Illegal is not a noun: Linguistic form for detection of pejorative nominalizations

Alexis Palmer, Melissa Robinson, Kristy Phillips
Department of Linguistics
University of North Texas
Denton, Texas, 76203, USA
{alexis.palmer,kristy.phillips}@unt.edu, melissa.robinson@my.unt.edu

Abstract

This paper focuses on a particular type of abusive language, targeting expressions in which typically neutral adjectives take on pejorative meaning when used as nouns - compare *gay people* to *the gays*. We first collect and analyze a corpus of hand-curated, expert-annotated pejorative nominalizations for four target adjectives: *female*, *gay*, *illegal*, and *poor*. We then collect a second corpus of automatically-extracted and POS-tagged, crowd-annotated tweets. For both corpora, we find support for the hypothesis that some adjectives, when nominalized, take on negative meaning. The targeted constructions are non-standard yet widely-used, and part-of-speech taggers mistag some nominal forms as adjectives. We implement a tool called NomCatcher to correct these mistaggings, and find that the same tool is effective for identifying new adjectives subject to transformation via nominalization into abusive language.

1 Introduction

Detection of abusive language tends to focus on identification of key words and character sequences that indicate expression of strongly negative attitudes toward individuals or groups of people (for example, Warner and Hirschberg, 2012; Waseem and Hovy, 2016; Nobata et al., 2016). Some key words, such as racial or ethnic slurs, are highly effective predictors, while other key words may signal contentious topics rather than actual abusive language. This second type of key word is semantically flexible. Depending on the context of individual occurrences, these words may be used abusively, neutrally, or even to express positive sentiment.

In this paper we focus on pejorative uses (i.e., uses expressing contempt, disapproval, or other negative sentiment) of words that are alternately neutral or pejorative, depending on their syntactic context. Specifically, we are interested in negatively-characterizing phrases such as *the blacks* or *the gays*. Formally, these expressions involve nominalization of adjectives, where one particular characteristic (e.g. homosexuality) becomes associated with a wide range of stereotypical notions (Wierzbiecka, 1986). Though these constructions are nothing new - the online Corpus of Historical American English,[1] for example, has one occurrence of *the blacks* as early as 1810 - they came to new public prominence during the 2016 U.S. Presidential election (Rebels, 2016; Liberman, 2016b,a).

A phrase like *the Mexicans* may not immediately register as pejorative, but the associated negative sentiment (1) becomes clear through contrast with a different type of noun phrase (2):

1. *I think **the Mexicans** are going to end up loving Donald Trump.* [cited in Liberman (2016b)]

2. *I think **the Mexican people** are going to end up loving Donald Trump.* [constructed]

In (2), *Mexican* is an adjective modifying the noun *people*; in (1), *Mexican* has been nominalized.[2]

This paper presents work in progress exploring the utility of linguistic form (i.e. particular syntactic constructions, discussed in Section 2) for automatically identifying this more subtle form of abusive language. We start by investigating a

[1] http://corpus.byu.edu/coha/

[2] The analysis of the form in (1) as nominal is supported by its compatibility with the nominal plural inflection *-s*.

hand-collected, expert-annotated corpus of nominal uses of *female*, *gay*, *illegal*, and *poor* (Section 3). For this data set and the four adjectives it targets, analysis shows a strong correspondence between nominal status and pejorative meaning.

Because our ultimate interest is in automatic detection of abusive language in unrestricted online data, we assemble a second corpus via automatic data extraction, use automatic part-of-speech (POS) labels as a proxy for linguistic form, and turn to the crowd for annotation (Section 4). This study again shows correspondence between negative sentiment and linguistic form, although the results are complicated by annotation issues.

Finally, we present two short investigations into the feasibility of the current approach for automatic detection of abusive language (Section 5). One interesting result shows that output from an automatic POS tagger can be used to identify new pejorative nominalizations in unrestricted data.

NOTE: This paper contains a number of examples of abusive and/or offensive language. These do not represent the views of the authors! Please proceed with caution and awareness.

2 Pejorative meaning and linguistic form

Locating pejorative meaning. Disentangling the pejorative load of an individual lexical item from the sentiment of the utterance in which it occurs is difficult, sometimes even impossible. The ultimate aim of the research agenda this paper contributes to is to mark *individual occurrences* of certain lexical items as pejorative or not. Sometimes the nominalizations of interest occur embedded in a clearly abusive context, as in example (6) below. In other instances, though, the context itself is relatively neutral, and use of the nominalization is precisely what shifts the utterance from neutral to pejorative (as in example (1) above).

