Because formality: The conjunction-noun construction in online text corpora[1]

Justin Bland (Virginia Tech), Matthias Raess (Ball State University)
& Kenneth Baclawski Jr (University of California, Berkeley)
<jebland@vt.edu>, <mraess@bsu.edu>, <kbaclawski@berkeley.edu>

1. Introduction

- *Because-noun* (1) is a shibboleth of colloquial Internet speech (Liberman 2012; Carey 2013, 2014; Bailey 2014; *WOTY* 2013).

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

  - Broadly defined as a prepositional use of *because* in utterance-final position, typically denoting superordinate topic.
  - Attested since at least 2010, but cf. Rehn (2015).

- Similar constructions have been reported with other conjunctions (2).

(2) *I didn't want to talk out loud, **thus text messaging**.* (McCulloch 2014)

  - "[S]ubordinating conjunctions as a class are appearing in a new type of construction" (McCulloch 2014).

Research questions

- **Question #1**: What is the origin of *because-noun*?
- **Question #2**: Is *because-noun* part of a broader, emerging *conjunction-noun* construction?

2. Background

The need for massive corpora

- While *because* is highly frequent, other comparable conjunctions are much less so, and *conjunction-noun* tokens are a small subset of these (Table 1; cf. Biber et al. 1998).

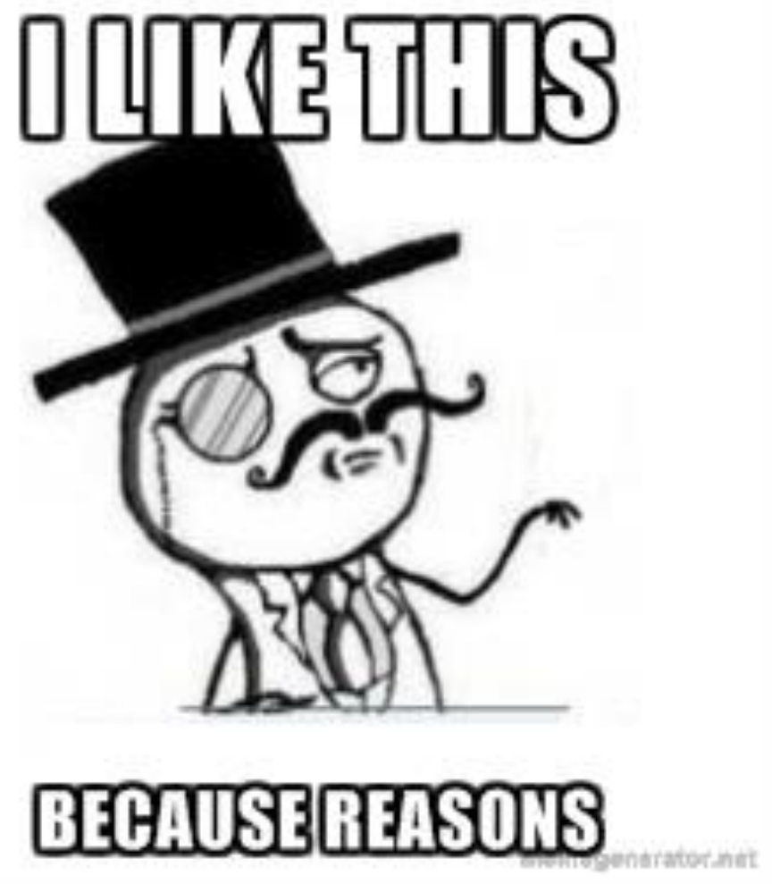
Conjunction	CONJ/million words	(CONJ+NP)/million words
because	1,985.492	6.647
although	92.605	0.084
unless	175.183	0.174

Table 1. Raw frequency of CONJ and CONJ+NP in Reddit Corpus, Jan.-Jun. 2014

- Diachronic corpora will allow us to investigate the history of the construction.
- Thus massive online corpora.

