# A Twitter-Based Study of Newly Formed Clippings in American English

Sravana Reddy, Joy Zhong, James Stanford

Dartmouth College

**Samantha Wolfe** @smileygirlsam 17 Dec

This little boy is so **presh**, he's like a cherub. So **adorbs**.

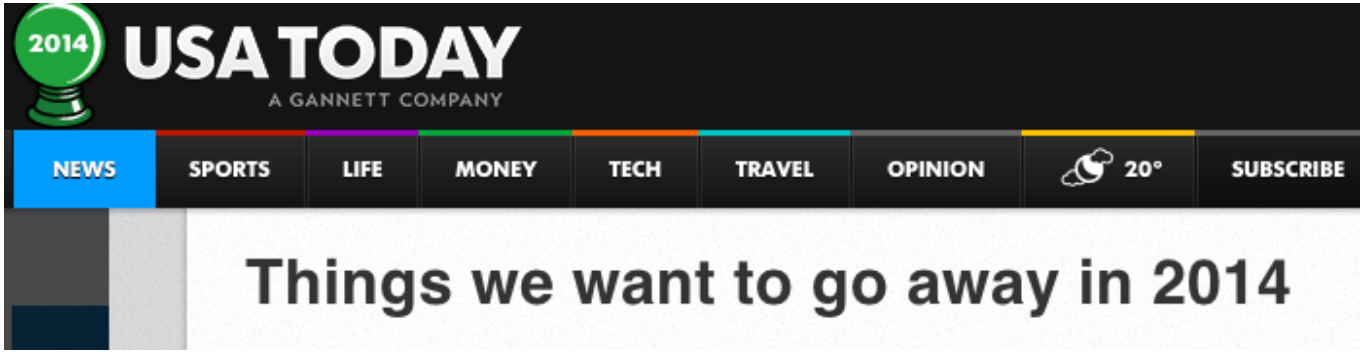
Expand Reply Retweet Favorite More

**Meg Partridge** @_MPartridge 26 Aug 12

@Connor_3592 We took free champagne, got me in everywhere, made a **totes awks** situation **totes hilar** and then got a free taxi home #MINT

**Vanessa, De Su Mama** @DeSuMama 4 Sep

Checked-in at @WestinDiplomat and lobby is **gorg**! Beach day w/**fam** tomorrow before fun w/friends at #nicheparent13 pic.twitter.com/dgZRpkOmjG




Related


Sprint honors

# Previous Work

Baclawski (2012)

A study of -s (‘adorbs’) in 40 Twitter users

# Research Questions

- Are these clippings just cyclical "slang"? (Eble 1996, 2004)
- Are they an increasingly productive process with new social meanings?
- Is this type of clipping more productive than past generations?
- What is the role of the –s suffix (*adorbs, awks, totes*)?
- Which speakers use it the most? Age, gender, ethnicity?

# This Study

*Hypothesis*:
**Women** are leading in the usage of these new clippings, and it is more **urban/suburban** than rural

Labov (1990, 2001), Trudgill (1972), Coates & Pichler (2011), Holmes & Meyerhoff (2003), Wolfram & Schilling-Estes (2006:155-6)

# Why use Twitter for American Dialect Research?

Each era applied contemporary technology…

- Kurath (1939)
- Hanley's recordings (1931-1937) (Purnell 2012)
- Chambers & Trudgill (1998)
- Labov, Ash & Boberg (2006)
- Kretzschmar (2009)

and many more

*Now*: Social Media analysis, computational modeling, Mechanical Turk

# Twitter for Sociolinguistics

– Eisenstein, O'Connor, Smith & Xing (2010)

US regional variation in lexical items

– Bamman, Eisenstein & Schnoebelen (2012)

Gendered language and networks

– Maybaum (2012)

Twitter terms

– Zappavigna (2013)

Twitter discourse and variation

– Doyle (2014)

Geographic distribution of "needs done"

# Methodology

- Collected 185 million geo-tagged tweets originating in the US (Jul-Nov 2013) by 893,024 users
- Automatically extracted a list of clippings
- For each word, created demographic profile of users
  - Gender
  - Population, median age, and ethnic distribution at the user's location
- Compared demographic features of clipping and its original form

# Extracting Clippings

- Rather than manually compiling list of clippings, automatically learn from Twitter data
- A clipping and its original form will be used in roughly similar contexts

# Extracting Clippings

- Represent every word type as vector of its left and right context
- Rank every word pair by context vector similarity
- Extract top ranked pairs where first three characters match

