The first decade of *because-NP*: 2007–2016

Justin Bland (The Ohio State University)
Kenneth Baclawski Jr. (University of California, Berkeley)
Matthias Raess (Ball State University)

Because-X

- Novel use of *because* to have non-CP or PP complements

(1) *But Iowa still wants to sell eggs to California, because money.* (Liberman 2012)

- Not simply deletion of *of* or use of *because* as a preposition (McCulloch 2014)

(2) a. *I'm gonna look for other schools this year, because :(!!* (Twitter)
b. *You've got to see this movie, because LOL.* (Twitter)

- Not limited to NP complements, despite the labels 'because-noun' and 'because-NP' (cf. 'because-X' in Bohmann 2016, Blamire 2017, a.o.)
  - We will use the label 'because-X' in this presentation, despite our title

Because-X

- *Because-X* is a marker of modern Internet slang, predominant in online forums (cf. Bland, Raess & Baclawski Jr. 2016)

- However, it coexists with a long history of *of-* or copula-deletion (Rehn 2015)

(3) a. *The wealthy, healthy, wise, famous and those favored by song, women and wine, all have, in individual instances, committed suicide because 'tired of life.'* (1898)
b. *Taboo connotes Greek ἅγος and ἅγιος, Latin sacer, holy or accursed because awesome.* (1918)

- Around 2011, it rapidly spread, ultimately being named the *WOTY* for 2013

Roadmap

1. Background on *because-X* and previous literature

2. Our previous results (Bland, Raess & Baclawski Jr. 2016)

3. Results from the Reddit and Twitter corpora
   - It arose in 2011, leveling off in mid-2012 to 2014 and persisting today
   - Reddit adopted *because-X* five or more months before Twitter
   - *Because-X* has a different character in Reddit (noun complements) and Twitter (interjections)
   - *Because reasons* has been and remains the most frequent *because-X*

4. Results from the social attitude survey
   - *Because-X* is linked to younger speakers and online media
   - It is not associated with gender or nationality

Previous literature

- Blog posts quickly noted the phenomenon and its general characteristics (Liberman 2012; Carey 2013, 2014; McCulloch 2014)

- Noted as the first non-lexical ADS Word of the Year (2013)

- Subsequent research has examined the syntax of *because-X*
  - Bailey (2014) on the syntactic distribution of *because-X* (247 participants)
  - Kanetani (2016) on the status of *because-X* complements as 'private expressions'
  - Blamire (2017) on *because-X* as a case-deletion phenomenon

- Some other studies have examined its distribution in online corpora
  - Schnoebelen (2014), Bohmann (2016)

Previous literature

- Schnoebelen (2014):
  - Twitter corpus (23,583 tokens of *because-X*, from one time slice)
  - *Because-X* is more prevalent among younger, female speakers in the US

- Bohmann (2016):
  - Twitter sample (12,751 tweets containing *because*, 803 tokens of *because-X*)
  - Does not find a correlation with colloquial, American, or computer-mediated speech
  - *Because-X* is used more in information-dense tweets (i.e. *of*-deletion)

- However, *because-X* is *rare*: 6.647/million words (Bland, Raess & Baclawski Jr. 2016)

- These studies do not investigate social meaning

- We need larger corpora and perception studies to further investigate the spread and social meaning of *because-X*

Our previous results

Bland, Raess & Baclawski Jr. (2016)

Compared Twitter, Reddit, and Wikipedia in order to investigate formality effects

- Twitter assumed to be less formal than Reddit
- Wikipedia used as a baseline

Results

- Evidence that *because-X* arose in 2011-2012
- *Because-X* used more on Twitter than on Reddit
- Examined other conjunctions like *although-X* and *unless-X*, but did not find that *because-X* was spreading to a more general CONJ-X

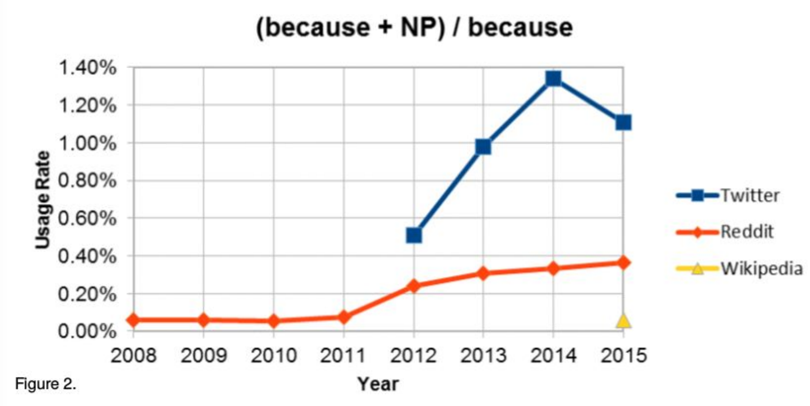


Figure 2.

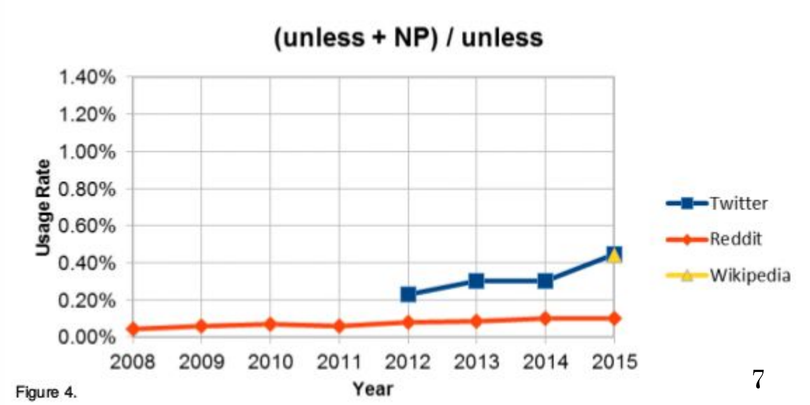


Figure 4.