The two corpora discussed in this paper differ in the care with which they distinguish between: a) pejorative meaning of a lexical item, and b) negative sentiment of an utterance. The PEJNOM corpus (Section 3) was annotated by one expert linguist. This annotator paid close attention to the location issue, developing guidelines for when to attribute pejorative meaning to a lexical item and when not to. Making this distinction requires closely examining the semantic contributions both of the targeted lexical item and of its context. The two must be separately interpreted.

The TWTARGETS corpus (Section 4), on the other hand, was annotated using crowd-sourcing, with simple instructions given to anonymous, amateur annotators. The annotations suggest that crowd annotators do not always make the distinction as carefully as we would like.

Relationship with sentiment analysis. There is a clear connection between this work and sentiment analysis, given that pejorative meaning is *by definition* the expression of negative sentiment. However, most methods for sentiment analysis target the level of the utterance or the entire document. Our analysis focuses in on the level of the individual lexical item, as we aim to automatically classify occurrences of particular target words as pejorative or non-pejorative.

Pejoration as a process. From a theoretical perspective, pejoration is a process by which lexical items acquire negative meaning. In the case of adjectival nominalization (i.e. for our target forms), pejoration occurs as certain adjectival forms begin to be used as nouns.

In our proposed process of ADJ→N pejoration, the first step is from adjective (e.g. *My **rich** aunt paid for my schooling*) to the zero plural form (e.g. ***The rich** should pay more taxes than the poor*).[3] The zero plural may be seen as an intermediate step between adjectival and nominal forms (Günther, to appear). So far these are standard forms, with no inherent pejorative meaning.

Pejoration happens when the word crosses the boundary from zero plural to true nominal forms. As Wierzbicka states, nouns (typically) refer to individuals or groups of individuals, and adjectives (generally) ascribe characteristics to individuals. In this nominalization, a kind or category of entity is formed around the (former) adjective. In addition to the single attribute denoted by the adjective, stereotypical properties become associated with the kind, such as *dumb* and *sexy* for the nominalized *blonde*.

Using Wierzbicka's theory, we take a step further in our analysis, arguing that a certain dehumanization or deindividualization can come with nominalization, as individuals are referred to not as complex human beings but by making reference to *a single characteristic* of the individual. Addi-

[3] This form is known as *zero plural* because it denotes plural reference without plural inflection on the noun.

tionally, the properties associated with the nominal forms often lack the human properties associated with more standard variants. Consider the semantic properties of *woman* and *female*.

- Woman: FEMALE, HUMAN, ADULT
- Female: BIOLOGICAL SEX

HUMAN is one of the properties of the word *woman*, but this is not the case for *female*.

Once the adjective has been nominalized, it can occur in different forms. In English, nominal forms vary with respect to definiteness and number (see Section 3.2 for examples). Some forms are more marked than others, and non-standard, bare plural uses like those in (3) are widely found in online environments.

3. *Our system is free and accessible to every citizen, **richs** and **poors**. #debate #presidentialdebate* [Twitter, 2016]

3 Corpus Study 1: Hand-curated data, annotated by an expert

Four data sets are used across the two studies; the number of instances in each appears in Table 1. For the PEJNOM corpus, one **instance** is one occurrence of a target adjective, within an utterance of 1-3 sentences. One utterance can contain more than one instance. For TWTARGETS and TWOPEN, one **instance** is one tweet.

Our first corpus study addresses a manually-collected data set. The data set was curated over a number of months by a graduate student in Linguistics with a theoretical interest in understanding why there is such a striking contrast between adjectival and (some) nominal uses of four adjectives: *female*, *gay*, *illegal*, and *poor*. The initial focus of this data set was to assemble a large collection of pejorative nominalizations, as an empirical foundation for linguistic analysis.

The original version of the corpus (PEJNOM-ORIG) focuses on identifying pejorative nominalizations, resulting in a thoroughly unbalanced data set. To expand the data set without collecting additional data, we annotate *all* occurrences of the four target forms in the corpus, not only those which triggered inclusion of instances in the corpus in the first place. This second annotation round added 444 instances to the corpus; the expanded version is named PEJNOM-EXP.

3.1 The corpus and the target forms

The PEJNOM-ORIG corpus was assembled from Twitter, Reddit, news articles and interviews, political debates, and video and written blogs. The majority of the data is written, though some spoken data was transcribed and included.

Each of the four target adjectives is most likely to occur in its negative/abusive form in particular environments related to the term. In order to find pejorative uses, selected topics revolving around immigration, anti-feminism, homophobia, and poverty were searched.

