#go awn: Sociophonetic Variation in Variant Spellings on Twitter

Rachael Tatman – rctatman@uw.edu
The University of Washington

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082.

Outline

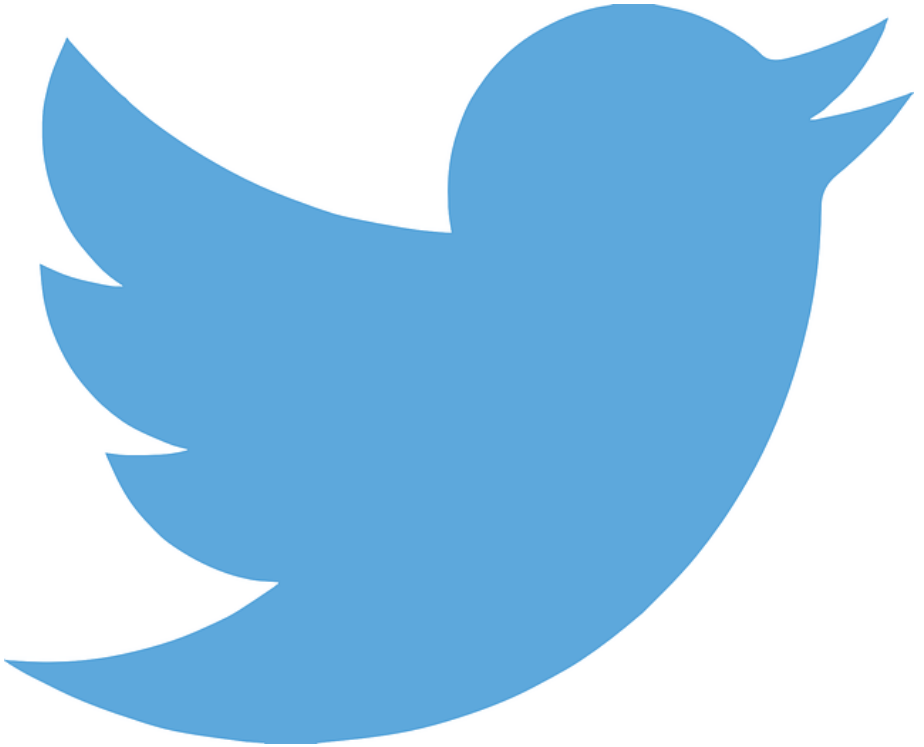
1) Research Questions
2) Background
3) Southern American English
4) Scottish English
5) Conclusion

Outline

1) **Research Questions**
2) Background
3) Southern American English
4) Scottish English
5) Conclusion

Research Questions

- Can we apply variationist methods to Twitter data?
- Do Twitter users use variant spellings to encode sociophonetic variation?
- How do variant spellings interact with style?



Outline

1) Research Question
2) **Background**
   - Computer Mediated Communication & Variationist Sociolinguistics
   - Advantages of Twitter Data
   - Variant Spellings
3) Southern American English
4) Scottish English
5) Style
6) Conclusion

CMC & Variationist Sociolinguistics

- Relatively few variationist studies of computer mediated communication (CMC)
- Two main reasons (Androutsopoulos 2006)
  - Unreliable or missing demographic information
  - Lack of phonetic/phonological information
- But...

CMC & Variationist Sociolinguistics

- Variation in computer mediated communication is systematic and mirrors that found in speech
- This has been shown for:
  - Text messaging (Thurlow & Brown 2003)
  - Internet Relay Chat (IRC) (Siebenhaar 2006)
  - Blogs (Herring & Paolillo 2006)
  - Instant messaging (Tagliamonte & Denis 2008)
  - Twitter (style accommodation) (Danescu-Niculescu-Mizil et al. 2010)
- Linguistic variation can also be used to identify user demographic information (Rao et. al 2010)

Using Twitter Data

Pros:

- Large quantity of data already available
- Fast data collection
- Reproducible research
- Limits the effects of the Observer's Paradox (Labov 1972)

Cons:

- Limited demographic information
- Limited control over data production
- Too much data (“firehose”)
- No phonetic data available

Using Twitter Data

Pros:

- Large quantity of data already available
- Fast data collection
- Reproducible research
- Limits the effects of the Observer's Paradox (Labov 1972)

Cons:

- Limited demographic information
- Limited control over data production
- Too much data (“firehose”)
- No phonetic data available

Variant Spellings

- Non-standard orthographic representations of words
- Also called “dialect orthography” (Krapp 1919)
- Spelling in CMC contexts is more variable, allowing for identity construction using variant spellings (Sebba 2003)
- Does it encode sociophonetic variation?