Our previous results

Need for further investigation

- Get monthly sample instead of yearly sample for more fine-grained analysis over time
- Normalize using corpus size, not occurrences of *because*
- Investigate most popular *because-Xs* in each corpus
- Use a survey to investigate demographic and attitudinal data not available in the corpora

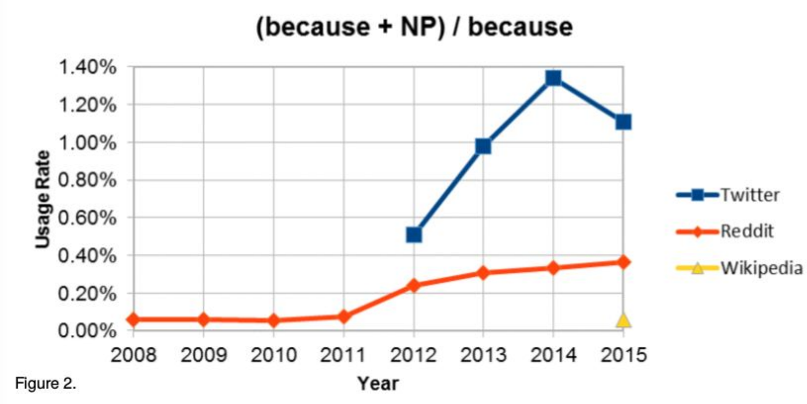


Figure 2.

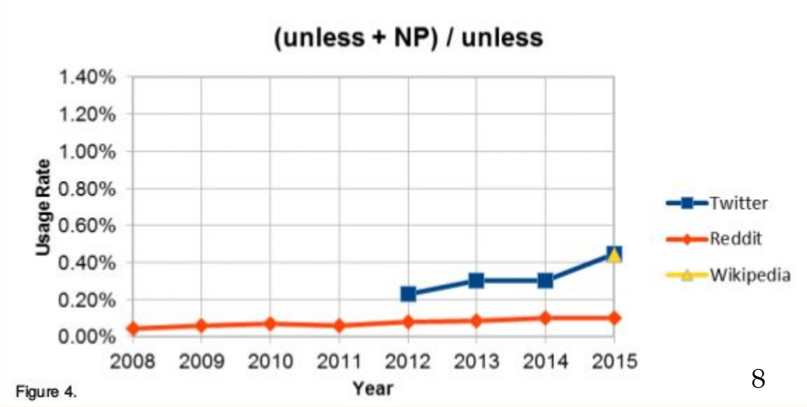


Figure 4.

Corpus data sources

Twitter Stream Grab corpus

- Approximately 1% of all publicly-available tweets since October 2011
- Used data from October 2011 to June 2016
- https://archive.org/details/twitterstream

Reddit Comments corpus

- 99.98% of all comments publicly posted to Reddit October 2007 to May 2015
- https://archive.org/details/2015_reddit_comments_corpus

Corpus filtering

Twitter

- Removed blank tweets.
- Removed native and naïve retweets.
- Removed tweets from shared accounts.
- Removed tweets from verified accounts.
- Removed tweets from users who had not set their language to English.
- Used Python's `guess_language` module to automatically detect tweet language; removed tweets that were not detected as English.
- Removed horoscope ads.

- Over **13 billion words** over 54 months
- Average of 243 million words per month

Reddit

- Used `guess_language` to automatically detect comment language; removed comments that were not detected as English.

- Over **47 billion words** over 92 months
- Average of 515 million words per month

Corpus analysis

Automatically tagged tweets/Reddit comments for part-of-speech

- ARK Twitter Part of Speech tagger (ver. 0.3) (Gimpel et al. 2011; Owoputi et al. 2012)
- Trained to handle non-standard orthography, lexis, syntax found on internet

Used script to automatically find tokens of *because-NP*, defined as a sequence of:

The word *because* tagged as P (prep. or subordinating conj.)

+

An NP

- One of the tag sequences: N, NN, DN, AN, DAN, ANN, AAN, ^, ^N, N^, ^^, A^, D^, DA^
- Screened out pronoun + verb contractions frequently mis-tagged as D (e.g. *they're, I'ma*)