The need for a range of formality

- *Because-noun* is bound up with formality.
  - Clearly innovative, informal, "fashionably slangy" (Carey 2013)
  - (#*because money*) in formal speech
- A range of formality is desirable.
  - **Twitter**: extremely informal, stream-of-consciousness (Russell 2013)[2]
  - **Reddit**: informal, conversational
  - **Wikipedia**: formal, academic English
- General assumption: Wikipedia >> Reddit >> Twitter for formality effects.



http://cdn.meme.am/instances/400x/36066874.jpg

3a. Methodology

Set of subordinating conjunctions

- For this study, only subordinating conjunctions not ambiguous with other uses (e.g. preposition, adverbial) were analyzed (cf. Quirk et al 1985).
  - ***albeit***
  - ***although*** (incl. variant spelling ***altho***)
  - ***because*** (incl. variant spellings ***bc, b/c, cuz, cus, coz, cos***)
  - ***lest***
  - ***unless*** (incl. variant spelling ***unles***)
  - ***whereas***
  - ***whereupon***
- Other conjunctions (e.g. *as*, *if*) will be left for future research.

3b. Methodology, cont'd.

Corpus data sources

- **Twitter**: sample of 3 months of tweets per year from the "spritzer" stream, 2012-2015, filtered to remove retweets, tweets from potential public figures or corporations, and tweets flagged as non-English (Internet Archive's Twitter Stream Grab).
- **Reddit**: sample of 6 months of comments per year (5 months in 2015), 2008-2015 (Internet Archive's Reddit Corpus).
- **Wikipedia**: systematic sample of 5% of Main namespace articles from 2015 (English Wikipedia dump).[3]

Data processing

- Corpora were automatically searched to find text containing a target conjunction.
- Text containing a target conjunction was automatically tagged for parts of speech (POS) using the ARK Twitter Part-of-Speech tagger, version 0.3 (Gimpel, et al. 2011; Owoputi, et al. 2012).
  - The ARK tagger was designed to handle Twitter-specific typography (e.g. #hashtag, @username) and was trained on Twitter data so it can handle the non-standard orthography, lexical items, and syntax commonly found on the internet.
- Search algorithm
  - Search terms were fashioned to minimize false positives, while targeting conjunction-noun sequences.
  - Algorithm identified tokens consisting of conjunctions followed by noun phrase elements then non-continuing punctuation (Schema 3).

(3)

{CONJ	**NP**	**PUNCT}**
{*because*, …}	{N, DN, AN, … }	{one or more of ?, !, ., ;}

  - Conjunctions were required to be tagged as P (subordinating conjunction or preposition) to avoid mis-identifying homonyms (e.g. *cuz* as an abbreviation of *cousin*).
  - The following NPs were considered: **N, NN, DN, AN, DAN, ANN, AAN, ^, ^N, N^, ^^, A^, D^, DA^** (N=common noun, ^=proper noun, D=determiner, A=adjective).
  - PRON+V contractions frequently mis-tagged as D (e.g. *they're*, *it's*, *I'ma*) were not accepted as part of NPs.[4]

- Precision: 91.2% (on Reddit Corpus), 89.8% (on Twitter Corpus)
  - Calculated from a hand-checked random sample of 300 tokens each from the Reddit and Twitter 2015 corpora.
  - Note: Recall was not tested at this time, so there are likely false negatives in the data set.

(4)

Accepted	**Blocked**
1 = *… because reasons.*	0 = *… because I've reasons.*
1 = *… because the red car!*	0 = *… because cars, I like.*
1 = *… because Barack Obama?*	0! = *… because reasons, right?*

4a. Results

Question #1: What is the origin of *because-noun*?

- Overall corpus sizes and usage rates:

Corpus	CONJ	CONJ+NP	(CONJ+NP)/CONJ
Twitter	4,106,450	40,983	0.998%
Reddit	63,742,859	180,765	0.284%
Wikipedia	59,032	68	0.115%

Table 2. Overall token counts, all years

- Relative frequency of (CONJ+NP)/CONJ by year:

Corpus	2008	2009	2010	2011	2012	2013	2014	2015
Twitter	–	–	–	–	0.50%	0.94%	1.29%	1.07%
Reddit	0.07%	0.07%	0.06%	0.08%	0.23%	0.29%	0.31%	0.34%
Wikipedia	–	–	–	–	–	–	–	0.12%

Table 3. (CONJ+NP)/CONJ frequencies by year

4b. Results, cont'd.

