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

GUY TABACHNICK LAUREL MACKENZIE

Univerza v Novi Gorici *New York University*

Corresponding author:

Laurel MacKenzie
New York University
10 Washington Pl
New York, NY, USA 10003
laurel.mackenzie@nyu.edu

Competing interest: The authors declare none.

ABSTRACT

Traditionally, verbs like *base* have combined with the preposition *on* to express a meaning of derivation (*based on*). However, many writing in a US context have noticed the rapid rise of *based off (of)* alongside *based on* (Curzan 2013; Behrens 2014; Janda 2020). In this paper, we document the relative increase of *off* in two English-language corpora in the verb *base* and six other verbs. The results show a clear real-time trend of increasing use of *off*, with some differences in the course of the change across different verbs. We also see an increase in use of *off* in apparent time, which we infer from the topical organization of comments in one of our corpora, the social media site Reddit.

Keywords: Variation, Prepositions, Lexical diffusion, Apparent-time study, Corpus methodology

1 Introduction

This paper studies variation between *on* and *off* as the prepositional complement of a select set of English verbs. One verb in which the variation has been well documented is *base*; (1) gives examples of the variants.[1]

(1) *Base on/off*

(a) I replied to your comment because you ***based it on*** a bunk article.

(b) So you didn't ***base it off of*** what the OP [original poster] said, you ***based it off of*** something in your head [. . .]

The *Oxford English Dictionary* (s.v. *base*) gives only examples with *on* (or *upon*) complements, dating back to 1776. But the variation demonstrated in (1) has received some attention in the linguistic literature, much of it observing rapid change in progress. Janda (2020) finds examples of forms like *based off (of)* from as early as 1980, but dates his first encounter with the *off* variant to ca. 2000, and suggests rapid change thereafter:

> [W]ithin a few years, the strength and breadth of this construction (in the sense of characterizing almost everyone below a certain age) had become evident. (596)

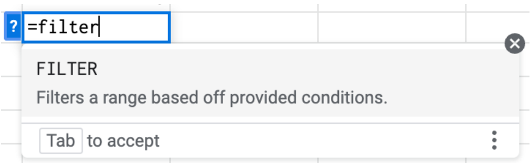
This is confirmed by Curzan (2013), who finds that *based off of* is rare in the Corpus of Contemporary American English (Davies, 2008–) but growing: in Google Books Ngram Corpus data from 2000 (Michel et al. 2011), *based on* outnumbers *based off*

[1] All of the numbered examples provided in this paper are from the r/Parenting subreddit of the Reddit corpus described in Section 3.1 unless a different subreddit source is noted.

of by 100,000:1, but by 2008, this has fallen to 10,000:1. Janda (2020: p. 597) finds that Google hits containing *based on* outnumber those containing *based off (of)* at a ratio of only 163:1, and the raw numbers of *based off (of)* hits are high, exceeding 50 million. Finally, Behrens (2014) provides a more qualitative assessment of the growing popularity of *based off of* (as opposed to *based on*):

> As of this writing, I hear it and see it written all the time from my students and from my younger colleagues; my older colleagues dismiss the structure as just plain wrong. (67)

Anecdotally, we note that the *off* variant is used in pop-up text in Google Sheets as of July 2024 (Figure 1).



=filter

FILTER

Filters a range based off provided conditions.

Tab to accept

Figure 1: *Based off* in pop-up text in Google Sheets.

(2) *Build on/off* Janda (2020: p. 597) observes that this variation between *on* and *off* can be found with other verbs, namely *derive, ground, justify, predicate, draw, go,* and *live*. Examples of this variation in still other verbs are provided in (2)–(7).[2]

[2] A reviewer suggests that speakers who allow a given verb to combine with both prepositions may assign phrases like *draw on* and *draw off* slightly different meanings. Such semantic differentiation is common in cases where two variants coexist (see, e.g., Traugott 2004 on “anti-synonymy”) and is even attested in less obviously semantically driven domains like inflectional morphology (Kiefer 1985: p. 108; Bermel & Knittl 2012; Tabachnick 2023: pp. 270–1). Accordingly, certain

(a) Peaceful parent, happy siblings should give you a good foundation to ***build on***.