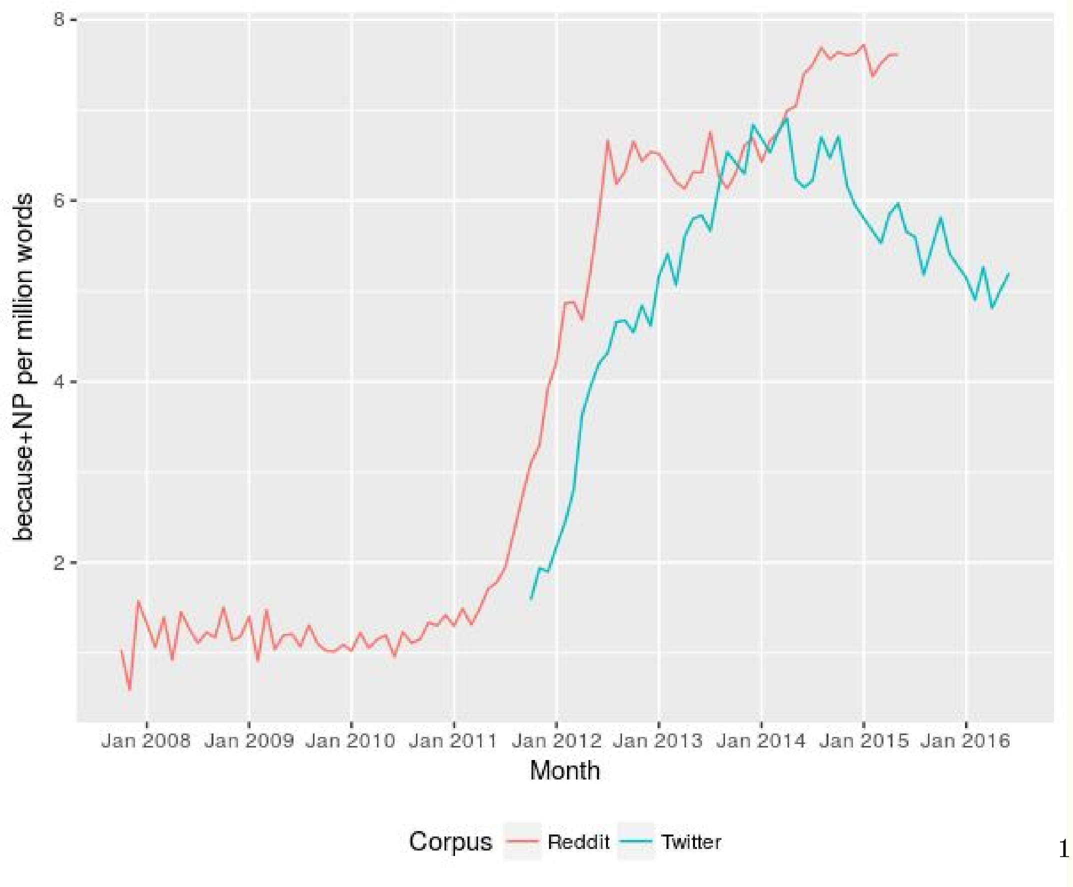
+

End of tweet/comment or clause-final punctuation

- One or more of ? ! . ;

Corpus results

- Confirmed *because-X* arose in 2011-2012
- Twitter and Reddit have similar maximum rates of *because-X* (contra our previous results)
- Reddit seems to have adopted *because-X* 5 or more months earlier than Twitter
- *Because-X* has persisted over time, but may be declining slightly
- Confirmed there was no larger CONJ-X phenomenon, e.g. *unless-X*, *although-NP*



Corpus results

Separate linear regression models for effect of month on monthly usage rate:

Twitter

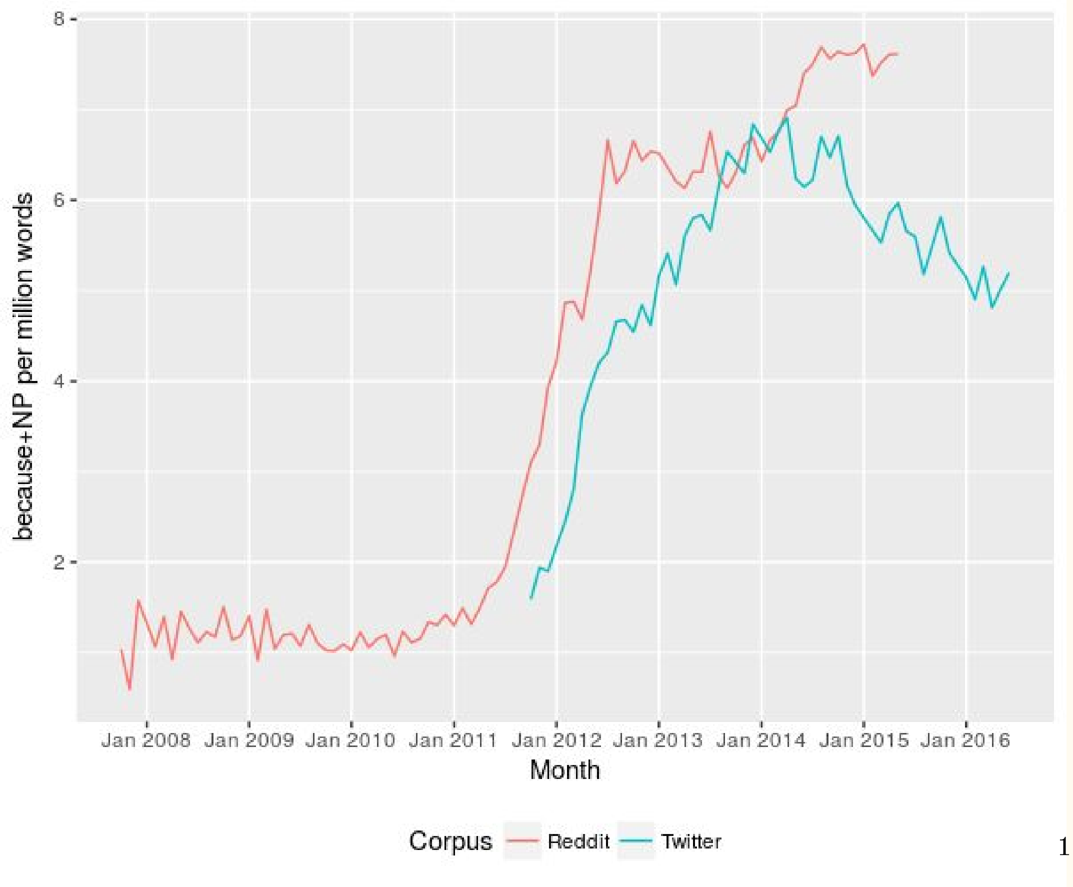
- Month* ($p = 8.55$e-06)

