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

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; Cary 2013; Bailey 2014; WOTY 2013).

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).

3. Methodology
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).

3b. Methodology, cont’d.
Corpus data sources
• Twitter: sample of 3 months of tweets per year from the “sprinter” 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).

4. Results, cont’d.
4a. Results
Question #1: What is the origin of because-noun?
• Overall corpus sizes and usage rates:

4b. Results, cont’d.
• Usage rates by year for although+NP and because+NP:

5. Conclusions (cuz tl/dr)
• Statistical analysis
• Textual data: ANOVA (response variable: relative frequency of (CONJ+NP): factors: year, corpus, conjunction)
• The results revealed significant main effects for conjunction, F(1, 6) = 8.94, p < .0001, and corpus, F(1, 2) = 23.76, p < .0001.
• Wikipedia
• Clearly a deaf ear of because-noun tokens (thus baseline).

References
• People have attempted to create an ebonics translation program online, although fall. (Redditt 2007)
• That’s because programming. (Redditt 2008)
• Ignored because anti-oum discrimination. (Redditt 2008)
• No, because SPACE LASERS? How cool is that! (Redditt 2009)
• May not know what it is (Redditt 2009)
• They didn’t take it for because Jesus. (Redditt 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 mnemonic use.
• Perhaps the earliest use of because-noun as an internet meme was because racecar, traced to the automotive blog, laps.com in early 2011 (knowyourmeme.com).
• Because racecar in our data from 2012 (Redditt).

• Future research
• Rule algorithm to reduce false positives and false negatives (e.g., because-A02, non-clause-final because-noun).
• Greater control of social variables (cf. geotagging, gender tagging).
• Questionnaire data for accessibility, sentiment analysis.
• More fine-grained time series analysis (e.g. by month).
• Analysis of the spread of because-noun through individual Subreddits.

Notes
• The evidence from the Archive is based on informal use of because-noun, and not necessarily formal colloquialism.
• The evidence from the Archive is based on informal use of because-noun, and not necessarily formal colloquialism.

Table 1. Frequency of usage in informal and formal corpora.

Table 2. Methodology.

Table 3. Because-noun noun-constructions in online corpora.

Table 4. Data processing.

Table 5. Conclusions.

Table 6. Results, cont’d.

Table 7. Methodology.

Table 8. Results, cont’d.

Table 9. Conclusions.

Table 10. Methodology.