Illegal. Data for *illegal* was primarily collected during the 2016 U.S. Presidential elections, with examples harvested from politicians during debates and interviews as well as online commentary from voters on political issues. Common topics were deportation, illegal immigration, and Donald Trump's border wall (e.g. 4).

4. *And those liberal SJWs don't want the wall.... And want to keep **illegals** in the US... Lmfao* [Reddit, June, 2016]

Female. Relevant forms of *female* are commonly found in Mens Rights blogs, specifically items tagged with MGTOW (Men Going Their Own Way). The Mens Rights movement is a collection of online groups that claim to exist to promote rights needed by men. However, within the MGTOW community, it is common for the discussion to focus on anti-feminist topics (e.g. 5). Other blogs with anti-feminist topics were also inspected for pejorative uses of *female*.

5. *As a gay shaman who has been victimized by a succession of narcissist **females**, MGTOW is giving me hope that the human race can survive the female psychopath.* [Youtube, 2015]

Gay. While most of the data for *gay* was collected from Twitter, anti-gay blogs and forums were inspected to find pejorative uses of *gay* (e.g. 6). The topics often center around gay marriage, gay rights, or hate crimes.

6. ***Gays** cannot reproduce, **gays** are not beneficial for humans in anyway and your love for them is without merit or reason.* [Reddit, 2014]

Poor. Pejorative examples of nominalized *poor* were found largely in satirical news articles focused on social topics, such as limits on welfare.

Data set	*# female*	*# gay*	*# illegal*	*# poor*	**All**
PejNom-Orig	715	149	564	241	1669
PejNom-Exp	1108	160	592	253	2113
TwTargets	200	200	200	200	800
TwOpen					56237

Table 1: Per data set, instances per target form.

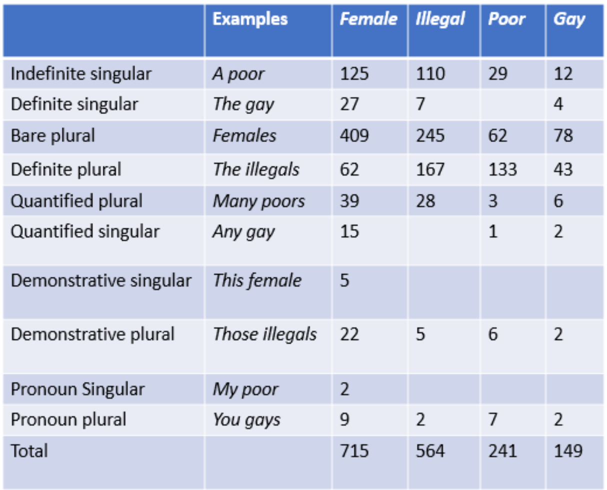
Additional examples were found on Twitter. The pejorative use of *poor* varies from the other target forms, as it is mostly used to voice a perceived attitude of another person or group, as in (7).

7. *"Hoover was in charge of the Great Depression, I only used words to say* ***poors*** *were dumb for paying taxes and staying poor." - Trump logic* [Twitter, Oct. 16, 2016]

Each instance in PejNom-Exp was annotated for two categories: **linguistic form** (3.2) and **pejorative meaning** (3.3).

3.2 A closer look at linguistic form

Each instance in the corpus is coded for its grammatical structure. The four main nominal forms are indefinite singular (*a gay*), definite singular (*the female*), bare plural (*poors*), and definite plural (*the illegals*). In order to make more fine-grained distinctions, additional categories were added, including demonstratives, quantifiers, and pronouns. Figure 1 shows the distribution of target forms across linguistic form categories, for PejNom-Orig.



	Examples	*Female*	*Illegal*	*Poor*	*Gay*
Indefinite singular	*A poor*	125	110	29	12
Definite singular	*The gay*	27	7		4
Bare plural	*Females*	409	245	62	78
Definite plural	*The illegals*	62	167	133	43
Quantified plural	*Many poors*	39	28	3	6
Quantified singular	*Any gay*	15		1	2
Demonstrative singular	*This female*	5			
Demonstrative plural	*Those illegals*	22	5	6	2
Pronoun Singular	*My poor*	2			
Pronoun plural	*You gays*	9	2	7	2
Total		715	564	241	149

Figure 1: Linguistic forms in PejNom-Orig.

The definite plural form is of particular interest. Acton (2014) argues that the definite plural structure can indicate the speaker's nonmembership in the group mentioned, as well as distancing the speaker from the group mentioned. In this case, the definite plural is a marked variant of the bare plural form. With this in mind, definite plurals are also coded when modified by a relative clause, as the relative clause may provide syntactic reasons for using the definite plural (e.g. 8).