Reddit

- Month* ($p <$ 2e-16)

Multiple linear regression model for effect of corpus and month on monthly usage rate, only for months where data is available for both corpora:

- Month* ($p <$ 2e-16)
- Corpus* ($p = 3.24$e-10)



Corpus study: Most common Xs

- *because reasons* has a top position in both corpora, confirming its use as the most common *because-X*
- Abbrevs and interjections preferred on Twitter
- Bare nouns preferred on Reddit
- Nouns on Twitter reference life situations and tastes; nouns on Reddit are more topical
- Xs used with *because-X* are often hashtag-like

	Twitter		**Reddit**	
1.	because yolo	2933	because reasons	13526
2.	because reasons	1050	because money	3743
3.	because lol	943	because boobs	3299
4.	because yes	644	because science	2753
5.	because yeah	613	because reddit	1593
6.	because school	501	because jesus	1412
7.	because life	482	because patriarchy	1395
8.	because no	390	because hey	1372
9.	because wow	331	because freedom	1345
10.	because damn	298	because god	1303
11.	because college	249	because yolo	1098
12.	because work	245	because internet	1047
13.	because duh	237	because yes	1037
14.	because food	236	because america	991
15.	because swag	233	because sex	958

Survey design

- Survey constructed using Qualtrics, distributed with Amazon Mechanical Turk
  - Native speakers of English from the US were recruited

- Participants were asked a variety of questions (following the survey)

- Demographic questions
  - Age, gender, state in the US, education, and others

- Internet usage questions (self-reported)
  - "Which social media sites do you visit/belong to?" (FaceBook, Twitter, Wikipedia, etc.)
  - "Which social media sites do you actively post to on a regular basis?"
  - Among others not discussed here (e.g. "How often do you check your social media?")

Survey design

- 118 participants (165 total, 45 did not complete the survey or failed gatekeeper tasks)
- 55 self-identified as female, 63 as male (participants given an open-ended prompt)
- Median age range: 26-35
- Median education completed: "Some college"
- Largely in line with typical demographics reported for MTurk (Ipeirotis 2010)

Survey design

- Participants were shown a sentence, then given sliding-scale prompts:

1. How likely is it that you would say this sentence? (1-100)
2. How likely is it that you would hear or read this sentence? (1-100)
3. Picture somebody saying this sentence. How old are they? (Young-Old)
4. … What is their gender? (Female-Male)
5. … Where are they from? (US-Abroad)
6. … Are they writing online or speaking in person? (Online-In person)

Survey design

- Participants were randomly assigned because-X or its non-elided counterpart (Interjections were included, but not reported here)

- Most examples were adapted from Twitter (Provenance does not have a significant effect)

(4) a. *2008 was an exciting year* ***because Obama.***
b. *2008 was an exciting year* ***because of Obama.***

(5) a. *I fell out of my chair at the movie,* ***because laughing so hard.***
b. *I fell out of my chair at the movie,* ***because I was laughing so hard.***

- 10 NP complements, with 5 instances of "because reasons" (varied weight/# of words)

- 5 VP, 5 adjective complements (cf. Bailey 2014)

Survey results: Controls

- Results from controls indicates that the survey design was successful

(6) *I can't go see the movie, because is stay are tonight parent my here.*

  - How likely is it that you would say this sentence? **(Median = 0/100)**
  - How likely is it that you would hear or read this sentence? **(Median = 0/100)**

- However, many were unwilling to admit using slang or Internet vocabulary

(7) *I'm going to the party tonight, because YOLO.*

  - How likely is it that you would say this sentence? **(Median = 1/100)**
  - How likely is it that you would hear or read this sentence? **(Median = 62.5/100)**

