

Flipping the *on/off* switch: Change in progress in the prepositional complements of verbs like *base*

Guy Tabachnick and Laurel MacKenzie

New York University

ADS 2023 Annual Conference,
January 8, 2023

- 1 Introduction: flipping the switch
- 2 Examples
- 3 Corpora
- 4 Results
 - Reddit
 - COCA
- 5 Discussion

The basic phenomenon

Many have noticed: rapid rise of *based off (of)* alongside *based on* (Curzan, 2013; Behrens, 2014; Janda, 2020)

The basic phenomenon

Many have noticed: rapid rise of *based off (of)* alongside *based on* (Curzan, 2013; Behrens, 2014; Janda, 2020)

- “Examples exist from as early as 1980, but many academics (like the present author) seem to have felt that their first (ca. 2000) encounter with, e.g., *The book was based off (of) the movie* was due to a unique student’s idiosyncrasy. Yet, within a few years, the strength and breadth of this construction (in the sense of characterizing almost everyone below a certain age) had become evident.” (Janda, 2020: 596)

Previous work: it's growing and here to stay

Curzan (2013): very rare phenomenon in corpora

Previous work: it's growing and here to stay

Curzan (2013): very rare phenomenon in corpora

- Google Ngrams: *based on* outnumbers *based off of* by 100,000:1 in 2000 and 10,000:1 in 2008
- Corpus of Contemporary American English (COCA): *based off of* appears once in academic writing and seven times in speech

Previous work: it's growing and here to stay

Curzan (2013): very rare phenomenon in corpora

- Google Ngrams: *based on* outnumbers *based off of* by 100,000:1 in 2000 and 10,000:1 in 2008
- Corpus of Contemporary American English (COCA): *based off of* appears once in academic writing and seven times in speech

Janda (2020): getting more popular quickly

Previous work: it's growing and here to stay

Curzan (2013): very rare phenomenon in corpora

- Google Ngrams: *based on* outnumbers *based off of* by 100,000:1 in 2000 and 10,000:1 in 2008
- Corpus of Contemporary American English (COCA): *based off of* appears once in academic writing and seven times in speech

Janda (2020): getting more popular quickly

- *based off (of)* has 15–35 million Google hits (1/163 the rate of *based on*)
- *It's based on the* 12 times more common than *It's based off (of) the*

Previous work: it's growing and here to stay

Curzan (2013): very rare phenomenon in corpora

- Google Ngrams: *based on* outnumbers *based off of* by 100,000:1 in 2000 and 10,000:1 in 2008
- Corpus of Contemporary American English (COCA): *based off of* appears once in academic writing and seven times in speech

Janda (2020): getting more popular quickly

- *based off (of)* has 15–35 million Google hits (1/163 the rate of *based on*)
- *It's based on the* 12 times more common than *It's based off (of) the*
- also seen in: *be derived, grounded, justified, predicated, draw, go, live*

A new study

Goal: a more detailed look at this variable

- informal texts
- across multiple verbs
- including *off* without *of*
- in real and apparent time

A new study

Goal: a more detailed look at this variable

- informal texts
- across multiple verbs
- including *off* without *of*
- in real and apparent time

Our study

- seven verbs: *base, build, capitalize, feed, profit, survive, thrive*

A new study

Goal: a more detailed look at this variable

- informal texts
- across multiple verbs
- including *off* without *of*
- in real and apparent time

Our study

- seven verbs: *base, build, capitalize, feed, profit, survive, thrive*
- two corpora

A new study

Goal: a more detailed look at this variable

- informal texts
- across multiple verbs
- including *off* without *of*
- in real and apparent time

Our study

- seven verbs: *base, build, capitalize, feed, profit, survive, thrive*
- two corpora
 - Reddit posts in college and parenting/pregnancy subreddits (Baumgartner, 2019; Chang et al., 2020)
 - COCA (Davies, 2008–)

A new study

Goal: a more detailed look at this variable

- informal texts
- across multiple verbs
- including *off* without *of*
- in real and **apparent time**

Our study

- seven verbs: *base, build, capitalize, feed, profit, survive, thrive*
- two corpora
 - Reddit posts in **college and parenting/pregnancy** subreddits (Baumgartner, 2019; Chang et al., 2020)
 - COCA (Davies, 2008–)

Results: preview

- *Off* is steadily increasing in real time (now over 10%) across different verbs
- Apparent-time shift of about 12 years between college and parenting posts
- *Profit* is far ahead of all other verbs in the shift

- 1 Introduction: flipping the switch
- 2 **Examples**
- 3 Corpora
- 4 Results
 - Reddit
 - COCA
- 5 Discussion

Introducing the examples

- All examples taken from the main parenting subreddit (r/Parenting)

Introducing the examples

- All examples taken from the main parenting subreddit (r/Parenting)
- The methods used to find these examples will be explained later!

Examples: *base*

- (1) I replied to your comment because you *based it on* a bunk article.
- (2) So you didn't base it off of what the OP [original poster] said, you *based it off of* something in your head [...]

Examples: *build*

- (3) You can't have a relationship *built on* power for decades and then suddenly turn it into a relationship *built on* respect by flipping a switch.
- (4) We all carry things that make us different, but saying that all heterosexual relationships are *built off* the same things is silly to me.

Examples: *capitalize*

- (5) Those people are only *capitalizing on* parents that are unprepared and worried about disappointing their kids.
- (6) They just don't seem to have a problem *capitalizing off of* our goodwill and never reciprocating, so that's the problem.

Examples: *feed*

- (7) Ignore the tantrums, she's *feeding on* them.
- (8) Punishment doesn't really help in those situations, because the kid is kind of *feeding off* the tension and conflict.

Examples: *profit*

- (9) You have legal rights since they are *profiting on* your sons image, no matter how little he may have been involved.
- (10) If you are *profiting off* my image without my knowledge in a space that isn't public, then you owe me compensation.