8. *Do* ***the #illegals who were given greencards supposedly by accident*** *factor into #HRC vetting #debates #Trumptrain* [Twitter, Oct. 2016]

The manual collection process used search terms on raw text. In order to locate definite and indefinite singular forms, while ruling out adjectival forms, we added selected verb forms (e.g. forms of copular *be*) to the search terms, targeting token sequences like *a poor is* (e.g. 9).

9. *yeah dude being poor happens from time to time, but being* ***A poor is*** *a way of life. LOL. :)* [Twitter, Jul. 13, 2016]

Utterances in which the target form is used in reference to itself (e.g. 10) are coded separately.

10. *sorry, but calling someone* ***an illegal*** *isn't racist! Illegal isn't a race* [Twitter, Jun 26, 2016]

Likewise, if the referent of the target form is nonhuman, such as *illegal* used to refer to illegal fireworks (11), or different from the expected referent, such as *illegal* for underaged drinkers (12), the instance is coded separately.

11. ***An illegal*** *went off on the ground and the sparks flew EVERYWHERE and one of them hit my forehead LOOOOOL* [Twitter, Jul. 4, 2015]

12. *Dunno if this is still true, but used to be* ***an ILLEGAL*** *wasn't considered a man unless he could finish 18 pack and drive home.* [Twitter, May 17, 2014]

Finally, instances with questionable spelling or other irregularities leading to ambiguity (e.g. 13) are excluded from the corpus.

13. *They are if* ***a poorz*** *has one or both.* [Wonkette.com, Sept. 2016]

3.3 Annotating pejorative meaning

Each instance is annotated for the presence of pejorative meaning, using four different labels: pejorative (PEJ), non-pejorative (NONP), uncertain (UNC), and satirical (SAT). The goal of this annotation is to capture **whether pejorative meaning is intended on the part of the speaker**.

What counts as pejorative? Through the course of annotation, the expert annotator refined her annotation guidelines, aiming to clarify precisely which factors trigger an annotation of PEJ. Some factors are consistent over all four target forms, while others are specific to one target form. The following factors signify pejorative uses of target forms; most are illustrated by examples:

- (14) negative adjective(s) modifying the target nominal form;
- (15) co-occurrence with phrases referring to particular stereotypes or behaviors associated with the relevant referent group (e.g. *freeloading* with an occurrence of *poor*);
- (16) appearance near negative verbs such as *hate* or *despise*, or negative phrases such as *get rid of* or *hardly any good*;
- coreference with other negative terms, such as *slut* for *female* or *wetback* for *illegal* was an indication for pejorative meaning as well;
- (17) other negative implications not tied to a specific lexical item or phrase.

14. *"You have the distinct odor of poverty. Trust me, I can smell you from here!* ***Sad filthy poors****." - Trump in PA* [Twitter, Oct. 10, 2016]

15. *Why don't* ***gays*** *like being girly? Cause* ***a gay*** *is normally called girly.* [Twitter, Aug. 13, 2016]

16. *Whites* ***hate illegals****. Blacks* ***hate illegals****. Native Americans* ***hate illegals****. Asians* ***hate illegals****. legals* ***hate illegals****.* [Reddit, May 2016]

17. *this feminist nonsense is to give every man the daily message that* ***A Man Needs a Female Like a Fish Needs a Lobotomy****.* [Youtube, 2016]

Some target forms have specific indicators of pejorative/non-pejorative meanings. For example, if *female* occurs while discussing gender-focused topics (e.g. 18) or in pro-feminist contexts, it tends to be non-pejorative.

18. *Estrogen makes* ***females*** *more emotionally driven on average compared to males.* [Youtube, 2016]

Characteristically, the pejorative form of *female* is often paired with somehow mis-matched gendered nouns: such as *man* rather than *male*. The "matched" counterpart of *female* is *male*; *man*'s counterpart should be *woman*. When *female* is used in direct contrast to *man*, the semantic mismatch signals pejorative meaning (19).

19. *The president of the United States, to me, should be* ***a man*** *not* ***a female****.* [CNN interview, 2015]

Non-pejorative instances. We extend the corpus by annotating *all* occurrences of the four target words. Most adjectival occurrences (e.g. 20) and zero plural forms are annotated as NONP.

