms across at least a dozen datasets [1], and this is an attempt to make a dozen datasets available for English POS tagging, an important NLP task for many downstream applications. We also present an example evaluation of the Stanford POS tagger across all these datasets. The three new domains – hip-hop lyrics, southeastern American Twitter and subtitles from the TV series *The Wire* – differ with varying degrees syntactically and lexically from newswire as well as in the communicative functions of the domains. In our suite, we ignore datasets that have been used for decades, such as the Brown Corpus, the Switchboard Corpus, and the English Penn Treebank, to avoid community-wide over-fitting, as well as fitting our models to 20^{th} century varieties of English. The 12 datasets are presented in Table 1. The data is available from `https://bitbucket.org/lowlands/`, which also contains pointers to other data sets for English newswire and Twitter.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890559.

2. NEW DATASETS

In this paper, we present three new English datasets from domains containing non-canonical language use reflecting aspects of African American Vernacular English (AAVE) such as syntactic variation, lexical items, abbreviations and phonologically-motivated spellings.

The three domains were chosen because of the heavy presence of AAVE features as well as for the individual characteristics of the domains, which too vary from those of the commonly-used evaluation datasets in NLP. All three domains are low in formality and high in individualistic expression, which is a desired quality for analyzing vernacular language uses. Much annotated NLP data, such as legal documents and newswire are formal in writing style and subject to high demands of standardization because of topic and the public distribution of texts from these domains. Lyrics, subtitles and tweets, while public, are not subject to as strict demands of standardization. Lyrics and subtitles showcase a certain subculture in a narrative form and wish to establish certain personas through language use (and mise-en-scène), while tweets are examples of phatic communication.

All three datasets were annotated by a trained linguist with experience in African American Vernacular English. We used the Universal Google tag set [10] with 12 categories[1] to ensure higher usability for POS tagger evaluation. [7] has shown that with this tag set, annotation guidelines are not necessary to obtain high-quality POS annotations, and we reach an inter-annotator agreement of 93.6%.

[1]ADJ (adjectives), ADP (adpositions), ADV (adverbs), CONJ (conjunctions), DET (determiners and articles), NOUN (nouns), NUM (numerals), PRON (pronouns), PRT (particles), "." (punctuation marks), VERB (verbs) and X (miscellaneous).

Table 1: The datasets in the evaluation suite

Domain	Reference	Sentences	Stanford
Answers	English Web Treebank	1,744	91.5%
Emails	English Web Treebank	2,450	90.7%
Newsgroups	English Web Treebank	1,195	91.6%
Newspaper	Dundee Treebank	51,502	93.2%
Reviews	English Web Treebank	1,906	92.5%
Spoken (child-directed)	CHILDES	5,222	94.0%
Spoken (interviews)	LOWLANDS	500	85.0%
Twitter	LOWLANDS	500	87.0%
Weblogs	English Web Treebank	1,015	92.1%
Lyrics	This work	509	77.7%
Subtitles	This work	1,074	83.3%
Twitter (AAVE)	This work	374	61.4%

In the following, we present the three data sets with an explanation of the collection process and the preprocessing of the data. We further describe the individual characteristics of the datasets and compare these with the commonly-used evaluation sets in NLP. Statistics on the three new datasets are provided in Table **??** with example sentences and annotations from each dataset in Table 3.

2.1 Hip-hop Lyrics

The annotated hip-hop lyrics dataset consists of lyrics from black American hip-hop artists[2] produced between 1993–2012. We collected the hip-hop lyrics dataset using the *Rap Genius* depository.[3] *Rap Genius* is a database of lyrics from various musical genres, where contributors can annotate semantic meanings of the lyrics. These contributions were used to annotate entities and constructions in the hip-hop lyrics which were unknown to the annotator.

The hip-hop lyrics dataset contains many non-canonical lexical entities such as *tecks* (Teck 9, a type of firearm[4]) and *loc'd out* (loco, crazy), phonologically-motivated spellings *fo', pimpin'* and *betta* and domain-specific entities used for rhythm and rhyme such as *badonkadonk-donk*.