Examples: *survive*

- (11) Mine *survives on* hot dogs and cheese sandwiches.
- (12) Also they feed them a very unhealthy diet(the kids don't know what water is all they drink is pop and juice) the 6 year old *survives off of* microwavable hot dogs and literally nothing else.

Examples: *thrive*

- (13) She also was likely traumatized by the repeated moving and insecure living arrangements - know that adage kids *thrive on* consistency?
- (14) Remember children *thrive off of* consistency, that is how they feel safe and calm.

Summary of examples

The variation extends to a range of contexts:

- Syntactic: active, passive, intervening pronoun, etc.
- Semantic: metaphorical and literal

- 1 Introduction: flipping the switch
- 2 Examples
- 3 Corpora**
- 4 Results
 - Reddit
 - COCA
- 5 Discussion

What is Reddit?

Reddit is a news and discussion website divided into topic-specific “subreddits”

- e.g. r/linguistics

What is Reddit?

Reddit is a news and discussion website divided into topic-specific “subreddits”

- e.g. r/linguistics

Discussions are grouped into threads under posts/links

- e.g. The Scots language Wikipedia is edited primarily by someone with limited knowledge of Scots

Reddit: parameters

- Source: selection from Pushshift.io Reddit Corpus (Baumgartner, 2019) through ConvoKit (Chang et al., 2020)

Reddit: parameters

- Source: selection from Pushshift.io Reddit Corpus (Baumgartner, 2019) through ConvoKit (Chang et al., 2020)
- Search: verb (pronoun) preposition
 - non-lemmatized, so search was string-based (e.g. *survive* = *survive* + *survived* + *survives* + *surviving*)

Reddit: parameters

- Source: selection from Pushshift.io Reddit Corpus (Baumgartner, 2019) through ConvoKit (Chang et al., 2020)
- Search: verb (pronoun) preposition
 - non-lemmatized, so search was string-based (e.g. *survive* = *survive* + *survived* + *survives* + *surviving*)
- Cleaning: string forms yielding many false positives were removed
 - *building* (*a building on campus*)
 - *feed(s)* (*I followed the news feed on YouTube*)
 - *profit(s)* (*I made a profit off of that*—also varies, but differently)

Reddit: parameters

- Source: selection from Pushshift.io Reddit Corpus (Baumgartner, 2019) through ConvoKit (Chang et al., 2020)
- Search: verb (pronoun) preposition
 - non-lemmatized, so search was string-based (e.g. *survive* = *survive* + *survived* + *survives* + *surviving*)
- Cleaning: string forms yielding many false positives were removed
 - *building* (*a building on campus*)
 - *feed(s)* (*I followed the news feed on YouTube*)
 - *profit(s)* (*I made a profit off of that*—also varies, but differently)
- ... too much data to clean manually, so other false positives left in
 - *wild animals who can survive on their own*

Reddit: parameters

- Source: selection from Pushshift.io Reddit Corpus (Baumgartner, 2019) through ConvoKit (Chang et al., 2020)
- Search: verb (pronoun) preposition
 - non-lemmatized, so search was string-based (e.g. *survive* = *survive* + *survived* + *survives* + *surviving*)
- Cleaning: string forms yielding many false positives were removed
 - *building* (*a building on campus*)
 - *feed(s)* (*I followed the news feed on YouTube*)
 - *profit(s)* (*I made a profit off of that*—also varies, but differently)
- ... too much data to clean manually, so other false positives left in
 - *wild animals who can survive on their own*

Total: 149,598 tokens from 2009–2018 (mostly 2012–2018)

Reddit: cohorts

Subreddits divided into two rough “age cohorts” assumed to have young(ish) speakers

Subreddits divided into two rough “age cohorts” assumed to have young(ish) speakers

- college
 - r/college
 - subreddits from individual colleges (Ding, 2018)

Subreddits divided into two rough “age cohorts” assumed to have young(ish) speakers

- college
 - r/college
 - subreddits from individual colleges (Ding, 2018)
- parenting
 - r/BabyBumps (pregnancy)
 - r/Parenting

Tokens from 2000–2019 divided by type

Tokens from 2000–2019 divided by type

- web and blog
- academic
- magazine, news, spoken
- fiction, movies, TV

Tokens from 2000–2019 divided by type

- ~~web and blog~~ from 2012 only
- academic
- magazine, news, spoken
- fiction, movies, TV

Tokens from 2000–2019 divided by type

- ~~web and blog~~ from 2012 only
- *academie* very little *off*
- magazine, news, spoken
- fiction, movies, TV

Tokens from 2000–2019 divided by type

- ~~web and blog~~ from 2012 only
- ~~academic~~ very little *off*
- magazine, news, spoken
- ~~fiction, movies, TV~~ very little data

Tokens from 2000–2019 divided by type

- ~~web and blog~~ from 2012 only
- ~~academic~~ very little *off*
- **magazine, news, spoken**
- ~~fiction, movies, TV~~ very little data

Total: 34,096 tokens from 2000–2019

Issues with Reddit corpus:

Corpus comparison

Issues with Reddit corpus:

- not lemmatized
 - certain strings removed (*building, feed(s), profit(s)*)

Issues with Reddit corpus:

- not lemmatized
 - certain strings removed (*building, feed(s), profit(s)*)
- no demographics for users (other than location of college)

Issues with Reddit corpus:

- not lemmatized
 - certain strings removed (*building, feed(s), profit(s)*)
- no demographics for users (other than location of college)

COCA is limited to American texts and true verbs, but has its own disadvantages

Issues with Reddit corpus:

- not lemmatized
 - certain strings removed (*building, feed(s), profit(s)*)
- no demographics for users (other than location of college)

COCA is limited to American texts and true verbs, but has its own disadvantages

- data quite sparse, especially for verbs other than *base*
- no reliable formality distinction, certainly no age cohorting

Main results: testing hypotheses on Reddit corpus

Main results: testing hypotheses on Reddit corpus

- real-time shift: 2018 *off* > 2012 *off*
- apparent-time shift: college *off* > parenting *off*