20. *Most of the arguments that I see against* ***gay marriage*** *invoke religious texts or figures.* [Reddit, 2015]

Satire/sarcasm. The satirical category (SAT) codes a different type of pejorative use. This category includes sarcastic uses and uses that voice the perceived attitude of a person, group, or society other than the speaker (see 21, for example). This tag is still considered to be pejorative, but is coded separately as it functions differently from blatant, explicitly negative uses. The SAT label occurs most frequently for *poor*, but does occur with other forms as well. Warner and Hirschberg (2012) also recognize sarcastic/satirical uses as a distinct category of abusive language.

21. *How dare* ***the poors*** *eat a steak! It offends my upper middle class sensibilities! Or something.* [Twitter, Oct. 20, 2016]

Uncertain. Lastly, the uncertain category (UNC) exists to capture instances for which the expert annotator did not feel confident choosing either PEJ or NONP. Often this is due to a limited amount of context, an unclear implication or sentence, or negative elements within questions, making it unclear whether pejorative meaning was intended on the part of the speaker.

Expert	#	Adj	ZP	Nom	Other	NoLF
ALL	2113	410	6	1649	41	7
%PEJ	1113	0.5	16.65	**66.8**	4.9	-
%NONP	564	**99.0**	16.65	9.3	9.8	-
%SAT	217	0.25	-	13.0	2.4	-
%UNC	181	0.25	-	10.9	-	-
%UNK	9	-	-	-	21.9	-
%NOAN	29	-	66.7	-	61.0	100.0

Table 2: Correspondence between pej. meaning and linguistic form, expert anno., PEJNOM-EXP.

22. *At work trying to explain how this man I know have **a gay** is so hard to explain especially without a good picture* [Twitter, Sept. 26, 2016]

3.4 Analysis: correspondence between pejorative meaning and linguistic form

Table 2 shows the correspondence between linguistic form (LF) and pejorative status for PEJNOM-EXP, taking only the annotations from the expert. For each LF category, the table shows the percentage of instances assigned to each of the four pejoration labels.

For this analysis, the fine-grained LF categories are collapsed into four categories. Adjectives and zero plurals, as expected, are overwhelmingly annotated as NONP, with all but 4 of the 410 adjectival occurrences of the four target forms. This is unsurprising, given the collection methodology used for the corpus, yet it confirms the expectation that these words are absent pejorative meaning when used as adjectives.

Of 1649 nominal occurrences across the four target forms, nearly 67% are annotated as PEJ, with the remaining instances spread across NONP (n=153), SAT (n=214), and UNC (n=180). An example of a non-pejorative nominal use is (23).

23. *It should not be understood as gay marriage (ie marriage for **gays**) but marriage that includes **gays** (ie the marriage is the same for all and is extended to **gays**), which is different.* [Reddit, 2015]

The category **Other** consists of those cases excluded from the main corpus (meta-references, non-human referents, etc.). Finally, a small number of cases in the corpus have no label either for LF or for pejorative meaning. These appear in the table as **NoLF**, UNK, and NOAN.

3.5 Analysis: multiple expert annotators

The PEJNOM-EXP corpus was annotated in its entirety by a single expert (**Annotator A**). To determine how replicable these annotations are, we recruited two additional expert annotators (**Annotators B1** and **B2**). All three are graduate students of linguistics. Neither B1 nor B2 had participated in this project before annotating.

Annotators B1 and B2 were given written annotation guidelines and asked to label (as PEJ or NONP) a subset of 121 instances, almost equally balanced across the four target forms. We call this data set PEJNOM-SUBSET.

Anno1 / **Anno2**	**A**		**B1**	
	%	K	%	K
B1	86.0%	0.717	–	
B2	71.9%	0.461	74.4%	0.499

Table 3: Agreement (% and Cohen's K) between expert annotators, PEJNOM-SUBSET.

Table 3 shows agreement figures for each pair of annotators, measured in both simple percent agreement and Cohen's Kappa.[4] We see that agreement between Annotator A and Annotator B1 is quite good, with K=0.717. Annotator B2 shows lower agreement with both of the other annotators, with Kappa scores of 0.461 and 0.499. Agreement across the three annotators (measured as Fleiss's Kappa) is a similarly modest 0.546.

Annotator	# PEJ	# NONP
A	54	67
B1	57	64
B2	84	37

Table 4: Ratings from multiple expert annotators (A=primary expert, B1&B2=additional experts), PEJNOM-SUBSET.

To better understand the differences between annotators, we look at the distributions of the two labels for each annotator (Table 4). It is clear that Annotator B2 is much more likely than the other two annotators to label instances as PEJ. This annotator seems to label based on the entire instance and not just the target form. We will see this behavior again in the crowd-sourced annotations described in Section 4.3.