totes
left: am, are, was, were, is..
right: okay, ok, adorbs, fine…

adorable
left: is, so, these, looks…
right: omg, with, dork,…

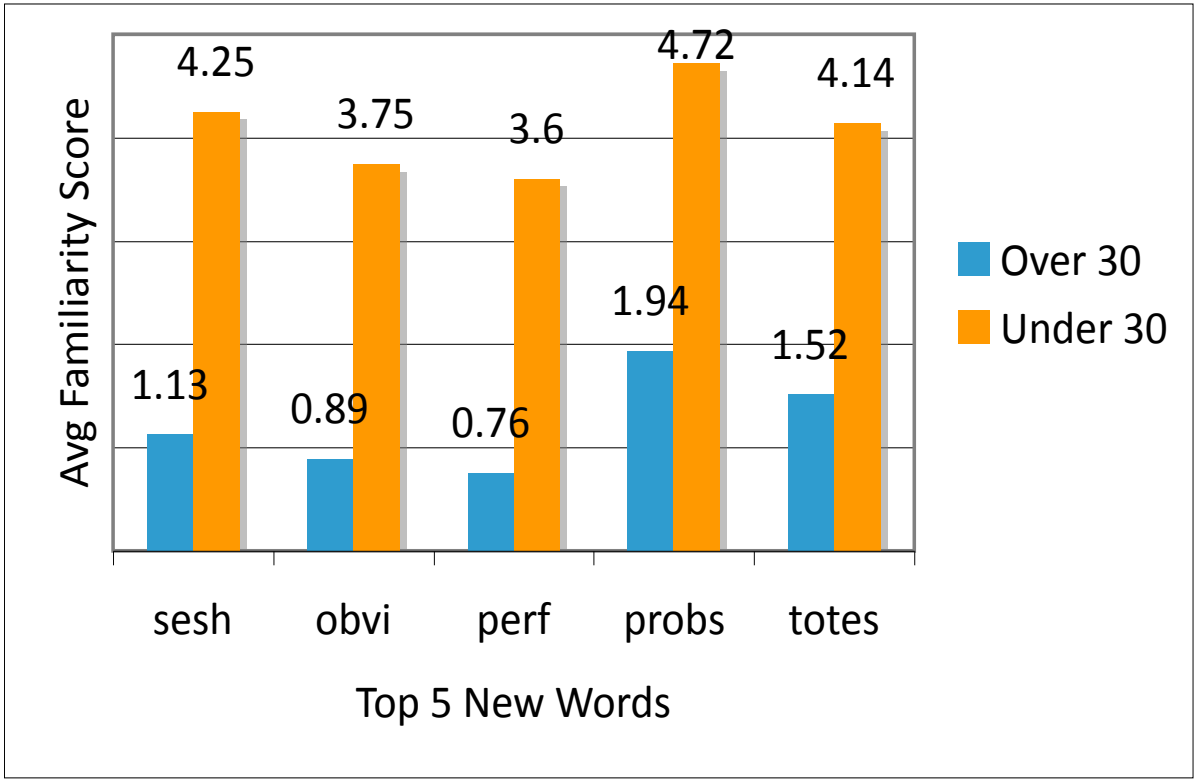
totally
left: am, are, was, were, is…
right: okay, fine, insane, not…

# Old vs. New Clippings

- Survey on Mechanical Turk
- Demographic questions: age, gender, location
- Rate familiarity with each clipping
  - Unfamiliar
  - Familiar, but I do not use it
  - I use it in speech only
  - I use it in writing only
  - I use it in speech and writing
- Same survey also conducted with Dartmouth undergraduate students

# Old vs. New Clippings

- Split survey respondents into ages 18-29 and 30+
- For each clipping, compute average familiarity score within the two age groups