Main results: testing hypotheses on Reddit corpus

- real-time shift: 2018 *off* > 2012 *off*
- apparent-time shift: college *off* > parenting *off*

Confirmation: COCA

Main results: testing hypotheses on Reddit corpus

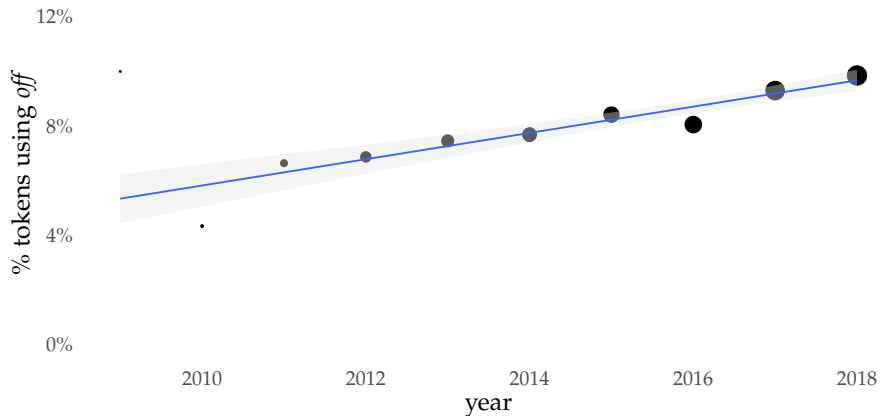
- real-time shift: 2018 *off* > 2012 *off*
- apparent-time shift: college *off* > parenting *off*

Confirmation: COCA

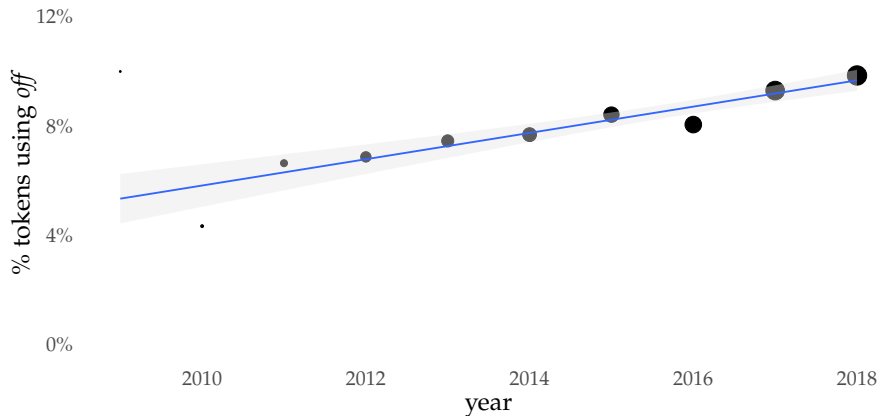
- basic real-time trend
- comparison of individual verbs

- 1 Introduction: flipping the switch
- 2 Examples
- 3 Corpora
- 4 Results**
 - Reddit
 - COCA
- 5 Discussion