Advertising image used by the Center for the Psychology of Women in Seattle. Image retrieved from http://psychologyofwomen.com/wings/ on April 21, 2015.

Variant Spellings

- Possibilities:
- Variant spellings are treated like lexical items with a different meaning than the standard spelling
  - Examples:
    - “go awf”: expression of approval and solidarity, used mainly by African American women (author's impression)
    - “hawt”: note that /hɑt/ is not produced as /hɔt/ by speakers without the low back merger (Labov, Ash & Boburg 2005)
- Variant spellings are used during style shifting as a way of encoding sociophonetic variation and can be extended to new lexical items
  - Examples:
    - “spelunkin”: used as song title: "Monster Spelunkin" (Tran & Velema 2014) . Unlikely to be a separate lexical item.

Outline

1) Research Question
2) Background
3) **Southern American English**
   - Methodology
   - Findings
4) Scottish English
5) Conclusion

Methodology

- One well-studied sociophonetic variable with a clear alternate spelling chosen
- High frequency words with that variable selected
- TwitteR package (Gentry & Gentry 2014) and Twitter public API used to find tweets which contained variant spellings of selected variable in high-frequency words
- Tweets sorted by hand
- Other variant spellings marked by hand
  - Do they contain other sociophonetic variables?
  - Do they pattern together in the same way they have been observed to in speech?

Methodology

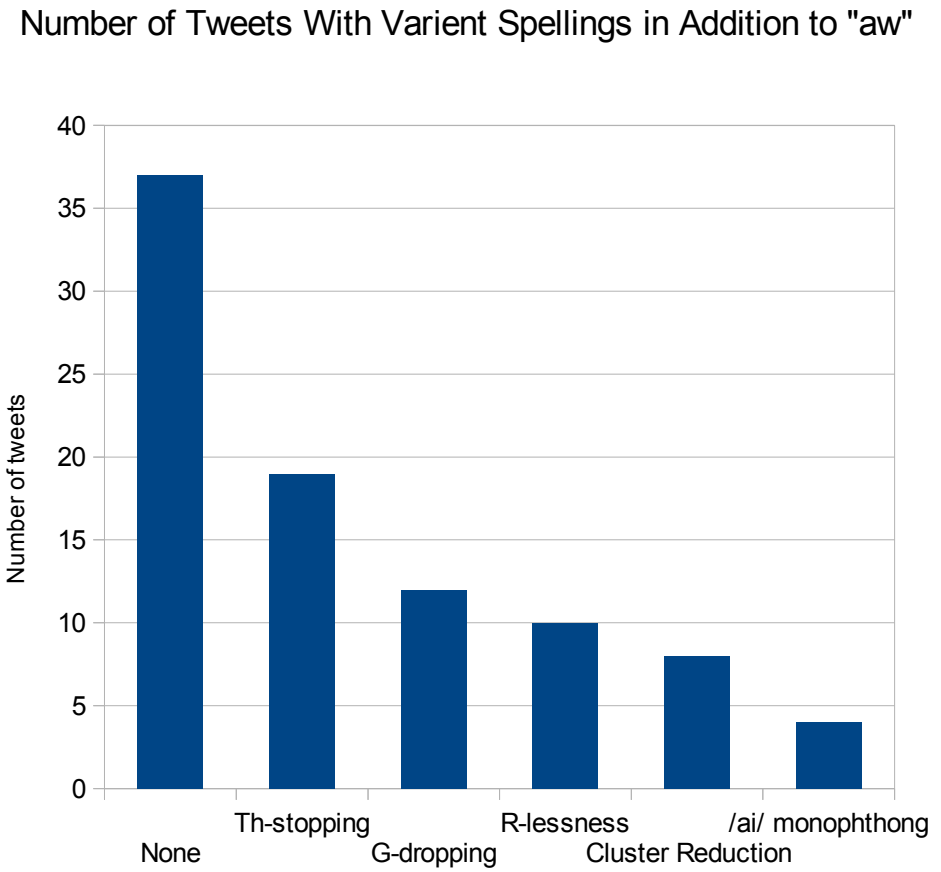
- One well-studied sociophonetic variable with a clear alternate spelling and unpredictable distribution chosen
  - Distinction between /ɑ/ and /ɔ/, with /ɔ/ spelled “aw”
  - For merged speakers, not possible to guess which is /ɔ/, as in “hawt”
  - Low back merger is a sociolinguistic marker of Southern American English and African American English (Labov, Ash & Boburg 2005)
- High frequency words with that variable selected
  - All of the 100 most-frequent English words with /ɔ/ form slected using the CMU pronouncing dictionary (Weide 1998)
- Twitter public API used to find recent tweets which contained variant spellings of selected variable in high-frequency words
  - Code available on author's github page
- Continued...