0 = Unfamiliar

3 = Familiar, but I do not use it

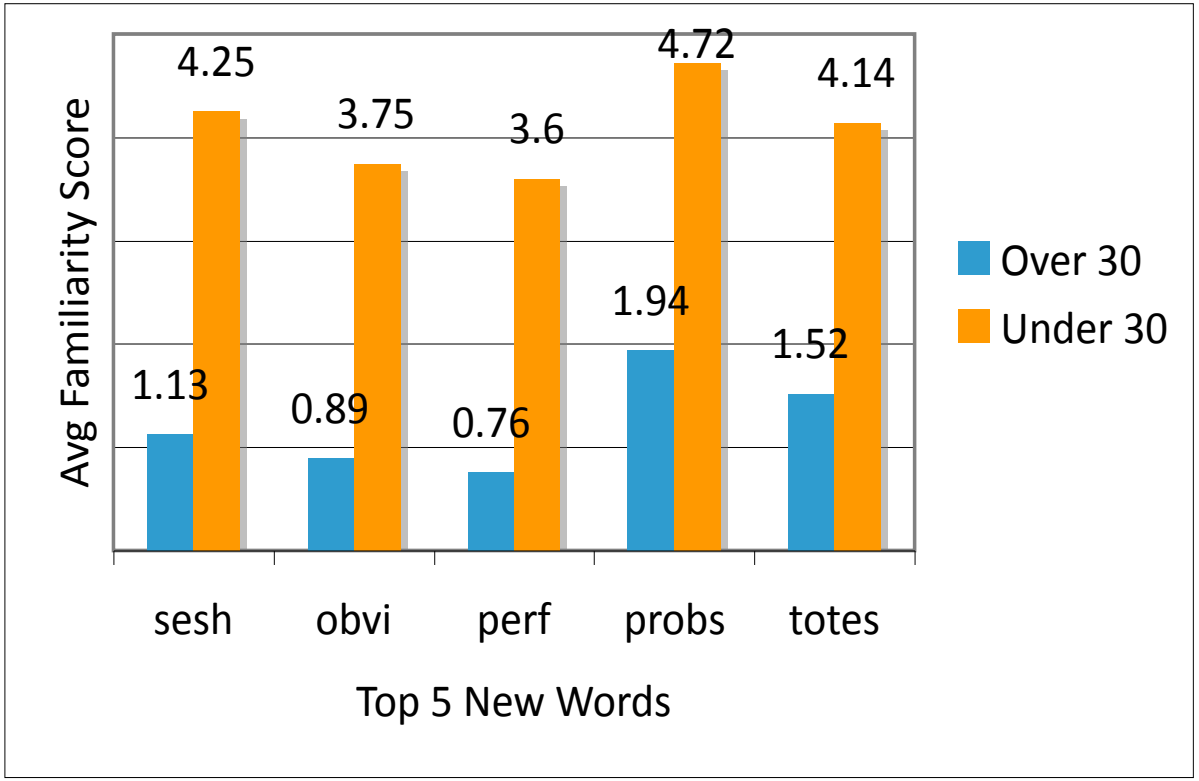
4 = I use it in speech only

5 = I use it in writing only

6 = I use it in speech and writing

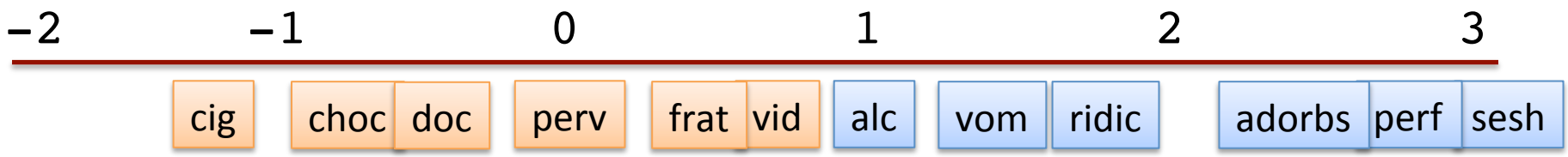
# Old vs. New Clippings

- Newness score for clipping
  = below 30 familiarity – above 30 familiarity



# Old vs. New Clippings

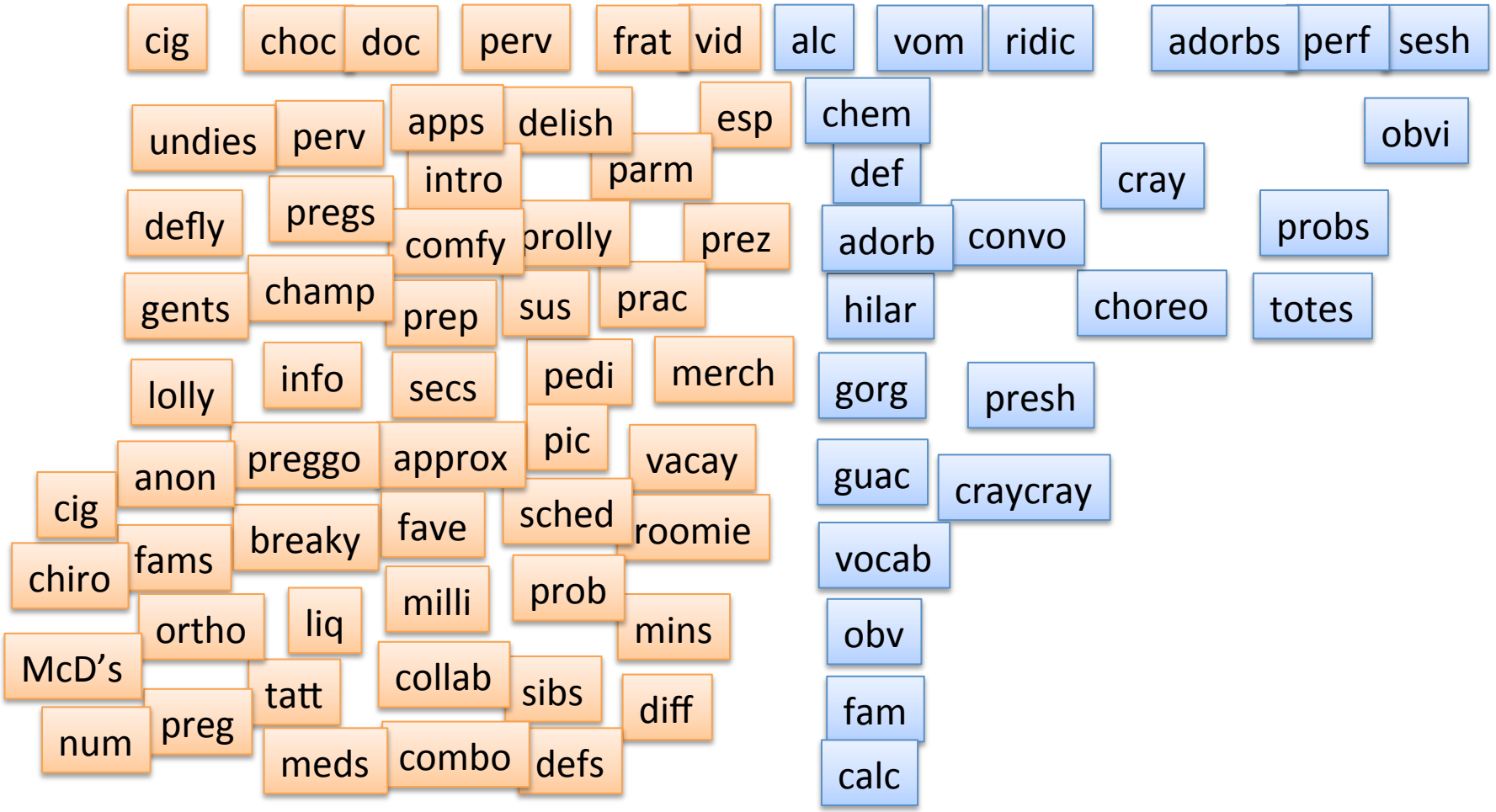
- Newness score for clipping
  = below 30 familiarity – above 30 familiarity
- Threshold at 1.0 newness score




Old clippings (55)

New clippings (25)