Reddit: real-time shift

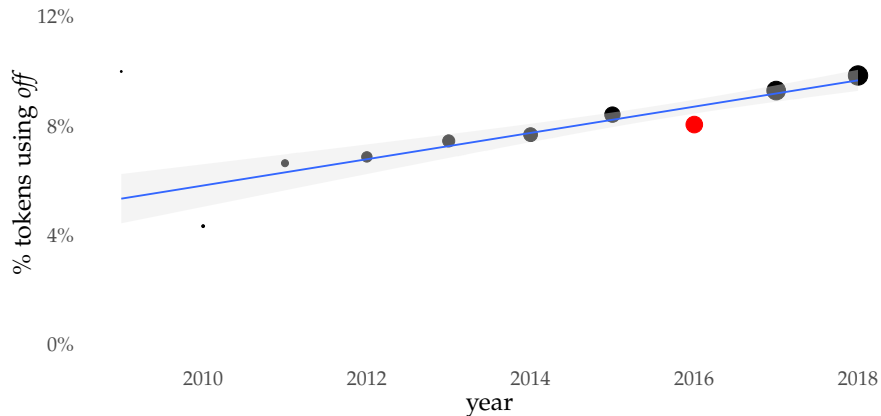


Reddit: real-time shift



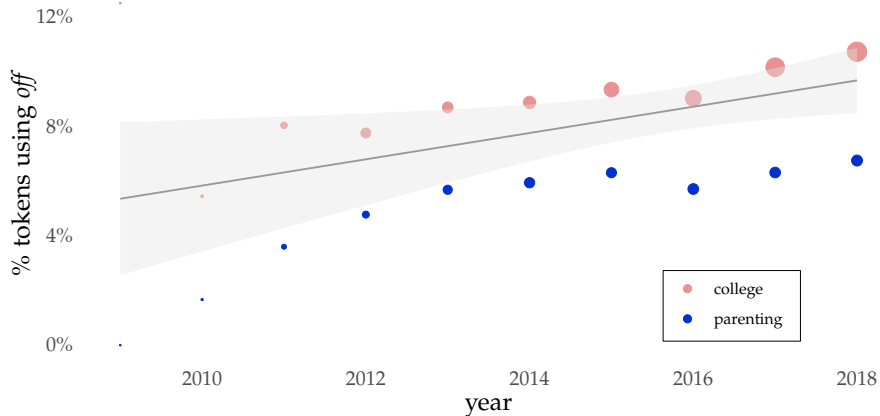
- steady rise from 7–10%

Reddit: real-time shift

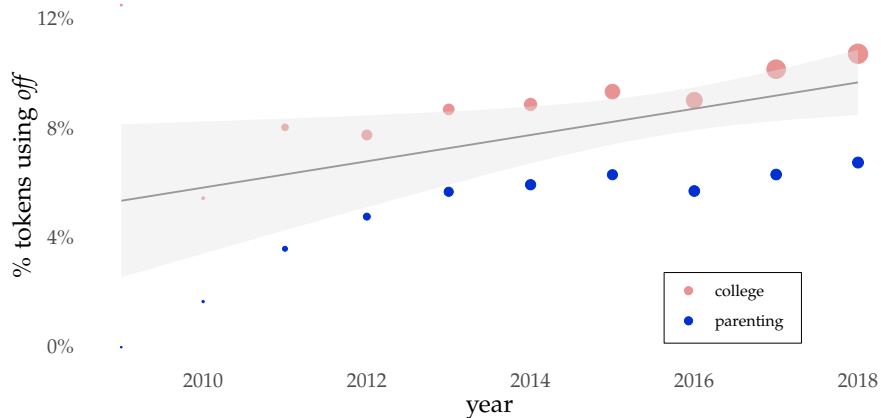


- steady rise from 7–10%
- no explanation for dip in 2016

Reddit: apparent-time shift

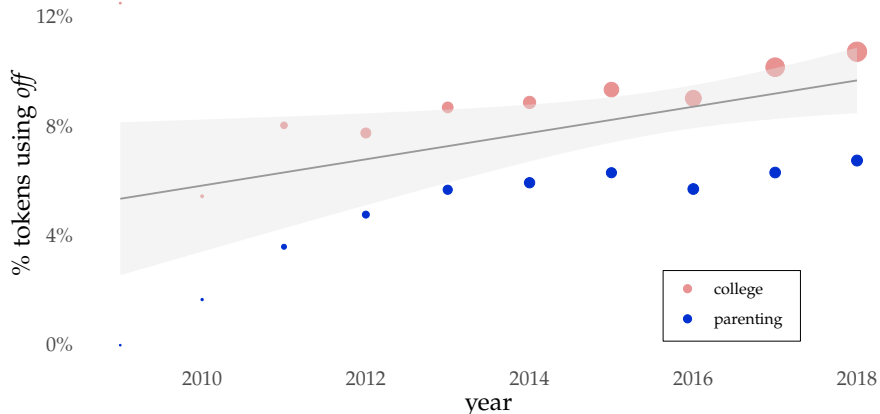


Reddit: apparent-time shift



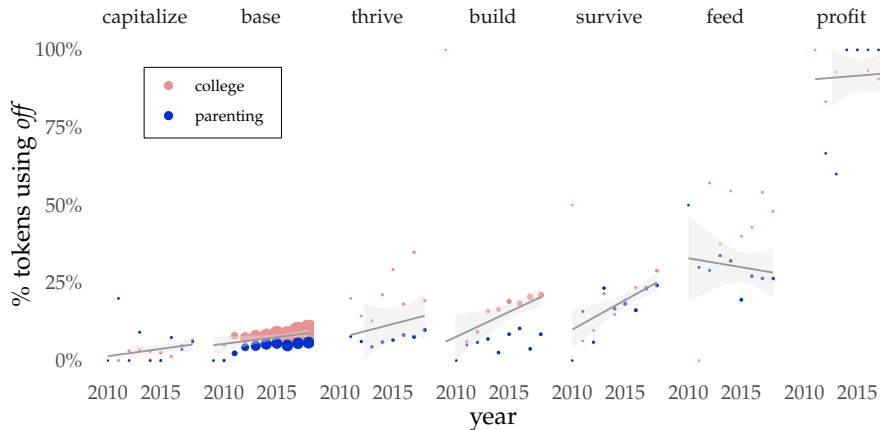
- parenting about 12 years behind college

Reddit: apparent-time shift

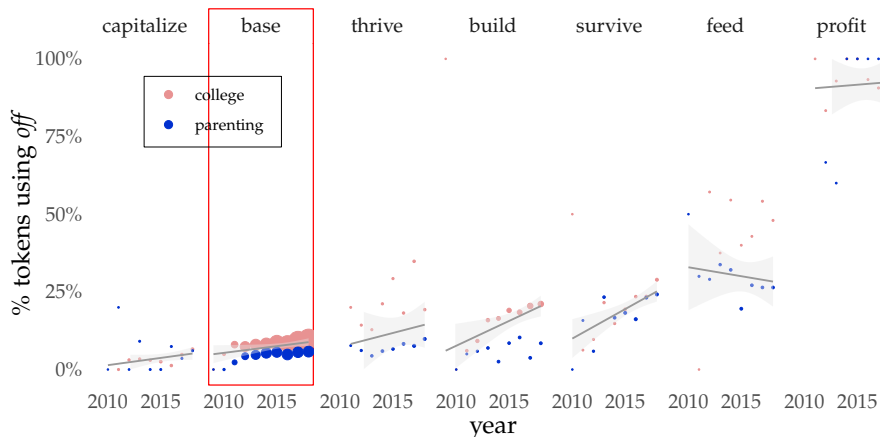


- **parenting** about 12 years behind **college**
 - r/BabyBumps about 8 years behind **college**
 - r/Parenting about 18 years behind **college**