- Usage rates by year for CONJ+NP and *because*+NP:

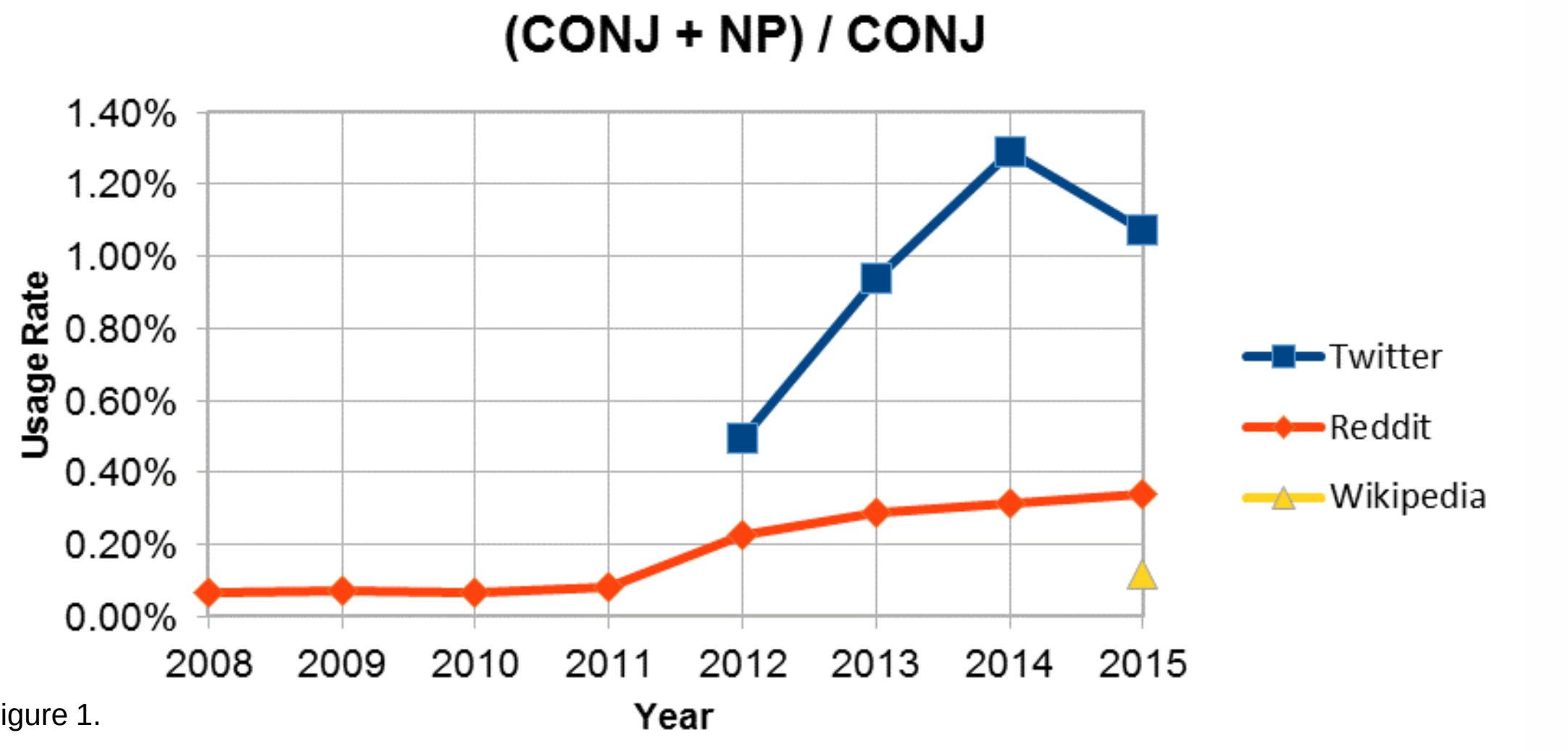


Figure 1.