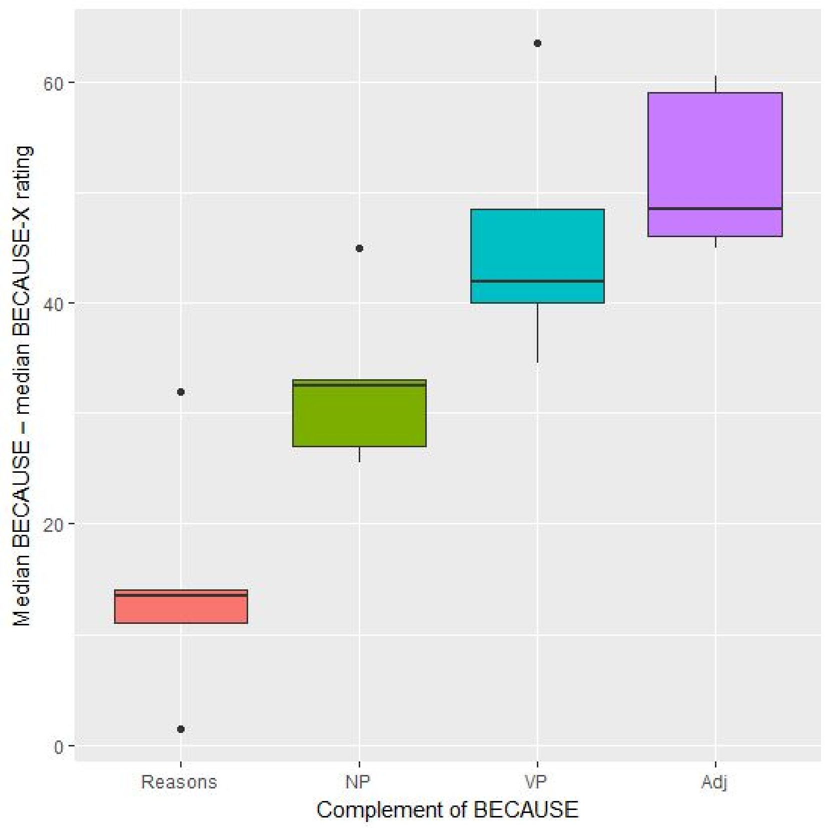
- We will focus on the "hear or read" prompt

Survey results: Syntactic distribution

- A score was given to the 20 prompts:
  (Median *because* rating – median *because-X* rating)

- Lower scores = more acceptable *because-X*

- Results:
  - Because-reasons stands out
    Category: Reasons ($\chi^2 = 16.75$, $p < 0.001$)
  - Other NP's are also highly rated
    Category: NP significant ($\chi^2 = 17.1$, $p < 0.001$)
  - VP's and Adj's are the lowest rated

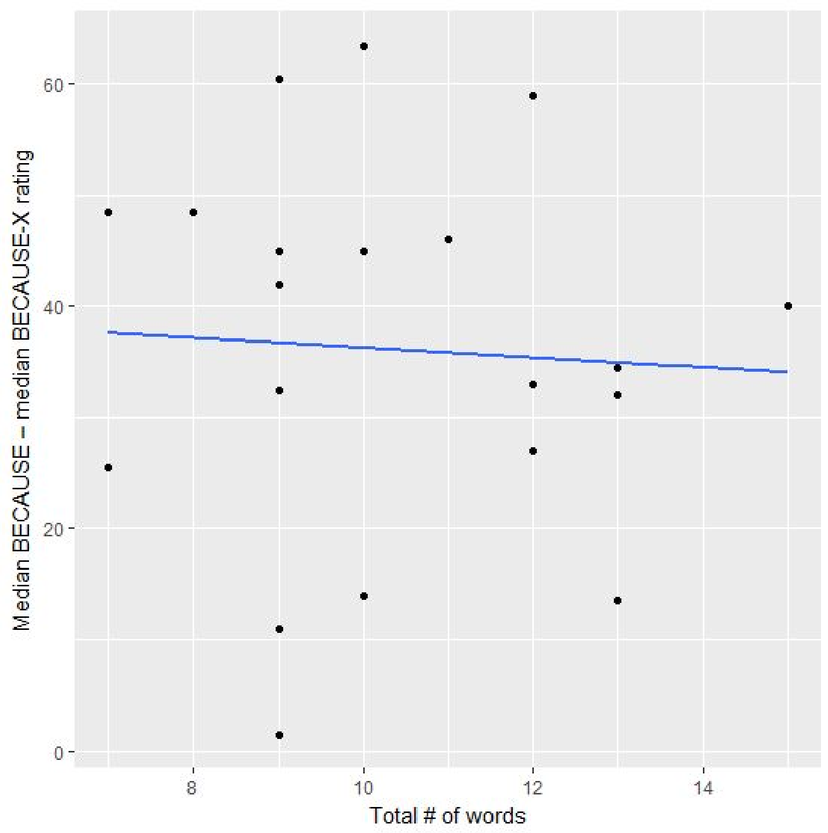


Survey results: Syntactic distribution

- Number of words in the sentence or complement of *because* do not significantly affect the results

- Test: Linear mixed effects model
  - Random effect for provenance of prompt
  - Likelihood ratio tests to find significance (lme4, ANOVA in R)
  - **# of total words, # of words in complement** n.s.