[4] Agreement computed in R using the `irr` package.

4 Corpus Study 2: Data harvested online and annotated by the crowd

The first corpus study confirms the hypothesis that these four adjectives, when nominalized, take on pejorative meaning. This result, though, comes with a giant caveat: the corpus was collected precisely to investigate pejorative nominalizations. To test this hypothesis in a less-biased setting, we build a second corpus of instances extracted automatically from Twitter using twarc.[5] To move closer to automatic detection of abusive language, LF is assigned by an automatic part-of-speech tagger, and annotation is done via crowd-sourcing.

4.1 The corpus

This corpus has two subsets: TwTargets and TwOpen. Both subcorpora were de-duplicated using twarc's built-in utilities.

TwTargets. The first subcorpus consists of tweets which contain at least one of the four target forms discussed in Section 3. Using twarc, we searched for tweets containing either the singular or plural form of the target forms.[6] The full TwTargets data set consists of the most recent 6000 tweets for each of the four target forms.

TwOpen. The second subcorpus consists of 100,000 English-language tweets with geocodes located within a 2000 mile radius of the geographic center of the United States.[7]

The larger data set is next pruned by length, keeping only tweets with more than six words. The six-word limit does not include usernames, URLs, hashtags, emoticons, cardinal numbers, or punctuation. The remaining roughly 56K tweets make up the TwOpen data set.

4.2 Approximating LF with POS tagging

The previous analysis suggests that, given good annotations, LF could serve as a reasonable baseline for identifying pejorative uses of certain adjectives. In an application setting, though, it is unreasonable to expect human-quality labeling of LF, so we turn to automatic POS taggers.

The particular set of constructions poses a challenge for automatic POS taggers, because these are lexical items occurring with a syntactic category (N) that is *not* the most likely category.

Tagger selection. Before selecting a tagger, we investigated several different options, running all taggers with default settings: the standard English POS tagging model from Stanford CoreNLP (Toutanova et al., 2003); the GATE Twitter POS tagger (Derczynski et al., 2013);[8] and TweetNLP (Owoputi et al., 2013).[9] For a small test suite (57 instances), TweetNLP with its native tag set gave the best results for the four target words, looking at both adjectival and nominal uses. The TweetNLP tag set is a coarse-grained tag set extended with Twitter-specific tags for elements like hashtags and URLs. Of interest for our task are the tags N for nouns and A for adjectives.

NomCatcher: tag correction for nominalizations. A number of the target nominalizations are wrongly labeled as A, in particular definite and indefinite singular instances. Plural instances are largely labeled correctly as N.

In order to perform analysis of whether nominalized adjectives are likely to be pejorative, it's essential that the nominalizations are tagged correctly. To this end we implement **NomCatcher**, a filter based on POS sequences, designed to identify and correct mistagged nominalizations.

In essence, NomCatcher searches for sequences that look like noun phrases lacking their head noun. NomCatcher targets any sequence with one or more article-like elements (tags D,S,O,$,Z) followed by some combination of the same tags, adjectives, and punctuation, and ending in an adjective. When this sequence is followed by end-of-sentence punctuation or a verb, NomCatcher changes the final A tag to N.

```
you_O can_V tell_V a_D gay_A is_V
from_P florida_^ just_R by_P
looking_V at_P them_O
```

In the example above, the tag for *gay* is changed from A to N. TweetNLP and NomCatcher are applied to both TwTargets and TwOpen.

4.3 Annotation by the crowd

For each of the four target forms, 200 instances were selected at random, evenly split between N and A. The instances were shuffled and split into

[5] https://github.com/DocNow/twarc

[6] Additional parameters: restricted to English-language tweets occurring in the prior 7 days, data downloaded on April 25th and 26th, 2017.

[7] Harvested on April 28th, 2017.

[8] https://gate.ac.uk/wiki/twitter-postagger.html

[9] http://www.cs.cmu.edu/ ark/TweetNLP/

n=800	5agree		4agree		3agree		NoMaj	
	14.5%		26.3%		46.9%		12.3%	
Sent.Label	N	A	N	A	N	A	N	A
NEG	72	16	61	51	71	67		
NEUT	11	8	34	49	83	120		
POS	3	6	4	12	14	20		
							47	51

Table 5: Degree of overlap between crowd annotators, per LF and per label, TWTARGETS.

Majority vote	**Adj** 400	**Noun** 400
%NEGATIVE	33.5	**51.0**
%NEUTRAL	**44.2**	32.0
%POSITIVE	**9.5**	5.2
%NOMAJ	12.8	11.8

Table 6: Correspondence between pejorative meaning and linguistic form, majority vote from crowd annotations, TWTARGETS.