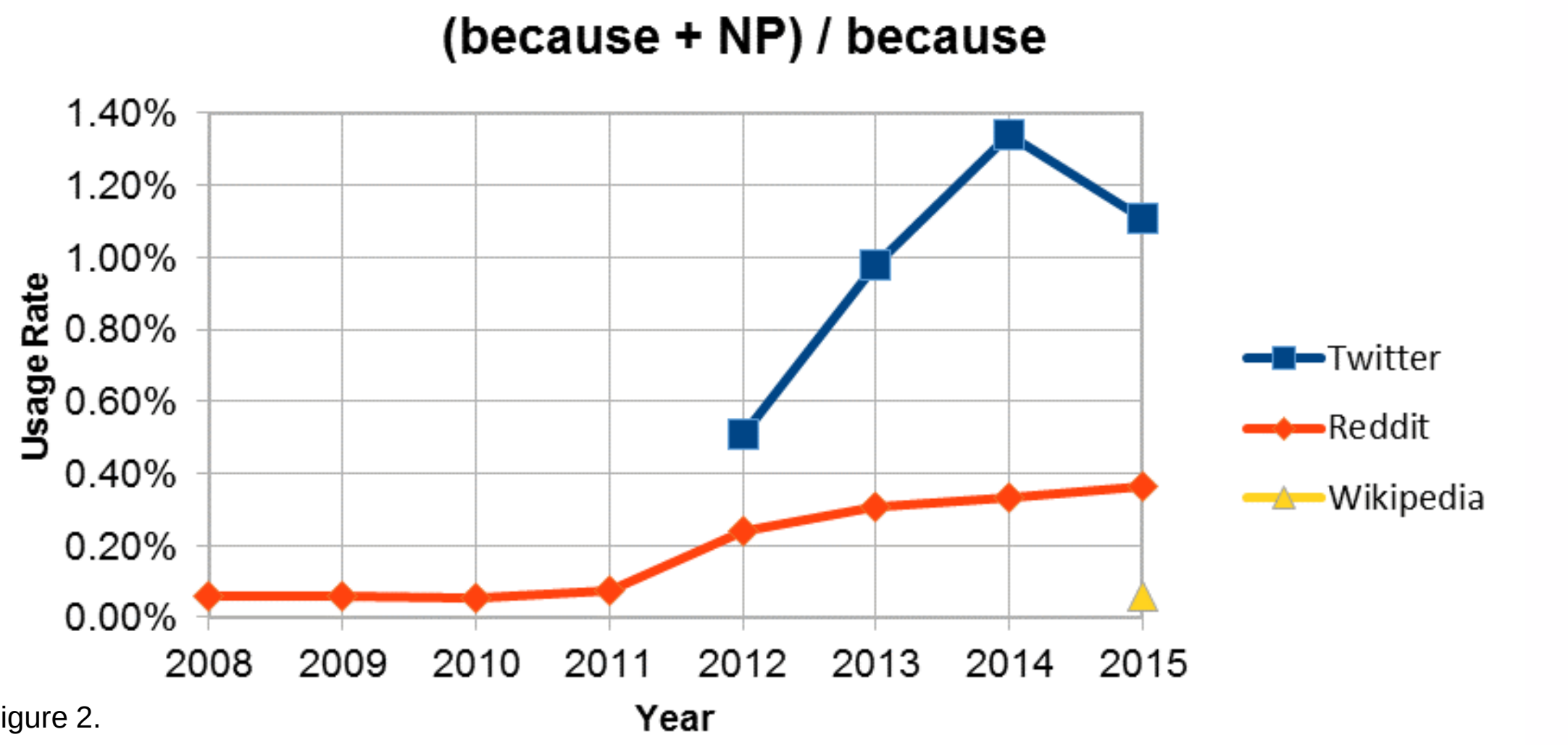


Figure 2.

Question #2: Is because-noun part of a broader, emerging conjunction-noun construction?

- Usage rates by year for *although*+NP and *unless*+NP:

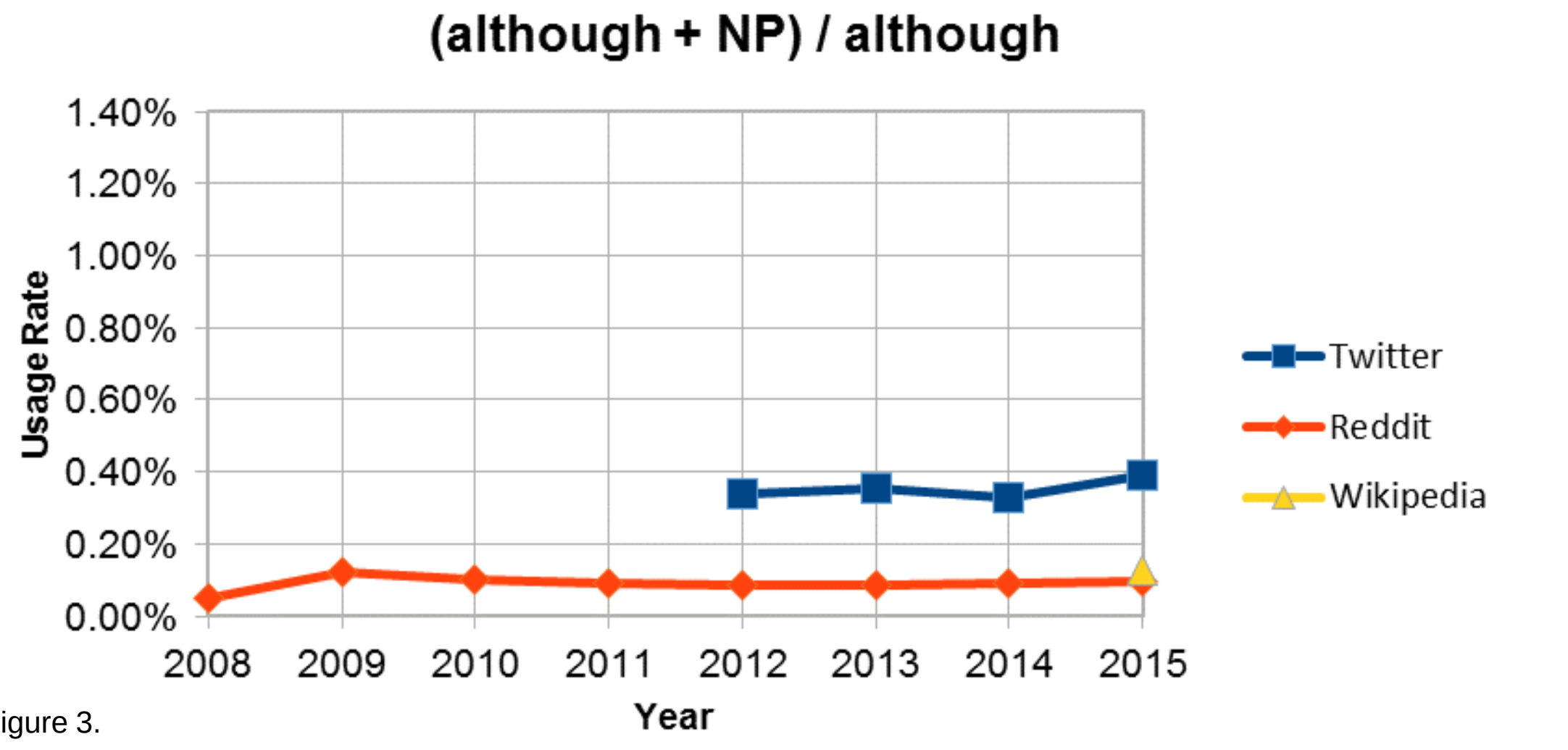


Figure 3.