Methodology

- Tweets sorted by hand
  - Removed tweets where the search variable occurred in the following:
    - Foreign words
    - Names/proper nouns
    - Universal resource locater (URL)
    - Clear typos:
      - “Awn thanks Merleen” for “Aww thanks Merleen” rather than “On thanks Merleen”
  - 74 tweets remained
- Other variant spellings marked by hand
  - Do they contain other sociophonetic variables?
  - Do they pattern together in the same way they have been observed to in speech?

Results

- 50% of tweets contained more than one sociolinguistics variables
- Other variables:
  - Th-stopping
  - G-dropping
  - R-lessness
  - Cluster reduction
  - /ai/ monophthongization
- Consistent with features found in Southern American and/or African American speech (Labov & Boburg 2005)



Results

Example:

Hype hayed foah dat becawse it was 8 bucks foah 2 yeahs and w da jets i like readin about da prospects ogay (JPG 2015)

"I paid for that because it was eight bucks for two years, and with the Jets [American football team] I like reading about the prospects, okay?"

- /ɔ/
- Th-stopping
- G-dropping
- R-lessness
- /ai/ monophthongization
  - Perhaps "hype"? "Like" is not
- Other
- Abbreviation

Results

- It seems that Twitter users are using multiple variant spellings together to encode phonetic variables
  - Consistent with phonological rather than lexcial use
  - Example: “hype hayed” returns one Google result
- This could be limited to one dialect, though
- Convergent findings are needed to verify the methodology

Outline

1) Research Question
2) Background
3) Southern American English
4) **Scottish English**
   - Methodology
   - Findings
5) Conclusion