The most characteristic trait of the hip-hop lyrics dataset is the absence of punctuation marks. As can be see in Table 4, only 6% of the tokens in the hip-hop lyrics dataset are punctuation marks, and this figure is even inflated, since we split rhyming words on hyphens (such as *badonkadonk-donk*). Excluding commas and hyphens, the percentage of end punctuation marks is a mere 1.4% of the tokens in this dataset. This lack of sentence separating punctuation marks poses a challenge to a POS tagger trained on newswire, where a punctuation mark is typical at the end of every sentence. It also posed a problem for us in determining where one sentence ends and the next begins.

The poetic nature of the domain plays on the ambiguity created by sentences flowing into each other, but for simplicity and to eliminate the fear of wrongly interpreting the intended meaning of the lyrics, we declare each line in the lyrics a sentence. Not only end punctuation is important for correct syntactic analysis, however. Consider the change in syntactic function of the word *this* (from determiner to pronoun) introduced by commas encircling *bitch* in the following sentence, again from Lil Wayne's "Gangstas and Pimps", *I'm in this bitch hold your pictures* (500 Degreez, 2002).

2.2 Subtitles from The Wire

The television series *The Wire*[5] is a fictional narrative portraying various criminal institutions in Baltimore, MD as well as the local law enforcement departments.

The dataset presented here is the full first episode of the first season of the show, which is concerned with the Baltimore drug scene. The dataset consists of dialogues between the characters on the show, and features both white and black police officers, prostitutes, politicians, gangsters, drug dealers, judges and junkies. While the vast majority of the characters are male, there is a handful of women in the episode and they too represent a variety of sociolinguistic categories.

Non-canonical language can be used as a resource in cultural products for creating personas and exhibiting social and cultural characteristics and affiliations. [14], who perform a linguistic analysis of the use of AAVE in *The Wire*, note that much of the dialog in *The Wire* is marked by AAVE features such as copula deletion, habitual *be*, completive *done* and continuative *steady*. They further remark that characters on the show can be seen as representatives of "genuine AAVE" speakers [14, p. 38].

We collected the subtitles using *opensubtitles.org*,[6] where subtitles are available for various television series and movies in a variety of languages. We manually controlled that the subtitles collected indeed reflected the dialogues in the episode. The data format used by *opensubtitles.org* contains time stamps for each utterance, which means that sentences can be split over several time stamps, especially if the utterance is long. Since we are interested in sentences rather than utterances, we join utterances together across time stamps, until end punctuation marks the end of the utterance. We split utterances that contain several sentences into separate sentences and tokenize these, separating all punctuation from words except apostrophes not used as discourse markers (e.g., *'em*).

[2] 2Pac, 50Cent, Birdman, COOLIO and the Gang, Lil Wayne, Missy Elliott and Precious Paris.

[3] `http://rap.genius.com`

[4] `http://www.urbandictionary.com/define.php?term=teck`

[5] HBO, 2002–2008. `http://www.hbo.com/the-wire`

[6] `http://www.opensubtitles.org/en/search`

The spoken origin of subtitles means that the data contains quintessential elements of spoken language such as interjections, cut-off sentences, contractions and exclamations. However, because *The Wire* is scripted, not all aspects of natural speech are seen in this data set. There are for instance no hesitations, corrections or false starts in our dataset. The form of dialogue introduces a higher quantity of certain syntactic structures than present in the commonly-used CMU Twitter corpus [9]. The *The Wire* dataset presented here contains, as example, ∼28% questions, where the CMU Twitter corpus only contains 11.4%.[7] Comparatively, the AAVE tweets dataset contains 8.5% questions while only ∼2% of the sentences in the hip-hop lyrics dataset are questions. The AAVE tweets dataset is described below. This skew in the distribution of syntactic forms is evidence of the necessity of using training data with a high degree of variation as well as evaluating on data from various domains.

We believe that subtitles can be a rich source of data for various NLP tasks because of the high degree of linguistic variation and syntactic structures in dialogues while being in a readily manageable format.