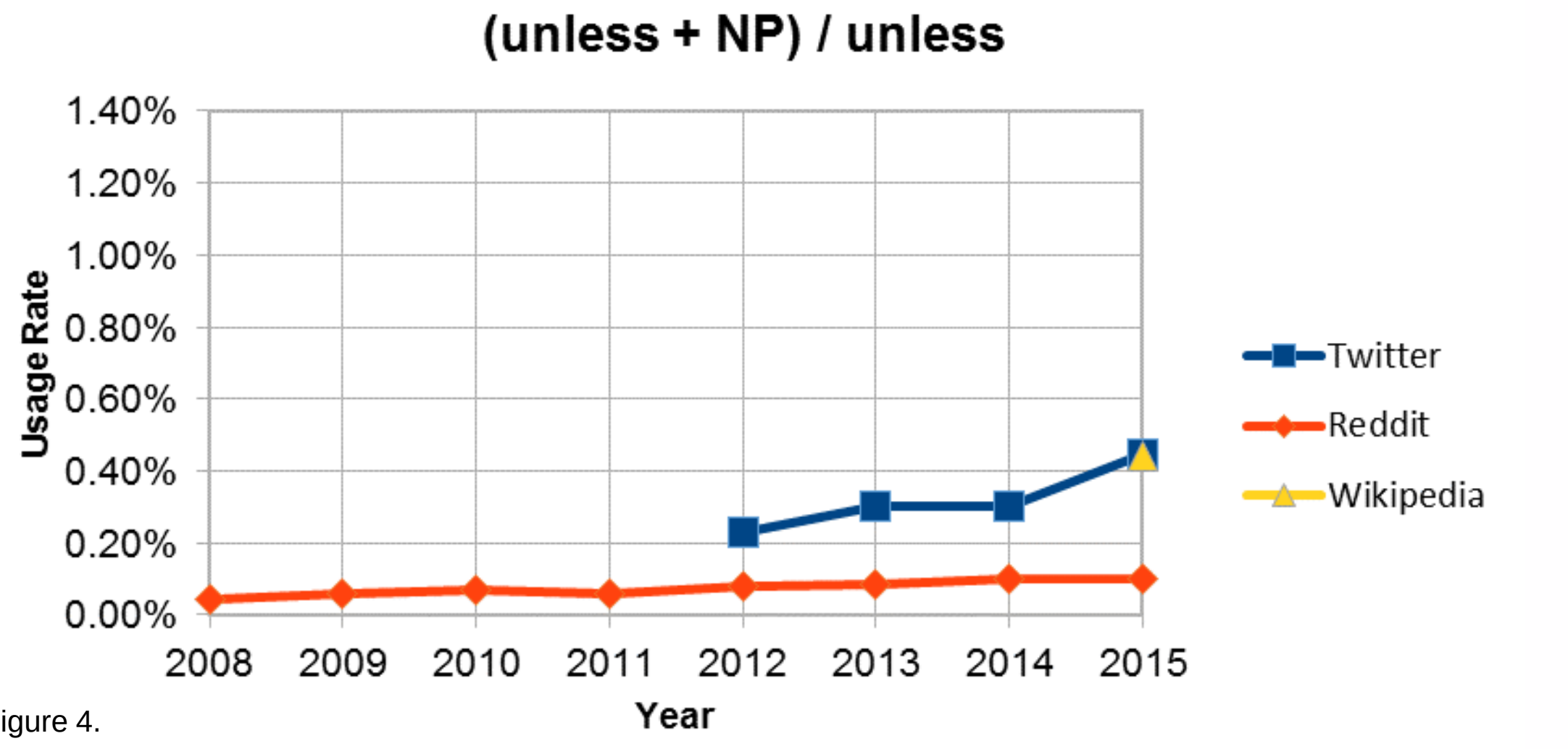


Figure 4.

- Percent of CONJ+NP constructions that involve *because*:

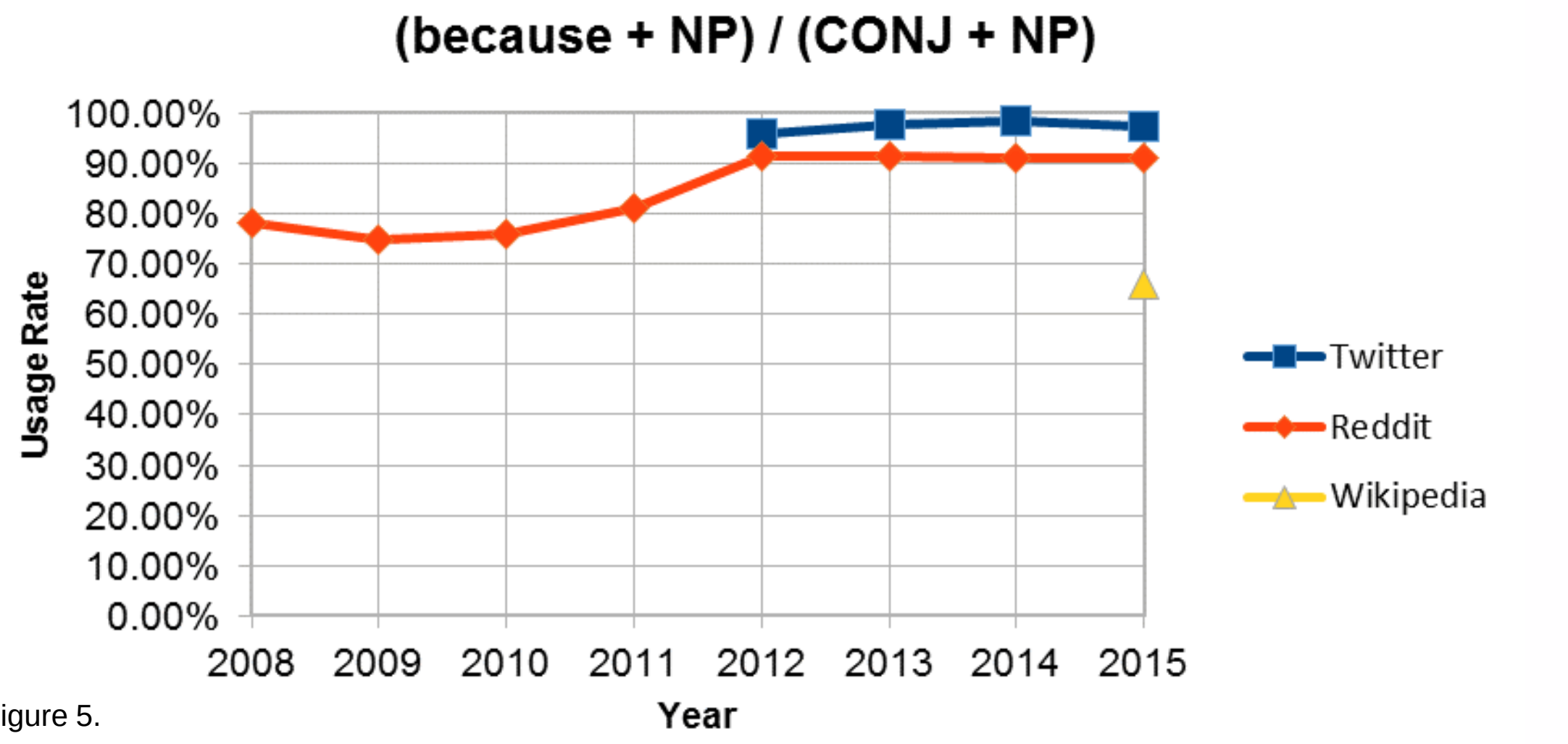


Figure 5.

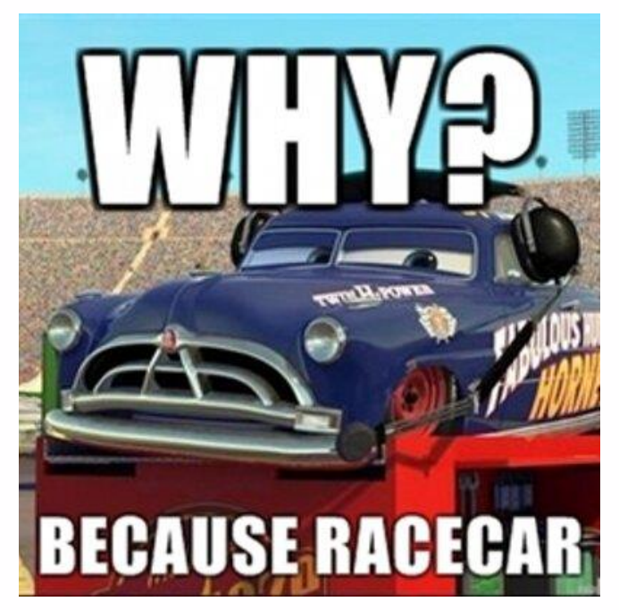
5. Conclusions (*cuz tl;dr*)

- Statistical analysis
  - Statistical test: ANOVA (response variable: relative frequency of (CONJ+NP)/CONJ, factors: year, corpus, conjunction)
  - The results revealed significant main effects for conjunction, $F(1, 6) = 8.94$, $p < .0001$, and corpus, $F(1, 1) = 23.76$, $p < .0001$.
- **Wikipedia**
  - Clearly a dearth of *because-noun* tokens (thus baseline).
- **Reddit**
  - Low level of *because-noun* from 2007-2011.
  - Marked increase from 2011-2015 (cf. **Question #1**).
  - Low level for other conjunctions throughout.
  - Credible instances exist very early, however (5a-f):