5 batches. Each batch was combined with 10 instances from PEJNOM-EXP, without considering the LF of the additional 10 instances. Each batch of 50 instances was labeled by 5 different annotators via the crowd-sourcing platform Amazon Mechanical Turk.[10] Participation was restricted to Amazon MT Masters only, and annotators were paid US$0.50/batch.

Annotators were instructed to indicate whether certain highlighted words (one word highlighted per instance) were used with POSITIVE, NEGATIVE, or NEUTRAL meaning. The following three examples were given as part of the instructions:

a POSITIVE: If you want the job done right, ask a **female** to do it.

b NEGATIVE: I don't understand why **females** think they know how to drive.

c NEUTRAL: My first pet ever was a **female** lizard.

Annotators were warned about potentially offensive data, told that the data would help develop systems for automatically detecting negative uses of words, and reminded to mark "sentiment for the word itself, not for the entire tweet."

Agreement between annotators. Table 5 presents detailed counts of the overlap between the 5 crowd annotators per instance. A clear majority vote (**3agree**) can be established for more than 85% of the 800 instances annotated by Turkers, and complete agreement (**5agree**) was reached for almost 15% of the cases. The full-agreement instances are mostly nouns, and mostly labeled NEG. Overall, the POS label is used infrequently, and crowd annotators tend to agree more on labels for nouns than for adjectives.

4.4 Analysis: correspondence between pejoration and linguistic form

Finally, we look at whether the hypothesis that nominalized occurrences of these adjectives tend to be used with negative meaning holds up in the non-expert setting.

Table 6 shows the correspondence between automatically-tagged LF and whether the majority vote of the annotators was Negative, Neutral, or Positive. For completeness, we include cases where no majority was reached.

Of the instances tagged with A, almost 54% are labeled as non-pejorative (majority vote: at least 3/5 annotators), counting both NEUT and POS as non-pejorative labels. 51% of the instances tagged with N are labeled as pejorative (NEG), with 37% receiving non-pejorative labels.

The numbers are small but encouraging, especially given that these are crowd-sourced annotations, annotators received no training, and no annotations were rejected. Despite clear instructions, in a number of cases it appears that annotators considered the sentiment of the entire tweet instead of just the word in question. For example, a majority vote of NEG was made for the following tweet:

24. *lol that's the best reason you could come up with in response to a group of* ***gays*** *supporting muslims?*

Nothing in the tweet itself suggests that this nominal use of **gays** is pejorative, and annotators were not given any additional context for the tweets.

5 Investigations

Expert annotations vs. crowd annotations. As a sanity check, 50 instances for each of the four target forms in PEJNOM-EXP were submitted for crowd-sourced annotation. Table 8 shows the mappings from expert annotations to the majority vote from the crowd.

This analysis is only suggestive, given that so few of the 200 PEJNOM-EXP instances in this batch have labels other than PEJ. We can note a

[10] https://www.mturk.com/mturk/

(1) Characteristics of individual humans	muslims, blacks, immigrants, riches, whites, sexists, homosexuals, feminists, fascists, blondes, illiterates, liberals, stupids
(2) Human, lexically pejorative	criminals, terrorists, rogues, racists
(3) Human-related, but unlikely to be pejorative	others, individuals, (10-year-)olds, browns (sports team)
(4) Non-human	news, rights, lives, extremes, likes, standards, tops, seconds, presents, riches, shorts, graphics, finals, nonprofits, offensives, positives, evils, ideals
(5) Verbs	owns, lives, opens, likes, lasts, tops, seconds, presents, grosses,

Table 7: Lexical items identified as potential pejorative nominalizations. Shown in plural form.

Expert	**CS: Neg**	**CS: Neut**	**CS: Pos**	**NoMaj**
PEJ	86	26	1	12
NONP	11	16	2	4
SAT	13	4	1	4
UNC	2	11	0	5
NOAN	2	0	0	0

Table 8: Correspondence between crowd-sourced labels (majority vote) and expert annotations, 200 instances from PEJNOM-EXP. Numbers are counts, not percentages.

few tendencies: PEJ instances are largely marked as NEG, NON-PEJ instances are divided between NEG and NEUT, sarcastic utterances tend to be labeled as NEG, and UNC cases either are marked as NEUT or fail to reach a majority.