Methodology

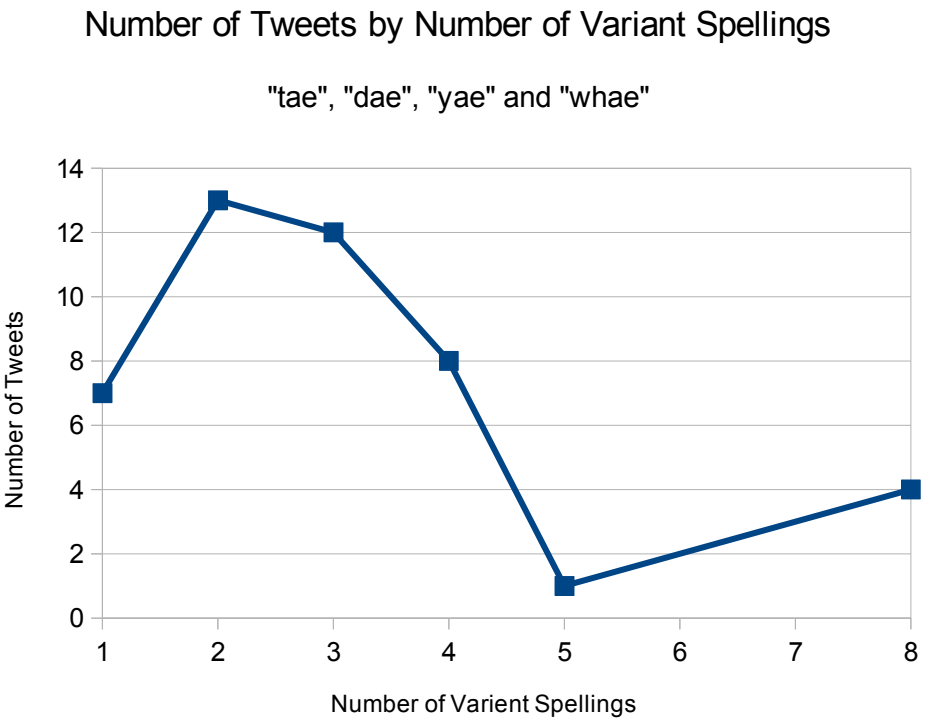
- One well-studied sociophonetic variable with a clear alternate spelling chosen
- High frequency words with that variable selected
- Twitter API used to find tweets which contained variant spellings of selected variable in high-frequency words
- Tweets sorted by hand
- Other variant spellings marked by hand

Methodology

- One well-studied sociophonetic variable with a clear alternate spelling chosen
  - [du] vowel produced [de] (Stuart-Smith 2004), commonly spelt “dae”
- High frequency words with that variable selected
  - All [u] words in the fifty most frequent English words (Davies 2011)
- Twitter API used to find tweets which contained variant spellings of selected variable in high-frequency words
- Tweets sorted by hand
  - 45 tweets remaining
- Other variant spellings marked by hand

Results

- 84% contained more than one variant spelling
- Average of 3 variant spellings per tweet
- Features:
  - [u] → [ʉ]
  - [ai] → [æ]
  - [l] vocalization: “fitba”
  - [ʊ] variant spellings
  - [ɔ] variant spellings
  - [ei] → [i]
- Consistent with features of Scottish Standard English (Stuart-Smith 2004, Renni 2001)



Results

Example:

dae ye ever look back oan how much time ye wasted oan someone nd wonder why naeone punched u in the heed (bj 2015)

"Do you ever look back on how much time you wasted on someone and wonder why no one punched you in the head?"

- [du] → [de]
- [u] → [ʉ]
- [ei] → [i]
- [ɔ]
- Abbreviation

Outline

1) Research Question
2) Background
3) Southern American English
4) Scottish English
5) **Conclusion**
   - Method
   - Style

Conclusion

- Do Twitter users use variant spellings to encode sociophonetic variation?
  - Yes, the use of variant spellings pattern with the sociophonetic variation observed in speech
- Can we apply variationist methods to Twitter data?
  - Yes! The method discussed here presents a principled way of looking at how Twitter users represent sociophonetic variation
  - Can be used to verify metalinguisitc awareness