(5a) *People have attempted to create an ebonics translation program online, **although fail**.* (Reddit 2007)
(5b) *That's **because programming**.* (Reddit 2008)
(5c) *Upvoted **because anti-scam discrimination**.* (Reddit 2008)
(5d) *No, **because SPACE LASERS**!! How cool is that!!* (Reddit 2009)
(5e) *May not though **cuz tl;dr*** (Reddit 2009)
(5f) *They didn't take her **because Jesus**.* (Reddit 2009)

- **Twitter**
  - Highest levels of *because-noun*, but no steady increase from 2011-2015.
  - Higher levels of other conjunctions, but never as high as *because-noun.*
  - The *conjunction-noun* construction only exists as such in the hyper-informal corpus; perhaps limited to language play (**Question #2**).

- *Because racecar*: an early memetic use
  - Perhaps the earliest use of *because-noun* as an internet meme was ***because racecar***, traced to the automotive blog Jalopnik.com in early 2011 (knowyourmeme.com).
  - *Because racecar* in our data from 2012 (Reddit).



https://pbs.twimg.com/profile_images/1268294539/WHY-Because-Racecar_400x400.jpg

- Future research
  - Refine algorithm to reduce false positives and false negatives (e.g. *because-ADJ*, non-sentence-final *because-noun*).
  - Greater control of social variables (cf. geotagging, gender tagging).
  - Questionnaire data for acceptability, sentiment analysis.
  - More fine-grained time series analysis (e.g. by month).
  - Analysis of the spread of *because-noun* through individual Subreddits.

Notes

[1] The authors would like to acknowledge Sravana Reddy and the students in the "Language Variation through the Lens of Web Data" class at the 2015 LSA Institute. We would also like to thank Bodo Winter and Andrew Wedel for their support and advice.

[2] For this poster, we gloss over demographic differences between Reddit and Twitter, which may be a lurking variable.

[3] Note that articles from the 2015 Wikipedia corpus may have been created or edited before 2015.

[4] The authors would like to point to three subtypes of *because-noun* construction for further research:

a.) *Because* followed by a sentential acronym. These examples were excluded, as they are genuinely ambiguous between the conjunctional and prepositional uses of *because* (6a-b).

(6a) Because-noun: *Eating chocolate animal crackers all day **because yolo**.* (Twitter)
(6b) Because-clause: *Eating chocolate animal crackers all day, **because you only live once**.*

b.) *Because* followed by symbols/emojis. Should these be classed as interjections (7a-b)?

(7a) *I'm gonna look for other schools this year **because** 😟 !!* (Twitter)
(7b) *They couldn't reasonably offer free tea bags **because $$$**.* (Reddit)

c.) Code-switching into or out of English. Some examples are clearly genuine uses of *because-noun* (8a-b), but others may imply errors or other constructions

(8a) *Fazer squats ao som de Beyonce **because goals**. [Doing squats to Beyonce because goals.]* [Portuguese and English] (Twitter)
(8b) *All my life i was told bigger eyes were cuter **because ulzzang*** [English and Korean] (Twitter)

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press.

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., & 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*, companion volume, Portland, OR. <http://www.ark.cs.cmu.edu/TweetNLP/gimpel+etal.acl11.pdf>

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., & Schneider, N. (2012). *Part-of-Speech Tagging for Twitter: Word Clusters and Other Advances*. Technical Report, Machine Learning Department, Carnegie Mellon University. CMU-ML-12-107.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Reddit Corpus [text corpus]. (2007-2015). Accessed August 1, 2015. <https://archive.org/details/2015_reddit_comments_corpus>

Reddy, S., Stanford, J.N., & Zhong, J. (2014). "A Twitter-Based Study of Newly Formed Clippings in American English." Paper presented at the American Dialect Society Annual Meeting.

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>

Russell, M.A. (2013). *Mining the Social Web*. Sebastopol, CA: O'Reilly Media.

Twitter Stream Grab [text corpus]. (2011-2015). Accessed December 1, 2105. <https://archive.org/details/twitterstream>

Wikipedia [text corpus]. (2015). Accessed August 1, 2015. <https://dumps.wikimedia.org/>