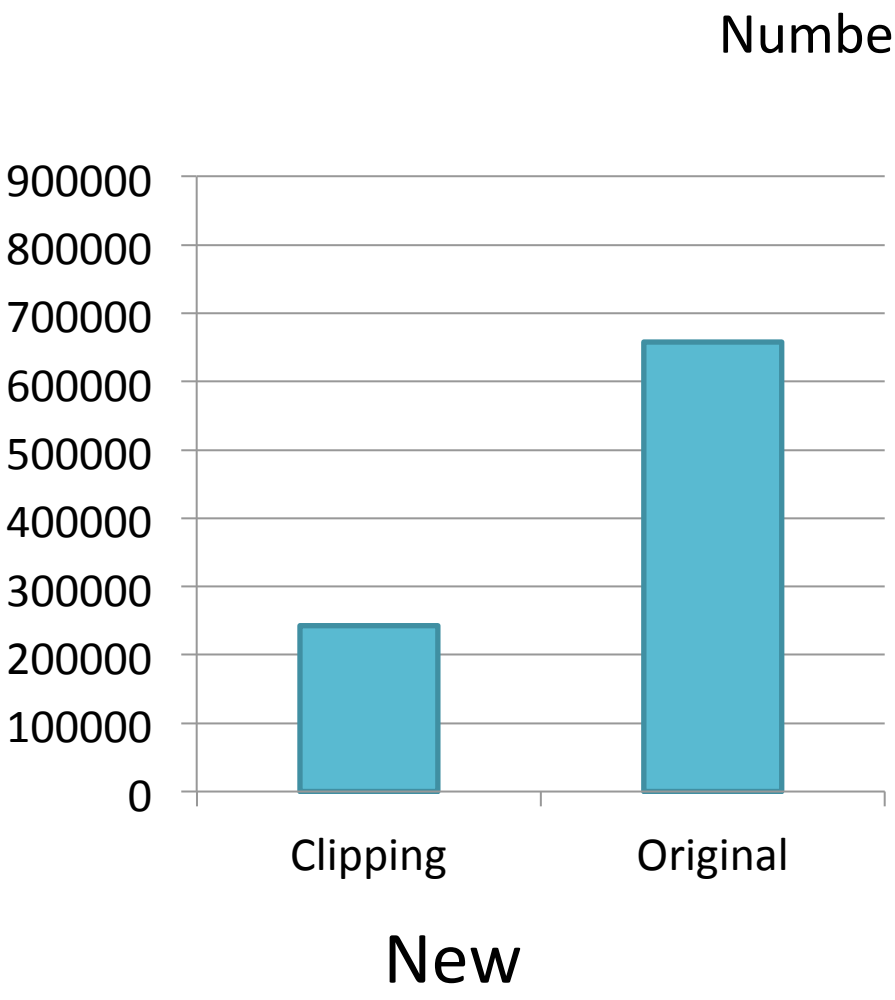
# Old vs. New Clippings

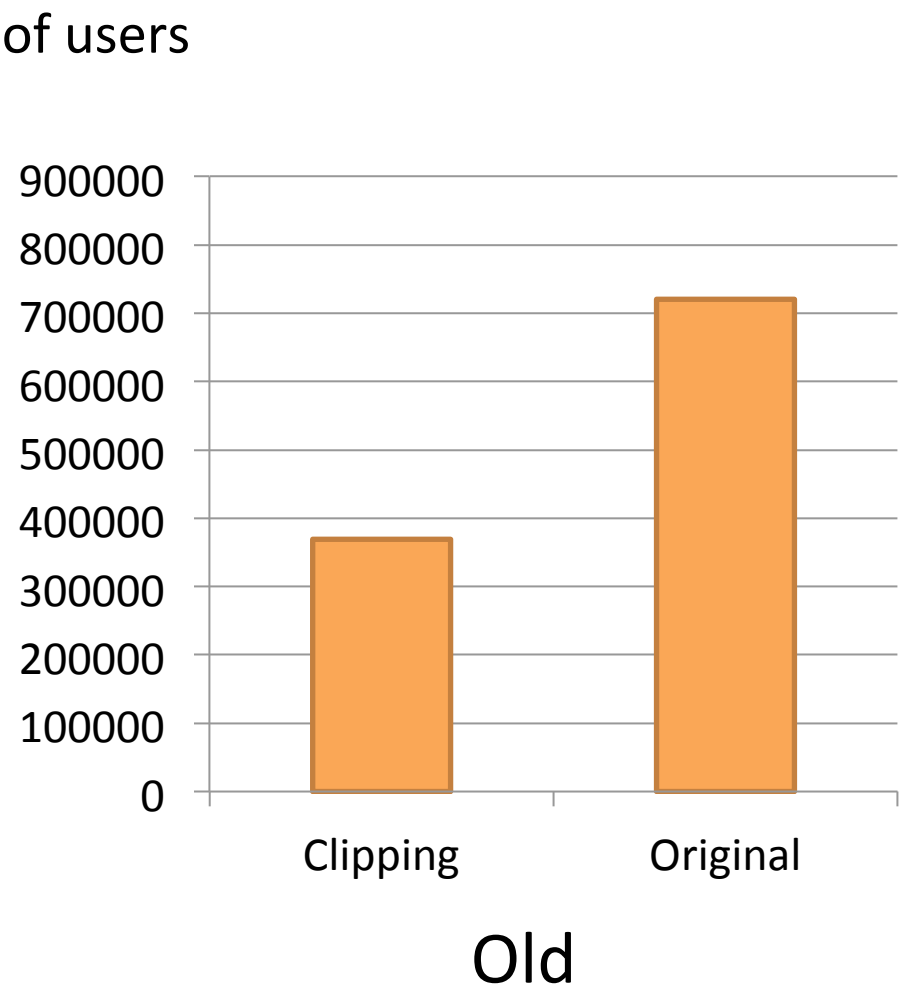



Old clippings (55)

New clippings (25)

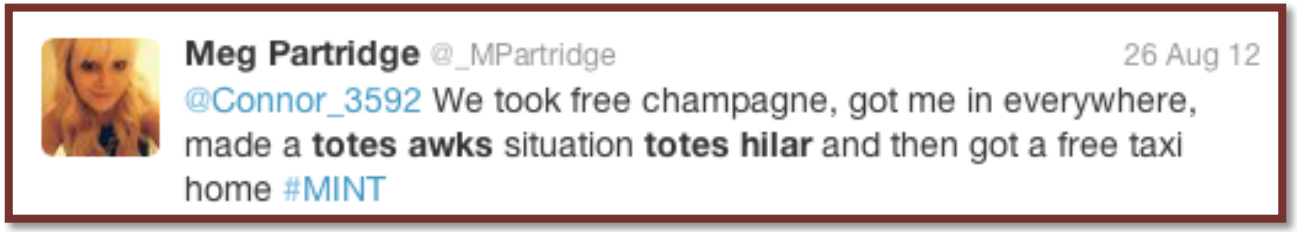
# Clippings on Twitter



# Demographic Analysis

- Gender
  (following Bamman et al.)
  - Most Twitter users report a name in addition to their pseudonym




  - Match first name against the Social Security Administration list of baby names born in 1995
  - About 2/3 of users have names in the SSA list and are assigned a gender

# Demographic Analysis

- Location

  (following Eisenstein et al.)