Table 2: Statistics on the new Ddatasets

Domain	Lyrics	Subtitles	Tweets
Sentences	509	1074	374
Sent. length	9.1	8.9	9.7
Types	1314	1519	1606
OOV types	9.8%	6.2%	11.0%
OOV tokens	12.1%	10.4%	22.2%

2.3 AAVE Tweets

Several part-of-speech annotated datasets from Twitter are currently available, but to the best of our knowledge, this is the first that has been geographically restricted to increase the degree of AAVE features. The tweets for the AAVE tweets dataset were collected through the Twitter API using the Python package, TwitterSearch,[8] between 2015-02-28–2015-03-03. The geo-coordinates provided in the metadata were used to exclude all tweets not posted within the American Gulf Coast states.[9] We chose to only include tweets from these states because of the higher percentage of black Americans in the population of this region,[10] and because it has been show that AAVE features are more prevalent in tweets from this area than from elsewhere in the United States [8]. Since we did not want to limit the data to urban language use, we chose states rather than cities as our search frame. This also enables comparative linguistic analyses of urban and rural uses of AAVE in this dataset. Five linguistic challenges of non-standard variation are endemic to online social media data, namely punctuation, capitalisation, spelling, vocabulary and syntactic structures [3]. These variations are also present in our AAVE tweets dataset along with emoticons, which can be seen in the higher frequency of the miscellaneous category, *X*, in this dataset as shown in Table 4.

Several studies have shown that besides syntactic structures and lexical entities, phonological variation is also present in tweets [2, 4, 5, 8].

[8], who focus on phonologically-motivated spellings of AAVE on Twitter, show that AAVE features such as interdental fricative mutation, derhotacization and backing in [str][11] is present on Twitter and that these features partially correlate with demographic information about the location of the tweeter.

2.4 Dataset characteristics

There are notable differences in what tags are likely to follow each other across the three new tdatasets. In the AAVE tweets dataset, for example, an instance of the miscellaneous class, X, is most often followed by another instance of the same category, whereas in the other two datasets, a word belonging to the miscellaneous class is most often followed by a punctuation mark. On Twitter, several punctuation marks can be used after each other, while this is less common in the other domains. These differences in syntactic constructions in the four domains illustrate the necessity of using multiple and varied data sets for evaluation of a POS tagger.

Table 2.2 shows the sizes of the new datasets, the number of word types and average tokens per sentence as well as the percentage of out-of-vocabulary (OOV) types and tokens in the three new datasets compared to the CMU Twitter corpus.

The highest percentage of both OOV types and tokens in both categories is not surprisingly in the AAVE tweets dataset, followed by the hip-hop lyrics dataset on both accounts, while the lowest in both categories is in the subtitles dataset.

While the hip-hop lyrics dataset has almost as high a percentage of OOV types as the AAVE tweets dataset, they are used less frequently. In the AAVE tweets dataset almost every 4th token is an out-of-vocabulary word.

Twitter is known for its short messages, but of the new datasets presented here, the AAVE tweets have the longest average sentence length. While this is interesting given Twitter's 140-characters restriction and the issues with developing NLP tools for Twitter [6, 3], it is not surprising. Hip-hop lyrics are essentially rapped poems to a beat, and the lyricists therefore also have strict restrictions on length as well as on rhythm and rhyme[12], and subtitles represent conversation and contains cut-off sentences, exclamations and short remarks[13].

The distributions of part-of-speech tags in the CMU Twitter corpus and the three new datasets are presented in Table 4. Two observations related to the distribution of the 12 POS tags are worth briefly mentioning here, as these ob-

[7]These figures are measured by counting sentences ending in a question mark, and they are therefore not the result of a discourse analysis of the data.

[8]`https://pypi.python.org/pypi/TwitterSearch/`

[9]North Carolina, South Carolina, Louisiana, Georgia, Mississippi, Tennessee, Arkansas, Alabama and Florida.

[10]U.S. Census Bureau, Census 2000 Redistricting Data (Public Law 94-171) `https://www.census.gov/prod/cen2010/briefs/c2010br-06.pdf`,`http://www.indexmundi.com/facts/united-states/quick-facts/all-states/black-population-percentage#chart`

[11]Uniquely to AAVE, word-initial [str] can be substituted by [skr] in words such as "street", "strip" and "strawberry"[8, 11, 12, 13]

[12]e.g., *Cadillac smoke* ***dro*** *just me and the* ***ho***. Lil Wayne feat. Birdman. "Gangstas and Pimps". *500 Degreez*, 2002

[13]e.g., *Yo.*. Example is taken from the subtitles dataset.