Reddit: verb-by-verb results

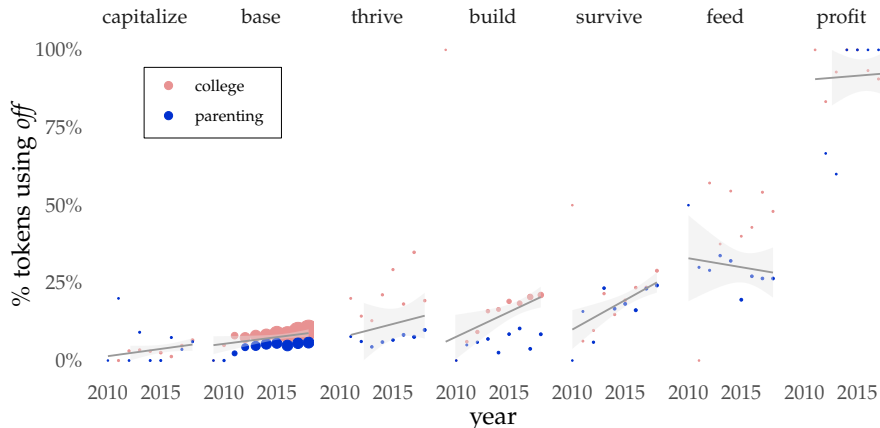


Reddit: verb-by-verb results



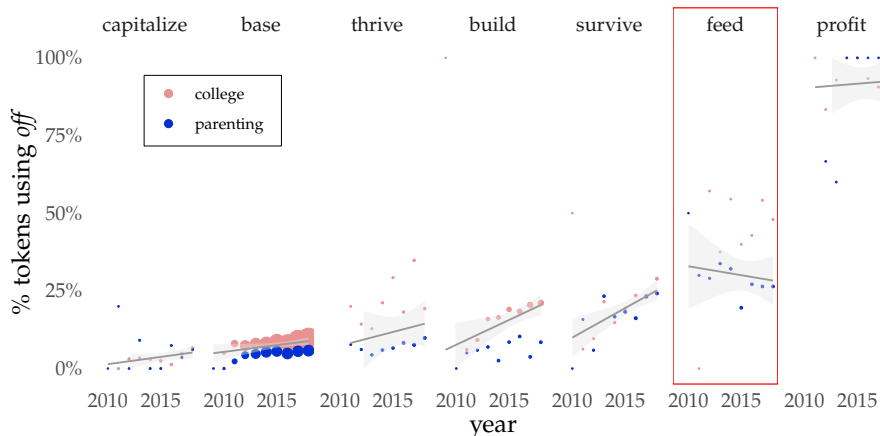
● most tokens are *base* (other verbs differ significantly from it)

Reddit: verb-by-verb results



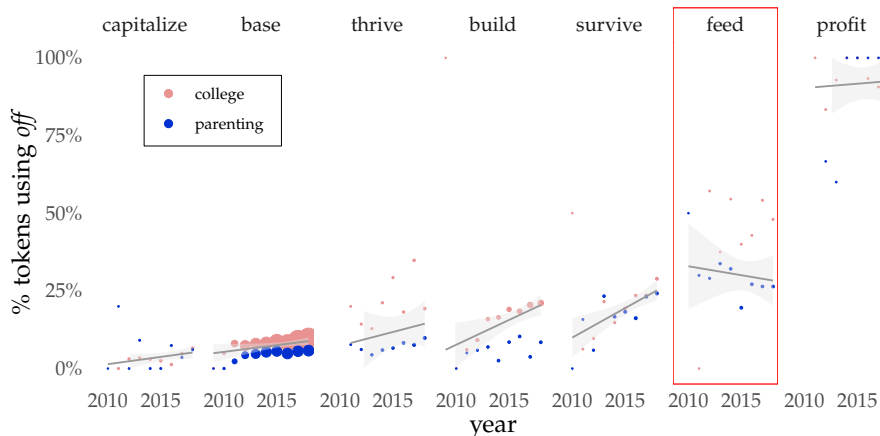
- most tokens are *base* (other verbs differ significantly from it)
- obvious outliers:

Reddit: verb-by-verb results



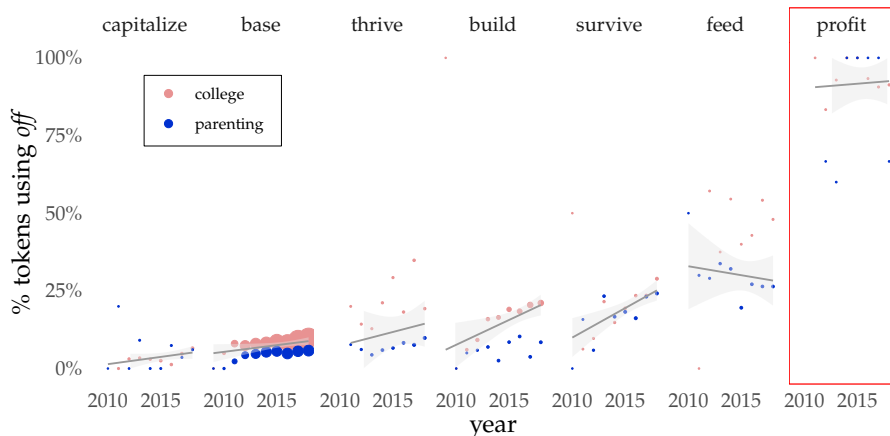
- most tokens are *base* (other verbs differ significantly from it)
- obvious outliers:
 - *feed*: constructions like *feed on demand*

Reddit: verb-by-verb results



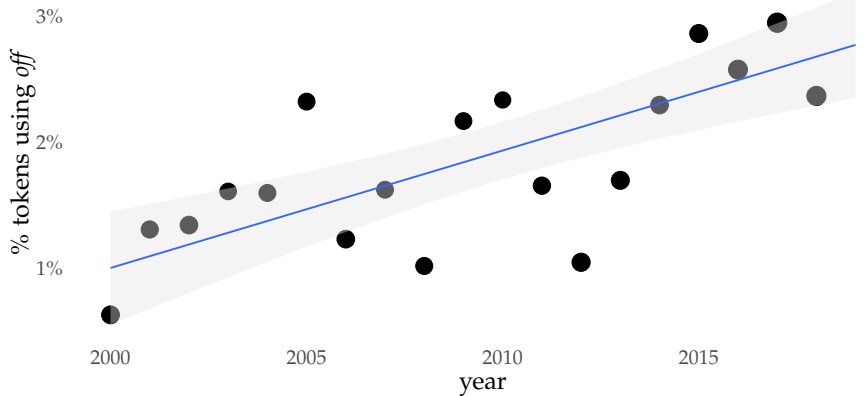
- most tokens are *base* (other verbs differ significantly from it)
- obvious outliers:
 - *feed*: constructions like *feed on demand* (for **parenting**)