- No effect of information density

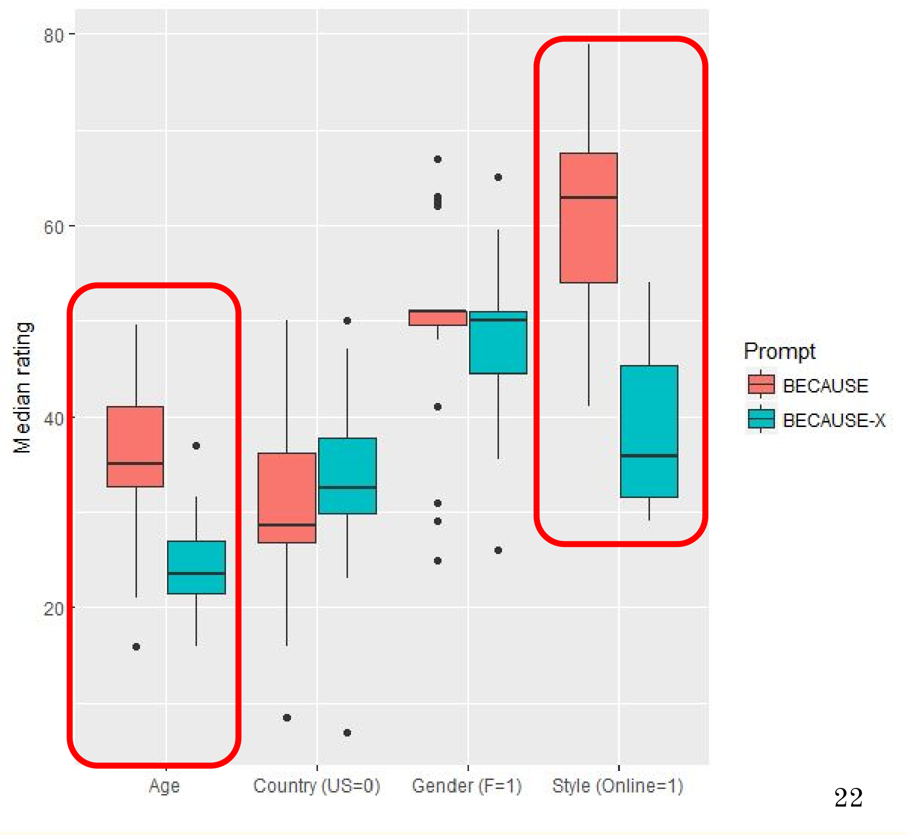


Survey results: Perception of *because-X*

- The median rating of perceived age, country, gender, and style was calculated for each prompt

- Rating of *because-X* correlates with perceived age and style

- Test: t-tests
  - Age: Younger speaker ($t = 6.04$, $p < 0.05$)
  - Style: Online ($t = 7.29$, $p < 0.05$)
  - Gender, Country (n.s.)



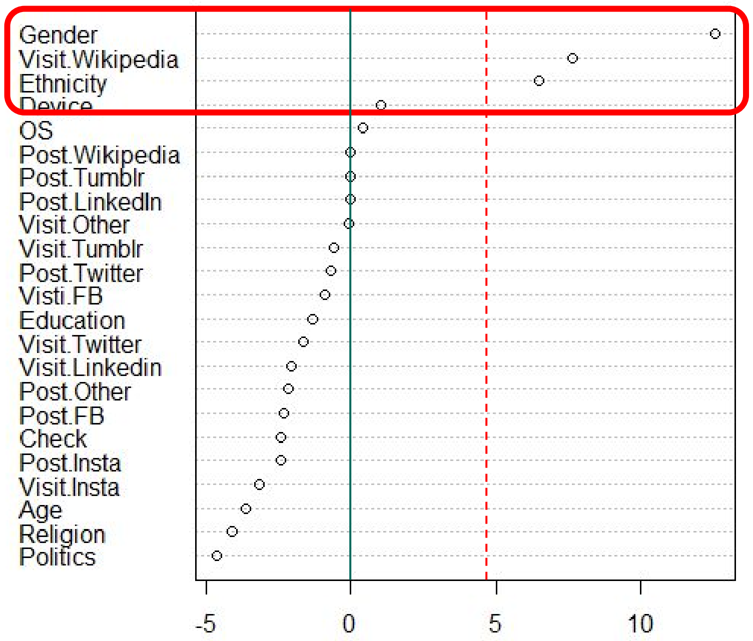
Survey results: Participant questionnaire

- The median rating difference was calculated for each participant
    - (Median *because* – median *because-X*)

- Random forests were run because of the high number of predictors and likely multicollinearity

  (cf. Tagliamonte & Baayen 2012, Shih 2011)

- Gender, Ethnicity, and Visit.Wikipedia stand out as the most likely predictors
    - "Which social media sites do you visit/belong to?" as opposed to:
    - "Which social media sites do you actively post to on a regular basis?"

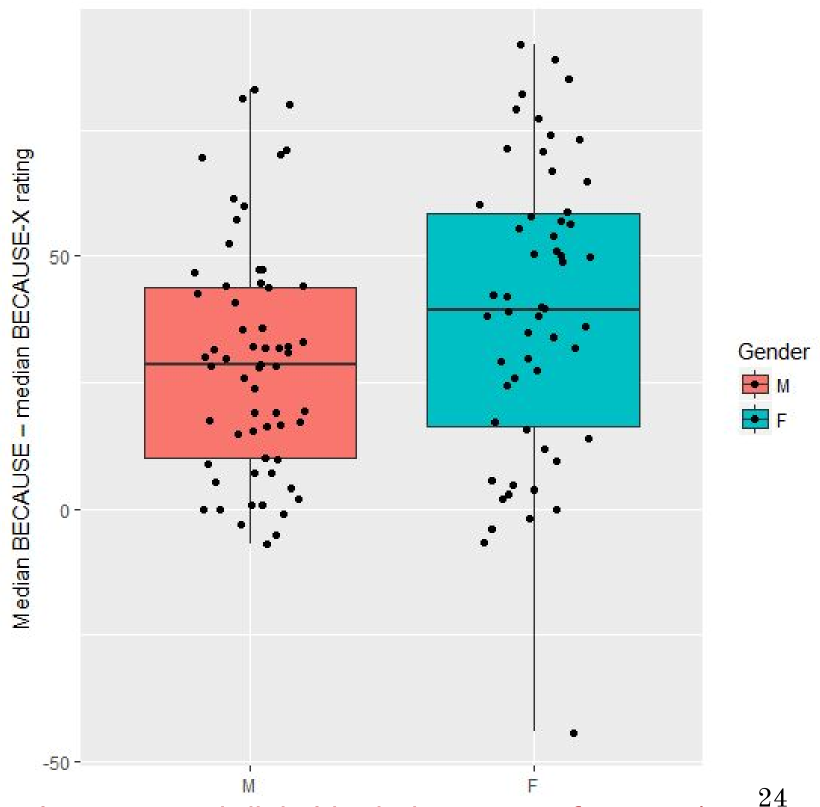


Survey results: Participant questionnaire

- Test: Linear regression
  - Ethnicity (n.s.)
  - Interactions (n.s.)