Identification of new pejorative nominalizations. Our analysis so far is restricted, treating just four adjectives. With NomCatcher (Section 4.2), we can quickly and automatically identify new adjectives that undergo the same kind of meaning shift. We apply NomCatcher to TWOPEN and to the hate speech corpus of (Waseem and Hovy, 2016), finding words whose POS tag is changed by NomCatcher from `A` to `N`.

From the 16K tweets in the hate speech corpus, NomCatcher's filter identifies 206 distinct lexical items. Some are good catches, but the majority are proper adjectives occurring between a determiner/article and a noun mistagged as `V`, as in `[their_D hypocritical_A whining_V]`. To narrow down the set of adjectives identified, a second filtering step is applied, checking the corpus for plural forms of the 206 words caught by NomCatcher. This step cuts the number of word types identified down to 43, which can be grouped as in Table 7.

Row 1 contains forms denoting human characteristics; these are the most likely to undergo semantic transformation to pejorative meaning. Row 2 contains human characteristics which are inherently pejorative. Row 3 is especially interesting; the two high-frequency forms (*others* and *individuals*) both avoid mentioning any particular characteristic. Rows 4 and 5 are not relevant for abusive language, as they are not referential to humans.

NomCatcher has similar results for our TWOPEN corpus; 314 lexical items are filtered down to 90. The categories remain the same, and the overlap with the words identified from the hate speech corpus is high.

6 Conclusions and future work

The aim of this work is to detect pejorative uses of lexical items that can be used either in completely harmless ways or in ways that are abusive and harmful. This is a challenging task, given that it relies on many layers of human interpretation.

Our approach focuses on the role of linguistic form, and our two corpus studies support the hypothesis that certain adjectives, when used as nouns, acquire pejorative meaning. The NomCatcher tool uses LF for quick identification of likely candidates for pejorative nominalization. Immediate next steps are to explore the effectiveness of sentiment analysis methods for this task.

As the work progresses, we will deepen the current analyses and expand the data sets, applying our methods to a large Reddit corpus, and eventually incorporate linguistic form into a full system for detecting abusive language online.

An exciting avenue for future inquiry is the role of sarcasm. Existing work identifying sarcasm on Twitter (Sulis et al., 2016; Ling and Klinger, 2016; Wang, 2013) finds that sarcastic tweets tend to express pejorative meaning with positive words. The sarcastic instances in our data show a different pattern, using pejorative nominalizations with other negative words to mock discriminatory mindsets, in the end conveying negative sentiment towards those who use this type of abusive language.

Acknowledgments

First and foremost, thanks to Patricia Cukor-Avila for many useful discussions, starting from the very beginning of this project. Thanks go to Mary Burke and Karla Villarreal for their short-notice expert annotations, and to the many unnamed crowd annotators involved in this project. We also thank our four anonymous reviewers for interesting comments and useful suggestions. This work was funded in part through research support to the first author from the UNT Department of Linguistics and the UNT College of Information.

References

Eric Acton. 2014. *Pragmatics and the social meaning of determiners (Doctoral dissertation)*. Ph.D. thesis, Stanford, CA.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics.

Christine Günther. to appear. The rich, the poor, the obvious – Arguing for an ellipsis approach to "adjectives used as nouns". In *NPs in English: past and present*. John Benjamins.

Mark Liberman. 2016a. "Ask the gays". http://languagelog.ldc.upenn.edu/nll/?p=26223.

Mark Liberman. 2016b. The NOUNs. http://languagelog.ldc.upenn.edu/nll/?p=26254.

Jennifer Ling and Roman Klinger. 2016. An empirical, quantitative analysis of the differences between sarcasm and irony. In *International Semantic Web Conference*. Springer, pages 203–216.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, WWW '16, pages 145–153. https://doi.org/10.1145/2872427.2883062.

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. Association for Computational Linguistics.

Latino Rebels. 2016. Maria Hinojosa Tells Latino Trump Surrogate on MSNBC's AM JOY That 'Illegals Is Not a Noun'. http://www.latinorebels.com/2016/10/29/maria-hinojosa-tells-latino-trump-surrogate-on-msnbcs-am-joy-that-illegals-is-not-a-noun-video/.

Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo. 2016. Figurative messages and affect in twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems* 108:132–143.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, pages 173–180.

Po-Ya Angela Wang. 2013. #Irony or #Sarcasm - A Quantitative and Qualitative Study Based on Twitter. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, Stroudsburg, PA, USA, LSM '12, pages 19–26. http://dl.acm.org/citation.cfm?id=2390374.2390377.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California.

Anna Wierzbiecka. 1986. What's in a noun? (or: How do nouns differ in meaning from adjectives?). *Studies in Language* 10(2):353–389.