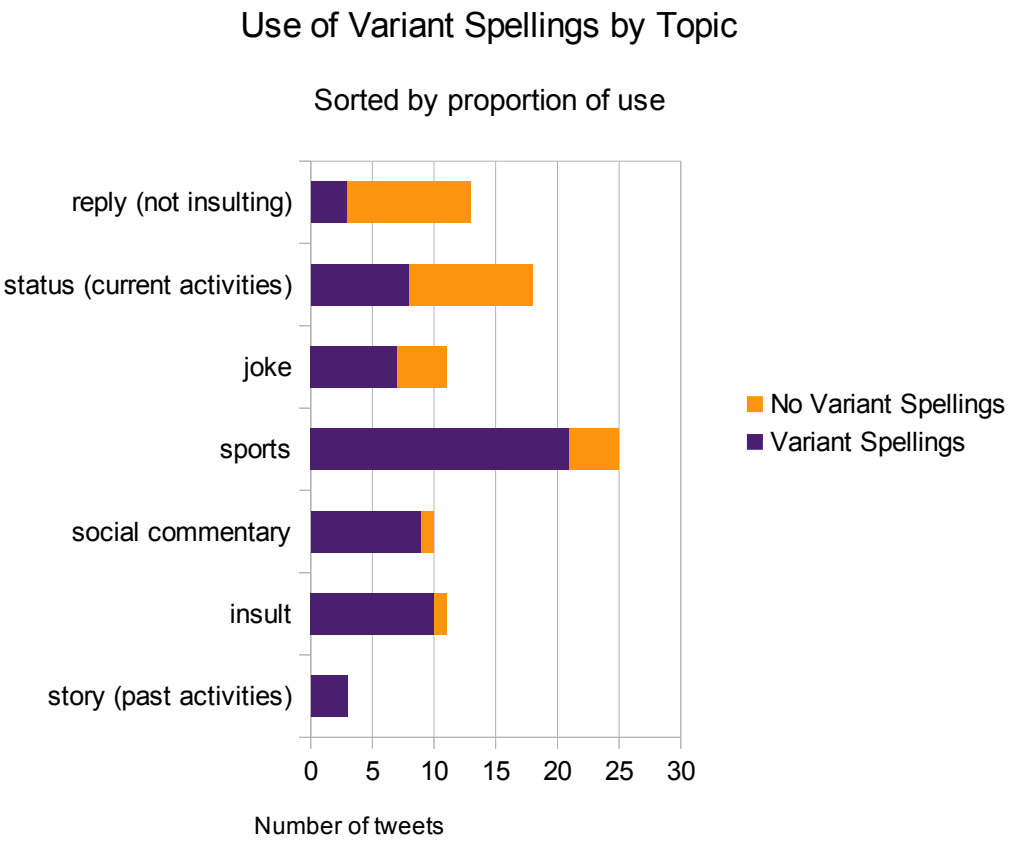
Conclusion: Style

- How does this interact with style?
- Case study: Twitter user BradleyKirkwood
  - https://twitter.com/BradleyKirkwood
  - 100 most recent tweets on April 23, 2015

Screenshot of BradleyKrikwood's Twitter Profile taken April 23, 2015.