Table 3: Example sentences from new datasets

Dataset	Annotated sentences
Lyrics	2Pac/NOUN cares/VERB ,/. if/CONJ don't/VERB nobody/PRON else/ADJ care/VERB I'm/PRON a/DET loc'd/ADJ out/PRT gangsta/NOUN set/NOUN trippin'/VERB banger/NOUN
Subtitles	Life/NOUN just/ADV be/VERB that/DET way/NOUN ,/. I/PRON guess/VERB ./. Couldn't/VERB help/VERB hisself/PRON ./.
Tweets (AAVE)	Dat/DET highlight/NOUN and/CONJ contour/NOUN doe/ADV URL/NOUN Aww/X damn/X I/PRON can't/VERB believe/VERB dat/PRON lol/X

servations points to deviations in the new datasets from the CMU Twitter corpus.

Firstly, it is clear that the amounts of punctuation marks and of the miscellaneous, *X* category vary from domain to domain. About 10% of both the CMU Twitter corpus and our AAVE tweets dataset are punctuation marks, while every 5^{th} token in the subtitles dataset and only 6% of the tokens in the hip-hop lyrics dataset are punctuation marks.

Secondly, the CMU Twitter corpus contains fewer determiners, pronouns and verbs than the three new datasets. The difference in the frequencies of nouns and determiners is smaller in the three new test sets(∼10%) than in the CMU Twitter corpus(∼14%). The highest distributions of verbs and pronouns are found in the hip-hop lyrics dataset, which is possibly due to the concise, poetic writing style, while the CMU Twitter corpus set contains a much smaller amount of pronouns than all the other data sets.

Table 4: Tag distribution per data set

		New datasets		
POS	**CMU**	Subtitles	Lyrics	Tweets (AAVE)
.	12%	21%	6%	10%
ADJ	5%	4%	5%	5%
ADP	9%	8%	9%	7%
ADV	5%	5%	4%	5%
CONJ	2%	2%	5%	2%
DET	6%	9%	11%	10%
NOUN	20%	18%	20%	22%
NUM	1%	1%	1%	1%
PRON	7%	14%	15%	12%
PRT	5%	2%	2%	1%
VERB	15%	17%	21%	17%
X	13%	1%	1%	8%
Total tokens	34.3k	4.2k	4.5k	5k

In closing, we believe the variation present in these three new datasets is relevant for testing the robustness of POS tagging systems, and we encourage researchers to include these datasets in the evaluation of their systems.

3. EXAMPLE EVALUATION

We present a very simple evaluation of the Stanford POS tagger in the right-most column in Table 1. We observe that results on the three new datasets are significantly lower than for any of the other datasets. This, in our view, demonstrates the need for evaluation data representing minority varieties of English such as African American Vernacular English.

4. ACKNOWLEDGMENTS

This research is funded by the ERC Starting Grant LOWLANDS No. 313695.

5. REFERENCES

[1] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[2] J. Eisenstein. Phonological factors in social media writing. In *Proceedings of NAACL Workshop on Language Analysis in Social Media*, 2013.

[3] J. Eisenstein. What to do about bad language on the Internet. In *Proceedings of NAACL-HLT*, 2013.

[4] J. Eisenstein. Systematic patterning of phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188, April 2015.

[5] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP*, 2010.

[6] J. Foster, Ö. Çetinoğlu, J. Wagner, J. L. Roux, J. Nivre, D. Hogan, and J. van Genabith. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proceedings of IJCNLP*, 2013.

[7] D. Hovy, B. Plank, and A. Søgaard. When POS datasets don't add up: Combatting sample bias. In *LREC*, 2014.

[8] A. Jørgensen, D. Hovy, and A. Søgaard. Challenges of studying and processing dialects in social media. In *Proceedings of the ACL Workshop on Noisy User-generated Text*, 2015.

[9] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, 2013.

[10] S. Petrov, D. Das, and R. McDonald. A universal part-of-speech tagset. In *Proceedings of LREC*, 2011.

[11] K. Pollock, Bailey, Berni, J. R. Fletcher, Hinton, and Weaver. Phonological features of African American Vernacular English (AAVE), 2001.

[12] J. Rickford. *African American Vernacular English.* Blackwell, Oxford and Malden, MA, 1999.

[13] E. R. Thomas. Phonological and phonetic characteristics of African American Vernacular English. *Language and Linguistics Compass*, 1(5):450–475, 2007.

[14] J. Trotta and O. Blyahher. *Game done changed*: A look at selected AAVE features in the TV series *The Wire. Moderna Språk*, 2011.