  – Tweets are geo-tagged with latitude/longitude

  – Map each geo-coordinate to one of 33000 Zip Code Tabulation Areas (ZCTAs)

  – Ignore users that tweet from more than one ZCTA

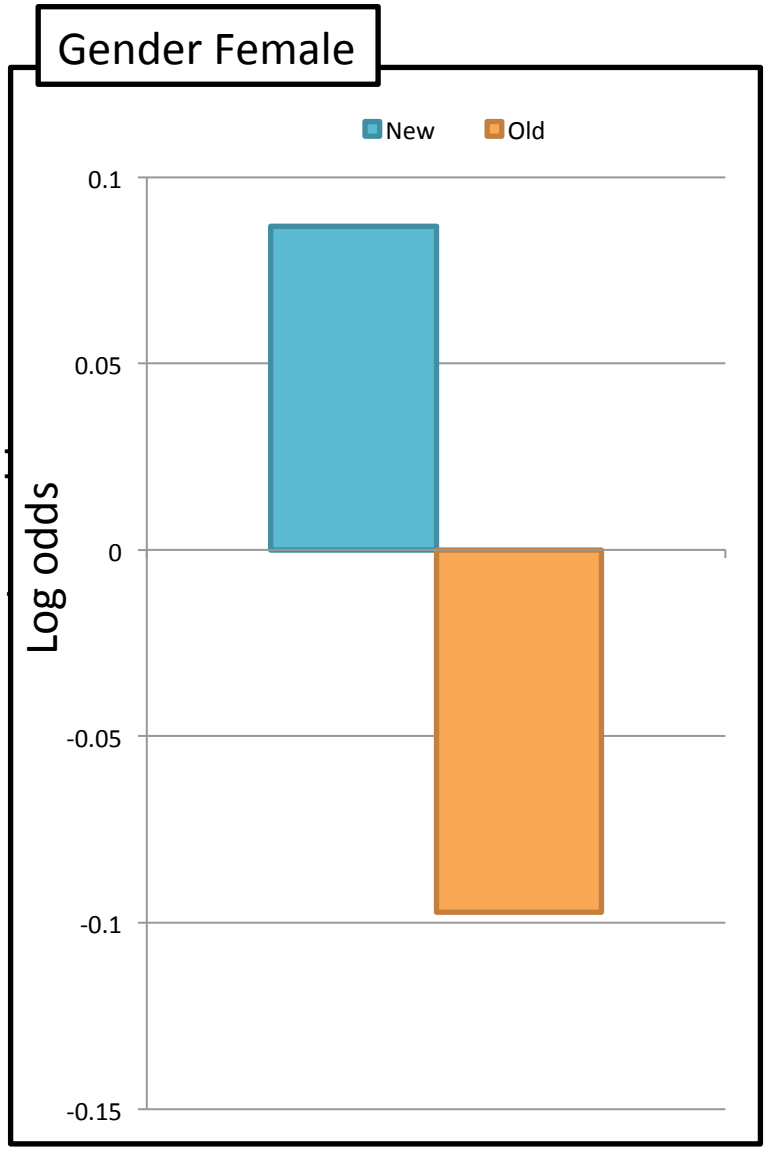
  – Get demographic attributes of ZCTAs from 2010 Census: Population, Median Age, White%, African American%, Asian%, Native American%, Hispanic%

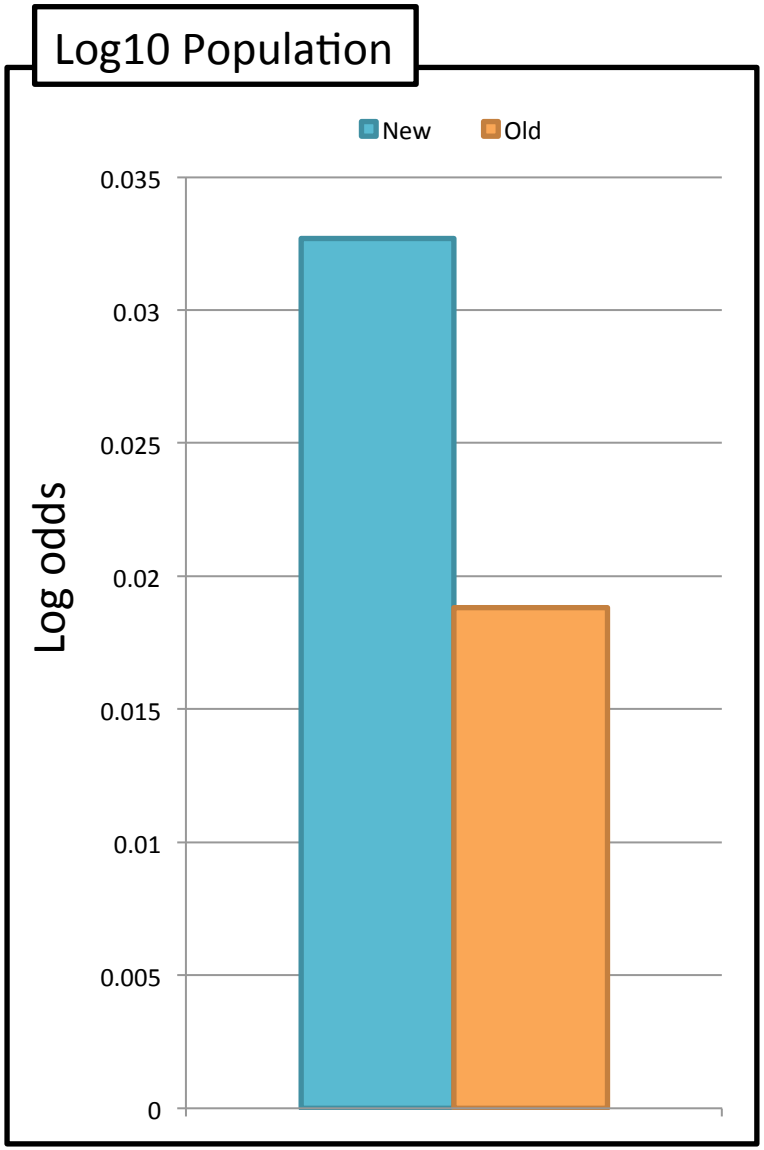
  – Each user is now associated with a demographic profile of their environment