Reddit: verb-by-verb results

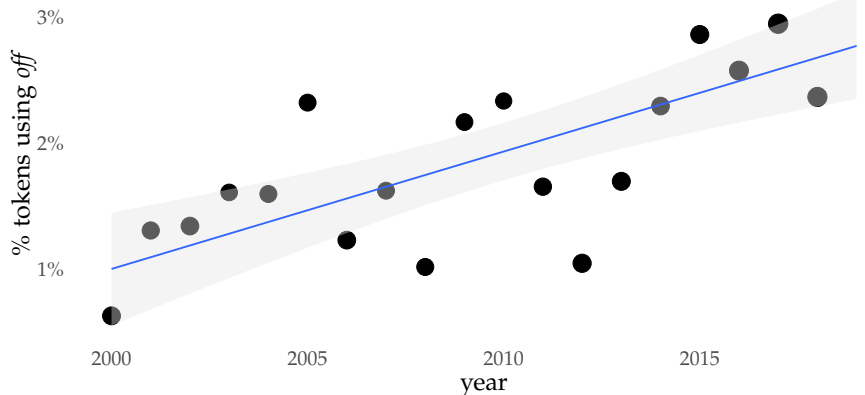


- most tokens are *base* (other verbs differ significantly from it)
- obvious outliers:
 - *feed*: constructions like *feed on demand* (for **parenting**)
 - *profit*: almost entirely *off* (with very few tokens)