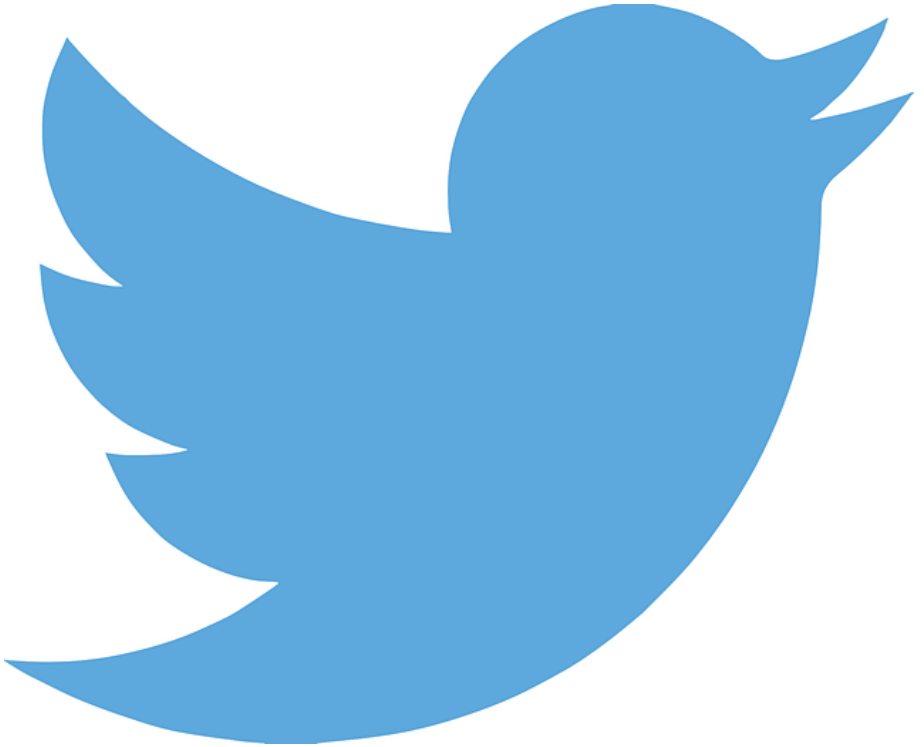
Conclusion: Style

- Tweets marked for use of variant spellings and topic by hand
- Most tweets used at least one variant spelling (64/100)
- Topic had a significant effect on variant spelling use
  - X^2 (6, N = 91) = 25.53, p <.001
- Use of variables shift with style
  - Sociolinguistic markers or stereotypes, not indicators (Labov 1972)
  - Performance registers? (Schilling-Estes 1998)
- Rich area for future research



Research Questions

- Can we apply variationist methods to Twitter data?
  - Yes
  - Method proposed here was applied to multiple dialects
- Do Twitter users use variant spellings to encode sociophonetic variation?
  - Yes, convergent evidence
- Do variant spellings interact with style?
  - Yes, area for future research



Thank you!

Rachael Tatman
The University of Washington

Contact: rctatman@uw.edu

Works Cited

Androutsopoulos, J. (2006). Introduction: Sociolinguistics and computer-mediated communication. Journal of Sociolinguistics, 10(4), 419-438.

bj [beccajamiesonx]. (2015, Apr 19). dae ye ever look back oan how much time ye wasted oan someone nd wonder why naeone punched u in the heed [Tweet]. Retrieved from https://twitter.com/beccajamiesonx/status/589927462903095296

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011, March). Mark my words!: linguistic style accommodation in social media. In Proceedings of the 20th international conference on World wide web (pp. 745-754). ACM.

Davies, M. (2011). Word frequency data from the Corpus of Contemporary American English (COCA).

Gentry, J., & Gentry, M. J. (2014). Package 'twitteR'.

Herring, S. C., & Paolillo, J. C. (2006). Gender and genre variation in weblogs. Journal of Sociolinguistics, 10(4), 439-459.

JPG [gencoj]. (2015, Feb 06). @OrdioMongo @___OJ___ hype hayed foah dat becawse it was 8 bucks foah 2 yeahs and w da jets i like readin about da kid prospects ogay [Tweet]. Retrieved from https://twitter.com/gencoj/status/563726906491957248

Labov, W. (1972). Sociolinguistic patterns. Philadelphia: University of Pennsylvania Press

Labov, W., Ash, S., & Boberg, C. (2005). The atlas of North American English: Phonetics, phonology and sound change. Walter de Gruyter.

Krapp, G. P. (1919). The pronunciation of standard English in America. Oxford University Press, American Branch.

Juhasz, B. J., Lai, Y. H., & Woodcock, M. L. (2014). A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. Behavior research methods, 1-16.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October). Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 37-44). ACM.

Rennie, S. (2001). The Electronic Scottish National Dictionary (eSND): Work in Progress. Literary and linguistic computing, 16(2), 153-160.

Sebba, M. (2003). Spelling rebellion. Pragmatics & beyond new series , 151-172.

Schilling-Estes, N. (1998). Investigating "self-conscious" speech: The performance register in Ocracoke English. Language in Society, 27(01), 53-83.

Siebenhaar, B. (2006). Code choice and code-switching in Swiss-German Internet Relay Chat rooms. Journal of Sociolinguistics, 10(4), 481-506.

Tagliamonte, S. A., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. American speech, 83(1), 3-34.

Stuart-Smith, J. (2004). Scottish English: phonology. A Handbook of Varieties of English, 1, 47-67.

Thurlow, C., & Brown, A. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. Discourse analysis online, 1(1), 30.

Tran, A. & Velema, S. (2014) Monster Spelunkin'. Monster Buddies Soundtrack.

Weide, R. L. (1998). The CMU pronouncing dictionary. URL: http://www. speech. cs. cmu. edu/cgibin/cmudict.