# Demographic Analysis

- Logistic regression
  - Predicted variable: clipping or original?
  - Features: demographic profile of users
    - Gender
    - Population
    - Median Age
    - Ethnicity

# New and Old Clippings

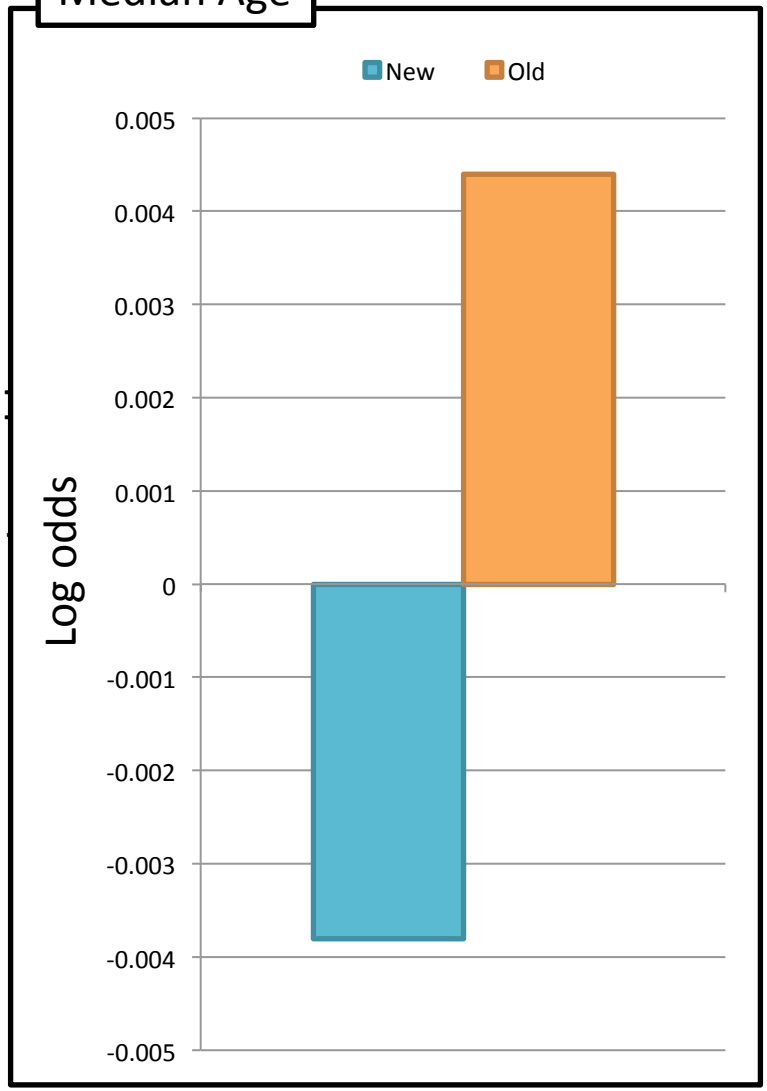



All factors shown are significant ($p<0.05$)