COCA: real-time shift

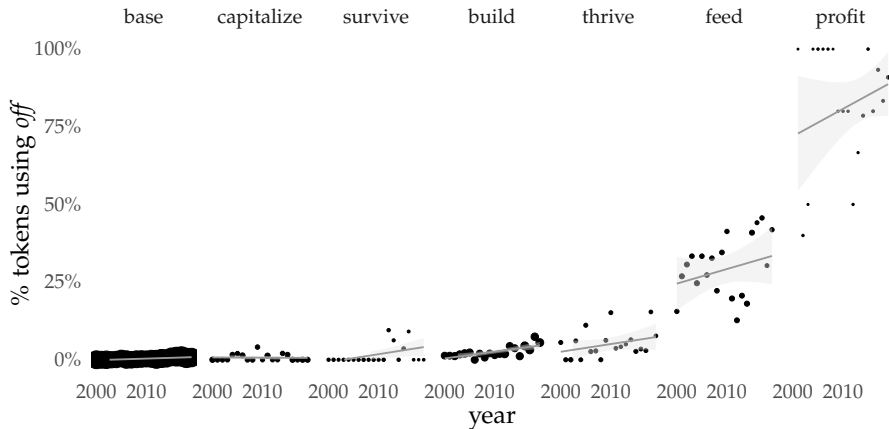


COCA: real-time shift

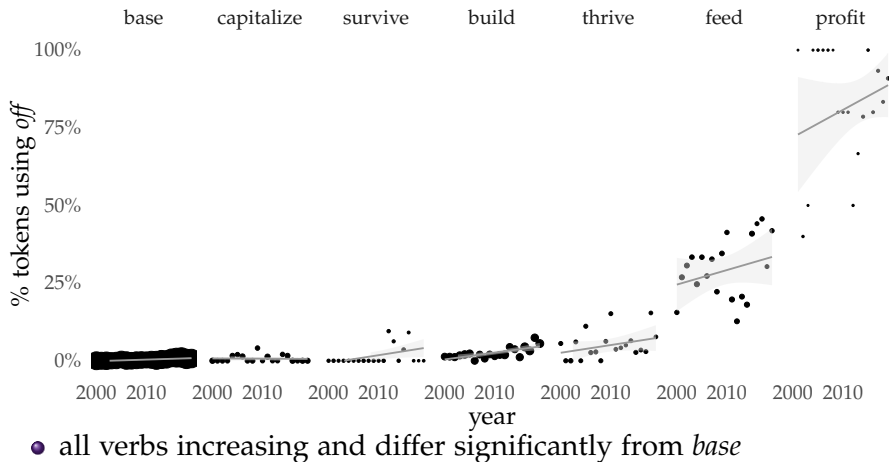


- rate of *off* is steadily (if noisily) rising, but much lower

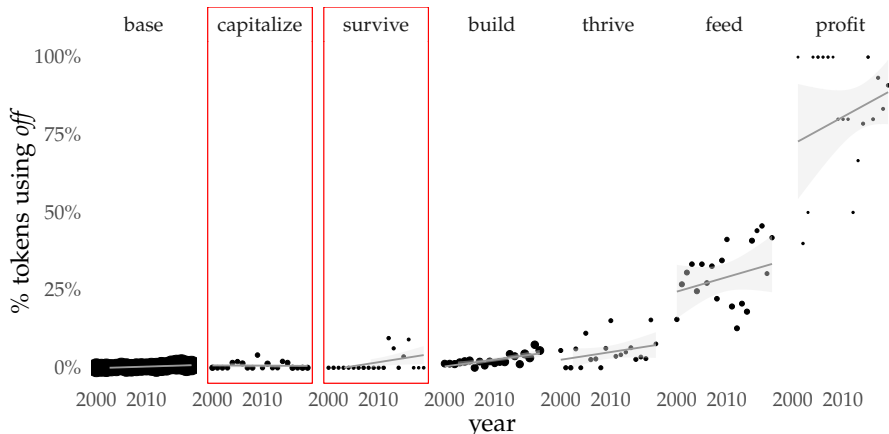
COCA: verb-by-verb results



COCA: verb-by-verb results

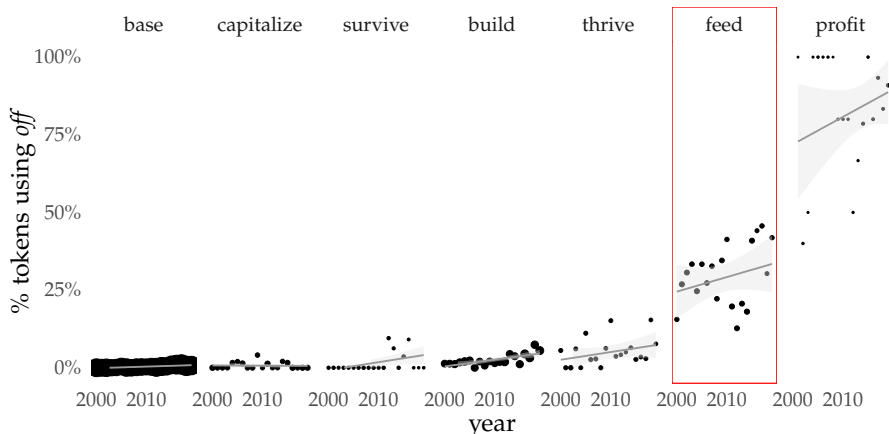


COCA: verb-by-verb results



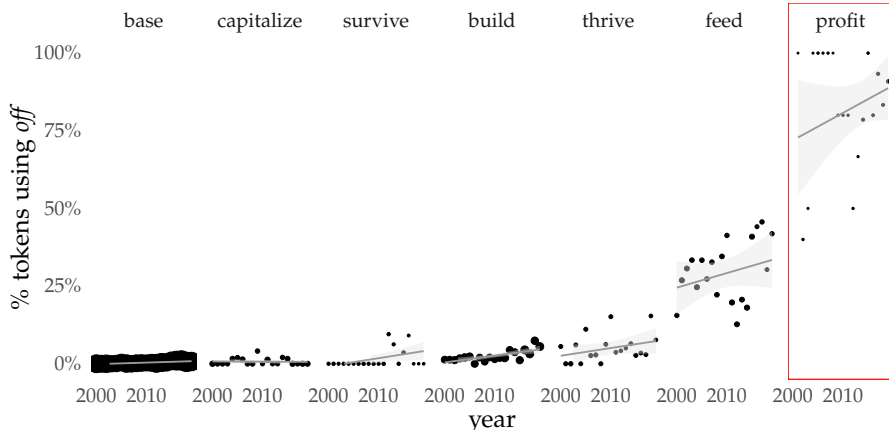
- all verbs increasing and differ significantly from *base*
- some verbs (*capitalize, survive*) near 0

COCA: verb-by-verb results



- all verbs increasing and differ significantly from *base*
- some verbs (*capitalize*, *survive*) near 0
- *feed* is better behaved than in Reddit

COCA: verb-by-verb results



- all verbs increasing and differ significantly from *base*
- some verbs (*capitalize*, *survive*) near 0
- *feed* is better behaved than in Reddit
- *profit* similarly high

Summary of results

Reddit:

Summary of results

Reddit:

- steady increase of *off* to $> 10\%$ by 2018

Summary of results

Reddit:

- steady increase of *off* to $> 10\%$ by 2018
- **parenting** posters are 12.25 years behind **college** posters

Summary of results

Reddit:

- steady increase of *off* to $> 10\%$ by 2018
- **parenting** posters are 12.25 years behind **college** posters
- individual verbs at different points in the shift
 - *feed* is much higher than most verbs, and trending downwards (for **parenting**)
 - *profit* takes almost entirely *off*

Summary of results

Reddit:

- steady increase of *off* to $> 10\%$ by 2018
- **parenting** posters are 12.25 years behind **college** posters
- individual verbs at different points in the shift
 - *feed* is much higher than most verbs, and trending downwards (for **parenting**)
 - *profit* takes almost entirely *off*

COCA:

Summary of results

Reddit:

- **steady increase of *off*** to $> 10\%$ by 2018
- **parenting** posters are 12.25 years behind **college** posters
- individual verbs at different points in the shift
 - *feed* is much higher than most verbs, and trending downwards (for **parenting**)
 - *profit* takes almost entirely *off*

COCA:

- **steady increase of *off*** to 2.5–3% in 2019

Summary of results

Reddit:

- steady increase of *off* to $> 10\%$ by 2018
- **parenting** posters are 12.25 years behind **college** posters
- individual verbs at different points in the shift
 - *feed* is much higher than most verbs, and trending downwards (for **parenting**)
 - ***profit* takes almost entirely off**

COCA:

- steady increase of *off* to 2.5–3% in 2019
- ***profit* takes almost entirely off**

Summary of results

Reddit:

- steady increase of *off* to $> 10\%$ by 2018
- **parenting** posters are 12.25 years behind **college** posters
- individual verbs at different points in the shift
 - ***feed* is much higher than most verbs, and trending downwards (for **parenting**)**
 - *profit* takes almost entirely *off*

COCA:

- steady increase of *off* to 2.5–3% in 2019
- *profit* takes almost entirely *off*
- ***feed* is much higher than most verbs, and trending upwards like the rest**

- 1 Introduction: flipping the switch
- 2 Examples
- 3 Corpora
- 4 Results
 - Reddit
 - COCA
- 5 Discussion

A methodological question: Is the Reddit corpus useful?

Despite its shortcomings, the Reddit corpus gave *reasonable, interpretable results*, and offers:

- a lot of text
- organized thematically in detail

Is *on/off* a single sociolinguistic variable?

Preliminary conclusion: yes

Is *on/off* a single sociolinguistic variable?

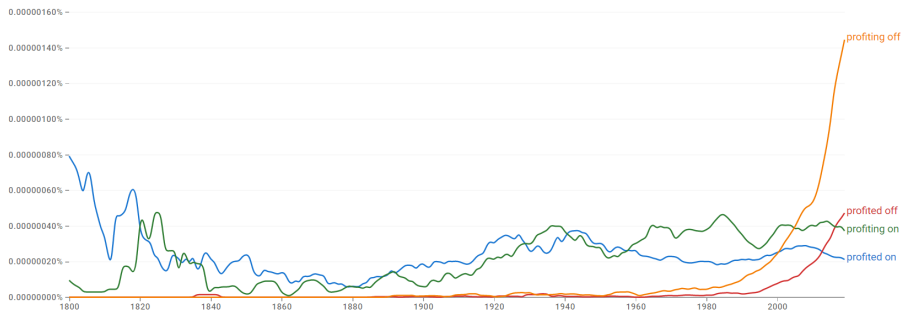
Preliminary conclusion: yes

- *On* and *off* are attested across multiple verbs and contexts
- Verbs have somewhat different rates (lexical diffusion?), but all are increasing
- *Profit* has a much higher *off* rate than others

Is *on/off* a single sociolinguistic variable?