  - Gender significant, such that Gender:Male is correlated with lower diff. for *because-X* (β = -10.9, $p < 0.05$)

- Why might male participants rate *because-X* higher?
  - A tentative hypothesis: the male-dominance of Reddit users and content
  - A Pew Research Center poll finds 71% of Reddit users to be male (2016)*



*http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

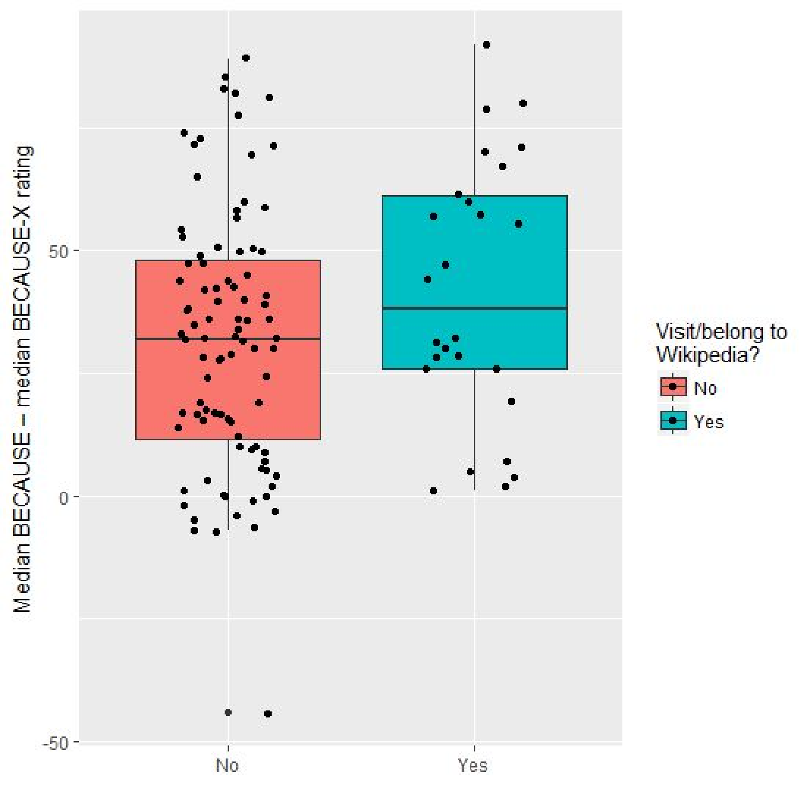
Survey results: Participant questionnaire

- Test: Linear regression
  - Ethnicity (n.s.)
  - Interactions (n.s.)

  - Gender significant, such that Gender:Male is correlated with lower diff. for *because-X* ($\beta = $ -10.9, $p < 0.05$)

  - Visit.Wikipedia significant, such that those who visit Wikipedia have higher diff.'s ($\beta = 12$, $p < 0.05$)

- Why Wikipedia?
  - Because prescriptivism?



Conclusions

- Corpus:
  - *Because-X* arose on Reddit in early 2011, followed by Twitter shortly after
  - The overall character of *because-X* seems to be different in Reddit and Twitter
  - *Because reasons* remains the *because-X* par excellence
- Survey:
  - *Because-X* is associated with younger speakers and online media, but not gender or nationality
- Future research:
  - More targeted research on the interaction between *because-X* and gender
  - More explanation of the differences between *because-X* on Reddit, Twitter, and elsewhere
  - Closer analysis of spread and social meaning in smaller online communities

References

Bailey, L. (2014). "'Because X: Syntactic restructuring, ellipsis, or 'internetese'?". LAGB 2014, 04/09/2014, University of Oxford.

Blamire, E. (2017). "A syntactic analysis of because x in English... because linguistics!" Presentation at the *Canadian Linguistic Association Annual General Meeting*, Toronto, ON. 2017.

Bland, J., Raess M., and Baclawski Jr., K. (2016). Because formality: The conjunction-noun construction in online text corpora. Poster presented at the American Dialect Society Annual Meeting, Washington, DC.

Bohmann, A. (2016). "Language change because Twitter? Factors motivating innovative uses of because in the English-speaking Twittersphere." In, L. Squires (ed.) *English in Computer-Mediated Communication*. De Gruyter.

Carey, S. (2013). "'Because' has become a preposition, because grammar." Blog post. Sentence first: An Irishman's blog about the English language. November 13, 2013. Accessed July 17, 2015. https://web.archive.org/web/20150707174851/https://stancarey.wordpress.com/2013/11/13/because-has-become-a-preposition-because-grammar/

Carey, S. (2014). "'Because' is the 2013 Word of the Year, because woo! Such win." Blog post. Sentence first: An Irishman's blog about the English language. January 4, 2014. Accessed July 17, 2015. https://web.archive.org/web/20150522082051/https://stancarey.wordpress.com/2014/01/04/because-is-the-2013-word-of-the-year-because-woo-such-win/

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Portland, OR. Companion volume.