# New and Old Clippings

## Median Age

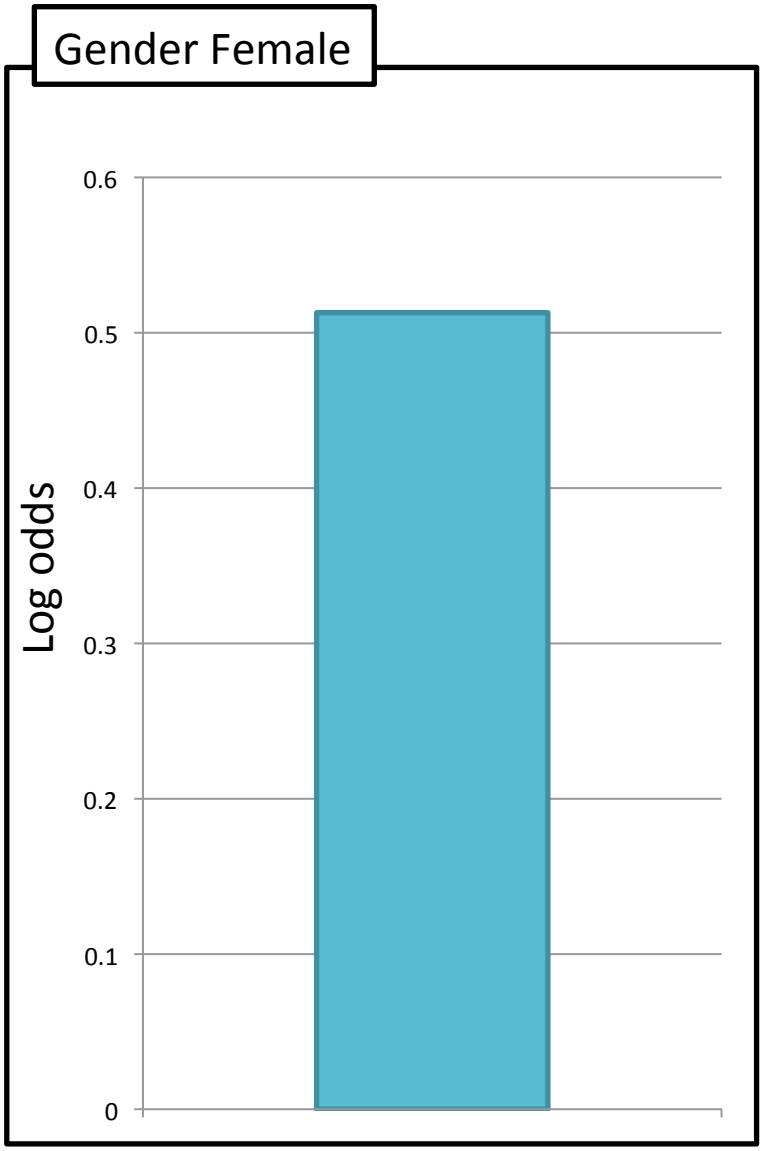


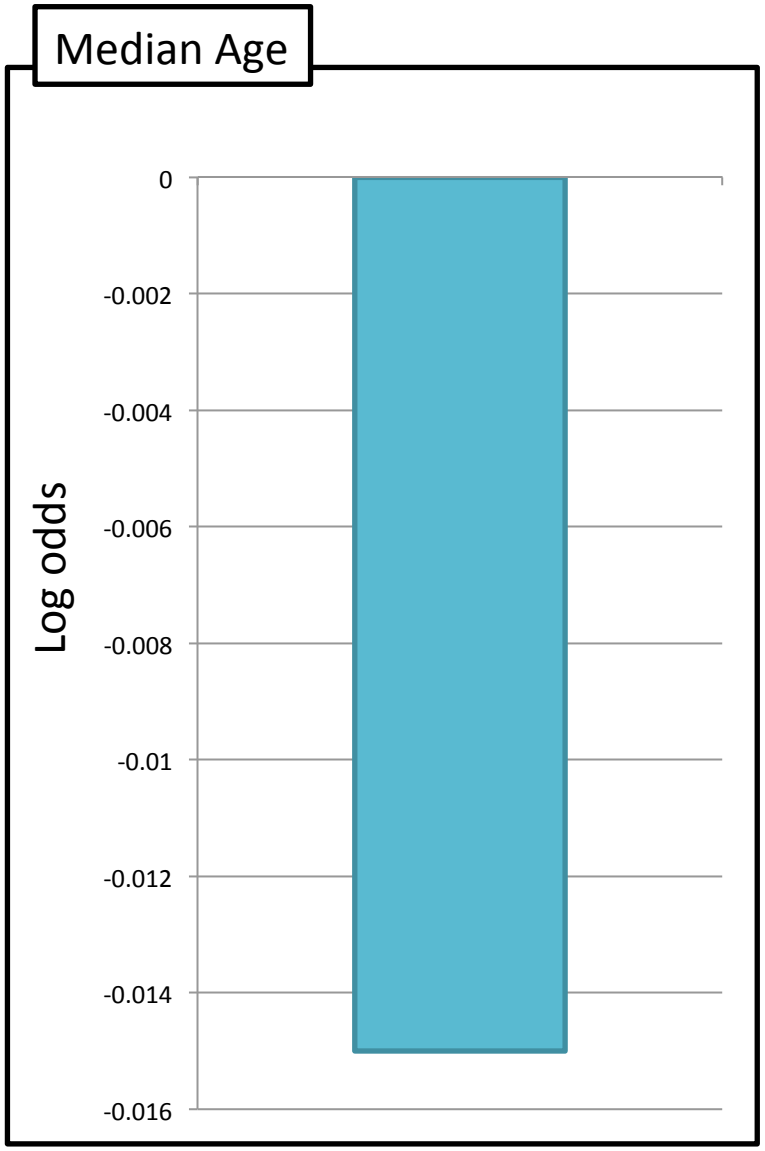

## Ethnicity

No Significant Effects

# Usage of -s suffix in clippings

adorbs/adorb,
probs/prob,
fams/fam,
awks/awk,
pregs/preg,
defs/def



# Conclusion

*Hypothesis Confirmed*:
**Women** are leading in the usage of these new clippings, and it is more **urban/suburban** than rural

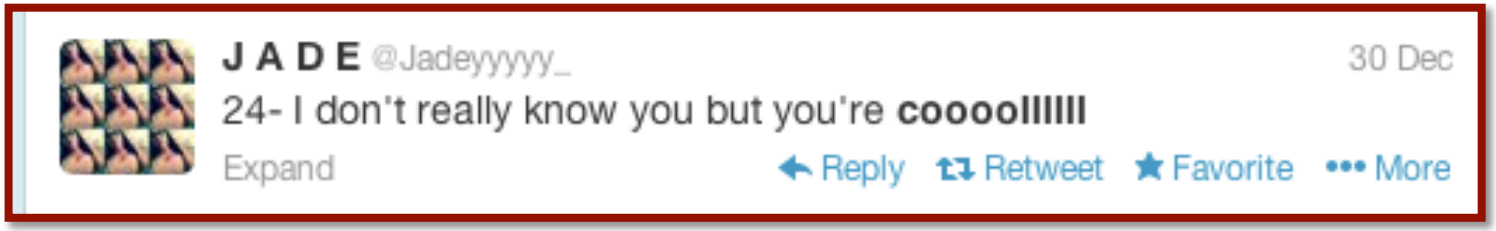
# Are clippings a Twitter artifact?

## They abound in long-form blog posts too…

I did my weekly grocery shop this morning and whilst I was loading the car I noticed a van driving slowly around the car park. A bit sus I thought. But then I saw the sign on the side, DVLA, I

I just tried to click on a theme that was pink, and was informed it was 75 dollars. That is *ridic*. I'll have to be figuring out how to make this pink all on my own.