Preliminary conclusion: yes

- *On* and *off* are attested across multiple verbs and contexts
- Verbs have somewhat different rates (lexical diffusion?), but all are increasing
- *Profit* has a much higher *off* rate than others
 - ... but a Google Ngram search suggests it's simply much further along in the change



What are its conditioning factors?

The verb definitely matters, but other things could be explored:

- Do syntactic factors (tense, intervening objects, etc.) matter?
- Do certain verbs prefer *off* rather than *off of*?
- Do literal and metaphorical uses of verbs like *thrive* behave differently?

What are its conditioning factors?

The verb definitely matters, but other things could be explored:

- Do syntactic factors (tense, intervening objects, etc.) matter?
- Do certain verbs prefer *off* rather than *off of*?
- Do literal and metaphorical uses of verbs like *thrive* behave differently?

Social factors (region, race, etc.)—needs different corpora

What's driving the shift?

“[T]he essence and puissance of the innovative construction is metaphorical: if one derives something from a source, then a crucial pathway between them leads from the source to the derivative”
(Janda, 2020: 597)

- *off* suggests **extraction**

What's driving the shift?

“[T]he essence and puissance of the innovative construction is metaphorical: if one derives something from a source, then a crucial pathway between them leads from the source to the derivative”

- *off* suggests **extraction**

“... Yet *basing* or *being based ON* portrays the implied motion as oriented in [the opposite] direction, and thus sounds more like planting a flagstaff downward into the ground.” (Janda, 2020: 597)

- alternately: *on* suggests **foundation**

What's driving the shift?

“[T]he essence and puissance of the innovative construction is metaphorical: if one derives something from a source, then a crucial pathway between them leads from the source to the derivative”

- *off* suggests **extraction**

“... Yet *basing* or *being based ON* portrays the implied motion as oriented in [the opposite] direction, and thus sounds more like planting a flagstaff downward into the ground.” (Janda, 2020: 597)

- alternately: *on* suggests **foundation**

Perhaps the shift from *on* to *off* is a change from a metaphor of foundation to one of extraction

- *profit* is particularly well-suited to this → earlier shifter
- the lexical semantics of *base* are compatible with foundation → later shifter

References I

Baumgartner, J. (2019). Reddit corpus. <https://files.pushshift.io/reddit/>.

Behrens, S. J. (2014). *Understanding Language Use in the Classroom: A Linguistic Guide for College Educators*. Multilingual Matters, Toronto.

Chang, J. P., Chiam, C., Fu, L., Wang, A., Zhang, J., and Danescu-Niculescu-Mizil, C. (2020). ConvoKit: A toolkit for the analysis of conversations. In Pietquin, O., Muresan, S., Chen, V., Kennington, C., Vandyke, D., Dethlefs, N., Inoue, K., Ekstedt, E., and Ultes, S., editors, *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60. Association for Computational Linguistics.

Curzan, A. (2013). Based off of what?

<https://web.archive.org/web/20131228114705/https://www.chronicle.com/blogs/linguafranca/2013/10/23/off-of-what/>.

Davies, M. (2008–). The Corpus of Contemporary English (COCA). <https://www.english-corpora.org/coca/>.

Ding, K. (2018). GitHub - karlding/college-subreddits: an eventually-complete list of college (and university) subreddits. <https://github.com/karlding/college-subreddits>.

Janda, R. D. (2020). Perturbations, practices, predictions, and postludes in a bioheuristic historical linguistics. In Janda, R. D., Joseph, B. D., and Vance, B. S., editors, *The Handbook of Historical Linguistics*, volume II, chapter 24, pages 523–650. Wiley Blackwell, Hoboken, NJ.

Model of Reddit corpus

	β coef	SE	Wald z	p
Intercept	-2.370	.012	-193.83	<.0001
Year (default: 2015)	.047	.005	8.86	<.0001
Verb (default: base)				
build	.736	.042	17.31	<.0001
capitalize	-.829	.194	-4.28	<.0001
feed	2.020	.079	25.60	<.0001
profit	4.863	.277	17.56	<.0001
survive	1.255	.069	18.13	<.0001
thrive	.703	.093	7.53	<.0001
Subreddit type (default: college)				
parenting	-.575	.024	-23.71	<.0001
Year * Subreddit type (default: 2015 * base)				
year * build	.065	.020	3.24	.0012
year * capitalize	.086	.090	0.96	.3361
year * feed	-.079	.038	-2.07	.0387
year * profit	-.023	.126	-0.18	.8557
year * survive	.062	.033	1.86	.0629
year * thrive	.044	.044	0.99	.3238

Model of COCA

	β coef	SE	Wald z	p
Intercept	-6.029	.147	-41.02	<.0001
Year (default: 2010)	.170	.023	7.37	<.0001
Verb (default: base)				
build	2.276	.190	11.97	<.0001
capitalize	1.032	.387	-2.66	.0077
feed	5.140	.162	31.82	<.0001
profit	7.498	.306	24.52	<.0001
survive	1.556	.615	2.53	.0114
thrive	3.053	.236	12.91	<.0001
Year * Subreddit type (default: 2010 * base)				
year * build	-.076	.031	-2.50	.0125
year * capitalize	-.181	.064	-2.85	.0044
year * feed	-.147	.026	-5.65	<.0001
year * profit	-.109	.050	-2.18	.0295
year * survive	.036	.112	0.32	.7491
year * thrive	-.115	.040	-2.92	.0035