Code For Sampling Tweets

```
# add your own words here
words <- c("awn", "awr", "awll", "yawr", "awlso", "wawnt",
"becawse")
TwitterData <- NULL
# will return 100 English Tweets for each word
for(i in 1:length(words)){
  word <- searchTwitter(words[i], n=100, lang = "en")
  word.df = do.call("rbind",lapply(word,as.data.frame))
  TwitterData <- rbind(TwitterData, word.df)
}
# save out your data to analyze later
write.csv(TwitterData, "TwitterData.csv")
```

Code available at: https://github.com/rctatman/TwitterVarientSpellings

Words used for SAE study

- (for excluded because “fawr” is a foreign word)
- on
- or
- all
- your
- also
- want
- because

Words used for SSE study
(with frequency rank)

- “to” - 7, 9
- “you” - 14
- “do” - 18
- “who” - 38

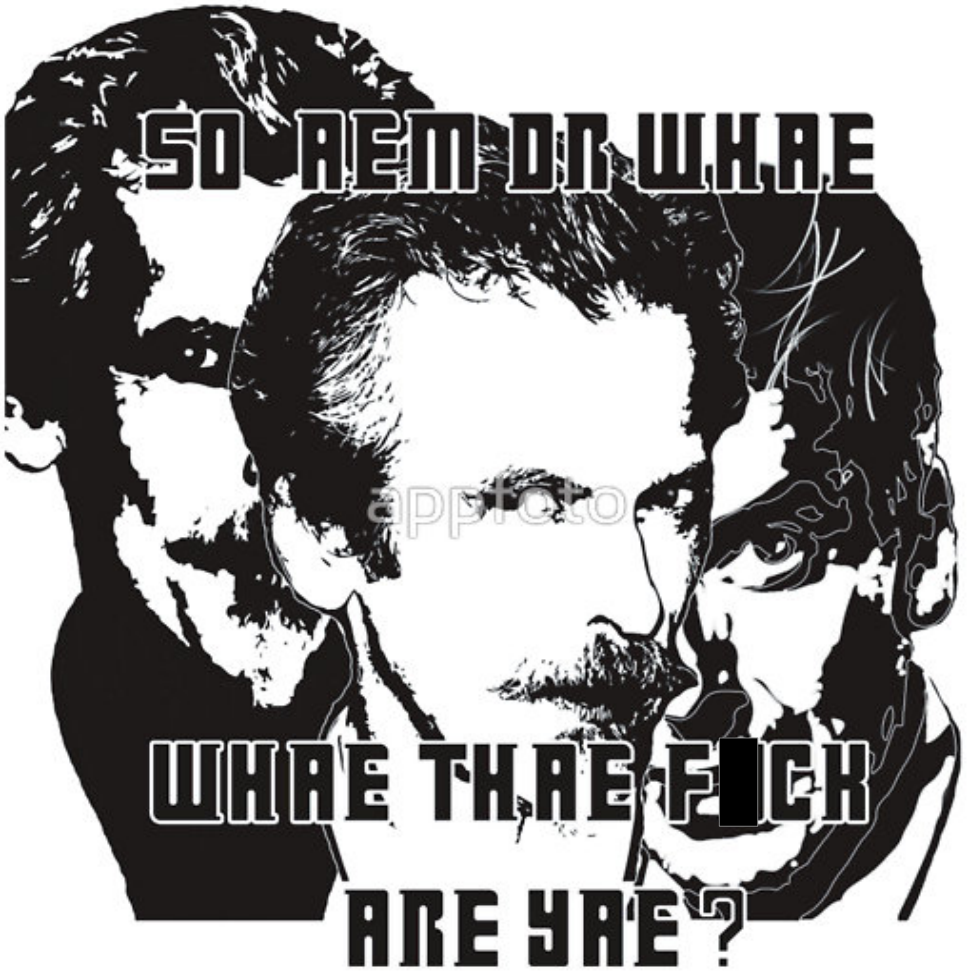


Image posted by Redbubble user appfoto. Retrieve April 23, 2015 from http://www.redbubble.com/people/appfoto/works/11160793-dr-who-whae-are-yae?p=sticker