## … and Twitter users often lengthen words



# Are clippings a Twitter artifact?

Baldwin et al. (2013) measure average word lengths in Twitter and different corpora

| Corpus | Word length | Sentence length |
|---|---|---|
| TWITTER-1 | 3.8±2.4 | 9.2±6.4 |
| TWITTER-2 | 3.8±2.4 | 9.0±6.3 |
| COMMENTS | 3.9±3.2 | 10.5±10.1 |
| FORUMS | 3.8±2.3 | 14.2±12.7 |
| BLOGS | 4.1±2.8 | 18.5±24.8 |

# Are clippings a Twitter artifact?

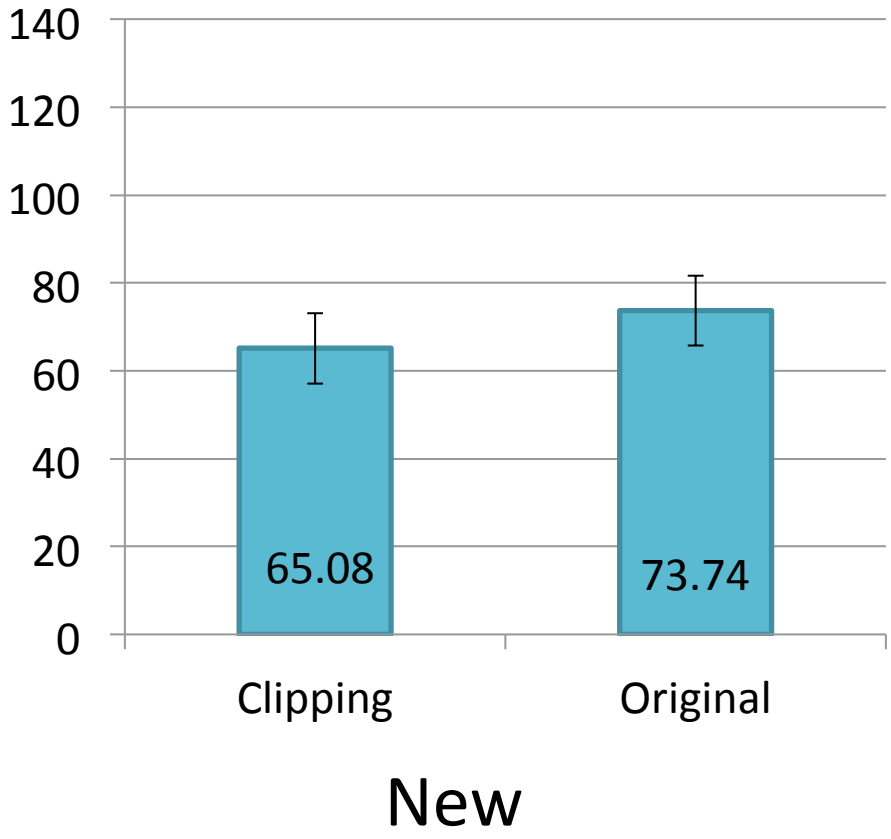
Eisenstein et al. (2013) find shortened forms are mainly used in tweets of length much less than 140 characters. Shortening is not used in order to fit length constraints!

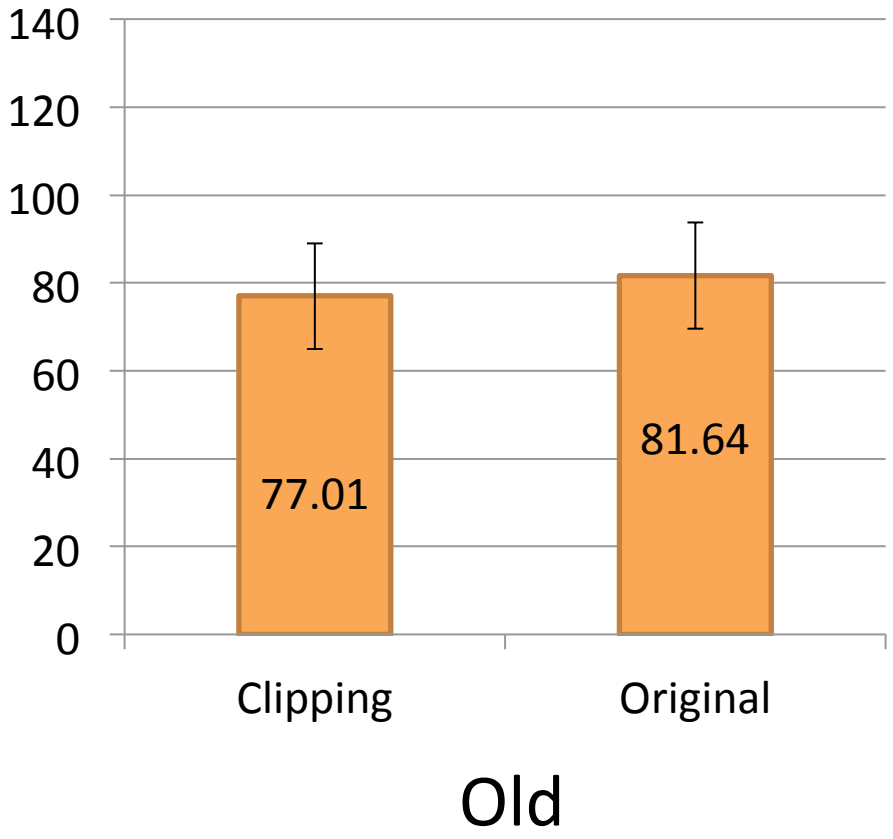
| standard | length | alternative | length |
|---|---|---|---|
| *your* | $85.1 \pm 0.4$ | *ur* | $81.9 \pm 0.6$ |
| *you're* | $90.0 \pm 0.1$ | | |
| *with* | $87.9 \pm 0.3$ | *wit* | $78.8 \pm 0.7$ |
| *going* | $82.7 \pm 0.5$ | *goin* | $72.2 \pm 1.0$ |
| *know* | $86.1 \pm 0.4$ | *kno* | $78.4 \pm 1.0$ |
| *about* | $88.9 \pm 0.4$ | *bout* | $74.5 \pm 0.7$ |

Table 1: Average length of messages containing standard forms and their shortenings

# Are clippings a Twitter artifact?

Our experiment: avg lengths of tweets containing clippings compared to tweets with original forms



# Future Work

- Track spread of clippings in Twitter over time
  - Will these clippings spread throughout the population?
  - Geographic/demographic dimensions of spread?
  - When did these clippings originate?
- Morpho-phonological study of clippings