Ipeirotis, P. (2010). "Demographics of Mechanical Turk." *NYU Working Paper No. CEDER-10-01*. Available at SSRN: https://ssrn.com/abstract=1585030

Kanetani, M. (2016) "A Note on the *Because* X Construction: With Special Reference to the X-Element." *Studies in Language and Literature* [Language] 70: 67-79.

Liberman, M. (2012). "Because NOUN." Blog post. Language Log. July 12, 2012. Accessed July 17, 2015. https://web.archive.org/web/20150317182710/http://languagelog.ldc.upenn.edu/nll/?p=4068

McCulloch, G. (2014). "Why the new "because" isn't a preposition (but is actually cooler)." Blog post. All Things Linguistic. January 4, 2014. Accessed July 17, 2015. https://web.archive.org/web/20150319210532/http://allthingslinguistic.com/post/ 72252671648/why-the-new-because-isnt-a-preposition-but-is

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., and Schneider, N. (2012). Part-of-speech tagging for Twitter: Word clusters and other advances. Technical report, Machine Learning Department, Carnegie Mellon University. CMI-ML-12-107.

Rehn, A. (2015). "Because Meaning: Language Change through Iconicity in Internet Speak." 2014 SURF Conference Proceedings. University of California, Berkeley: Summer Undergraduate Research Fellowships. https://escholarship.org/uc/item/0r44d2bh

Schnoebelen, T. (2014). "Innovating because innovation." *Idibon*. Accessed at https://corplinguistics.wordpress.com/2014/01/15/innovating-because-innovation/

Reddit Comments [text corpus]. (2007-2015). Accessed 2017. https://archive.org/details/2015_reddit_comments_corpus

Tagliamonte, S. and Baayen, H. (2012). "Models, forests and trees of York English: Was/were variation as a case study for statistical practice." *Language Variation and*

Change 24: 135-178.

Twitter Stream Grab [text corpus]. (2011-2016). Accessed 2017. https://archive.org/details/twitterstream

Acknowledgements

Many thanks to the following people for their helpful comments, suggestions, and advice: Sravana Reddy and the students in Language Variation through the Lens of Social Media at the 2015 Linguistic Institute, the audience at American Dialect Society 2016, Marie-Catherine de Marneffe, Kathryn Campbell-Kibler, and Bodo Winter.

Thanks!

Justin Bland (bland.97@osu.edu)

Kenneth Baclawski Jr. (kbaclawski@berkeley.edu)

Matthias Raess (mraess@bsu.edu)

Appendix: Interjections & emojis

- Interjections and emojis were among the highest rated *because-X* prompts in our survey (highest med. rating of because+Reasons/NP: 59.5/100)

(9) a. *You've got to see this movie, because LOL.* (Med. rating: 60/100)

b. *She's working overtime this week, because $$$.* (Med. rating: 63/100)

Appendix: Predictors of other ratings

- We ran random forest analyses for ratings of the perceived Age, Gender, Country, and Style of the prompts
  - Each participant was given scores for their median ratings of perceived Age/Gender/Country/Style for *because* prompts and *because-X* prompts
  - Participants ended up with four difference scores:
    Age difference rating = Median Age rating for *because* – median Age rating for *because-X*,
- **Age difference:** Age of the participant is a significant predictor, such that older participants rated *because-X* speakers to be younger ($p < 0.05$, Est. $= 2.1$)
- **Gender difference:** Gender of the participant is a sig. predictor, such that participants rated *because-X* speakers to be of their own gender ($p < 0.001$, Est. $= 10.95$)

Appendix: Predictors of other ratings

- We ran random forest analyses for ratings of the perceived Age, Gender, Country, and Style of the prompts
  - Each participant was given scores for their median ratings of perceived Age/Gender/Country/Style for *because* prompts and *because-X* prompts
  - Participants ended up with four difference scores:
    Age difference rating = Median Age rating for *because* – median Age rating for *because-X*,
- **Country difference:** Post.FB is a significant predictor, such that participants who report frequently posting on FaceBook rate *because-X* speakers to be more foreign ($p < 0.05$, Est. = –10.27)
- **Style difference:** Age, Post.FB, and Visit.Wikipedia were all sig. predictors:
  - Older participants rate *because-X* as more online ($p < 0.05$, Est. = +4.8)
  - Participants who post on FaceBook rate *because-X* as less online ($p < 0.05$, Est. = –11.67)
  - Participants who visit Wikipedia also rate *because-X* as less online ($p < 0.05$, Est. = –13.1)

Appendix: Survey results: Comments

- Our participants seem to be of two populations:
  - Those who interpret *because-X* as an Internet phenomenon
  - Those who interpret *because-X* as *of*-/copula-deletion (i.e. those who visit Wikipedia)

- This is borne out in comments
  - Participants were asked to give an optional comment after each prompt

- Internet speech
  - "It sounds like a meme", "It sounds a little like internet meme speak"
  - "I could imagine seeing this on 4chan"

- *Of*-deletion
  - "Since improper English, I would guess that a foreigner would say it"
  - "Maybe something someone would say in a rush"
  - "It would have to be a child, someone who doesn't speak the language very well or maybe someone who got cut off before they could finish whatever they were about to say"