(b) We don't use all of it, but it gave us a good foundation to ***build off of***.

(3) *Capitalize on/off*

(a) Those people are only ***capitalizing on*** parents that are unprepared and worried about disappointing their kids.

(b) They just don't seem to have a problem ***capitalizing off of*** our goodwill and never reciprocating, so that's the problem.

(4) *Feed on/off*

(a) Ignore the tantrums, she's ***feeding on*** them.

(b) It sounds like he's ***feeding off*** your stress.

(5) *Profit on/off*

(a) You have legal rights since they are ***profiting on*** your son[']s image, no matter how little he may have been involved.

(b) If you are ***profiting off*** my image without my knowledge in a space that isn't public, then you owe me compensation.

(6) *Survive on/off*

(a) For three months we ***survived on*** our credit cards, then when the credit ran out, we burned through the savings we had set aside for remodeling.

semantic distinctions do not necessarily place the two prepositions outside the envelope of variation. As far as we can tell, the verbs shown below and discussed in this paper have, at least, widely overlapping meanings when combined with both *on* and *off*.

(b) We ***survived very well off*** my dad[']s salary so it wasn't for the money, just for something to do.

(7) *Thrive on/off*

(a) She also was likely traumatized by the repeated moving and insecure living arrangements - know that adage kids ***thrive on*** consistency?

(b) Remember children ***thrive off of*** consistency, that is how they feel safe and calm.

Examples (1)–(7) demonstrate that the variation occurs in a variety of tenses and aspects, with and without intervening object pronouns and adverbs.

Our contributions in this paper are as follows. First, we present novel quantitative evidence for variation and change in the prepositional complements of the verbs given in (1)–(7). We provide real-time evidence for increased used of *off (of)* in informal written language, drawing on data from the online discussion forum Reddit, and in the Corpus of Contemporary American English (Davies, 2008–). This builds on Janda's and Curzan's corpus studies by presenting data on verbs beyond *base*, and by including *off* with and without the following *of*. Second, we present a proof of concept for a methodology that uses the structure of Reddit to infer the ages of authors, thus also providing apparent-time evidence for increased use of *off (of)*. Although the Reddit corpus has been used for sociolinguistic research before (e.g. Flesch 2019; CH-Wang & Jurgens 2021; Brook & Blamire 2023), its potential for inferring demographics is underutilized. Thus, our study confirms that Reddit, whose enormous size makes it a valuable potential source of sociolinguistic data, can be used (with caution) to study demographic factors like age (and, likely, geography) despite lacking overt demographic metadata for most of its users.

2 Variation and change in prepositions elsewhere in English

Variation and change in prepositions has been attested elsewhere in English. The *based on/off* variable is reminiscent of variation in the complement of *different*, for which *from*, *than*, and *to* are all attested, with geographical and social conditioning of their use (Iyeiri, Yaguchi & Okabe 2004; Mair 2007). Behrens & Mercer (2007), Behrens (2014), and Schlüter (2022) give additional examples of preposition variation in fixed expressions in English, some of which can be observed in the writing of contemporary native American English speakers – such as *have concerns on* (standardly *about*) and *look forward for* (standardly *to*) – and others of which show regional variation – such as *chat with*, which skews North American, versus *chat to*, which skews British. We know of only one instance of prepositional variation which may be showing the kind of rapid change observed for *base*, and it is very lexically restricted: the emergence of *on accident* as a competitor to *by accident*, which is rapidly increasing in apparent time (Barratt 2006).

The *off* variant of our variable is implicated in another case of prepositional variation: the variable presence of *of* after *off*. This variation is fairly widespread in English, appearing not just with *off* (e.g. *get off (of) the bus, the islands off (of) the coast*), but also with other prepositions and words that take *of*-headed complements: *out (of) the window, all (of) the children, not that big (of) a deal* (Estling 1999, 2000; Nylund & Seals 2010; Vartiainen & Höglund 2020). This variation between *of* and ∅ shows social and geographical conditioning, though the specifics depend on the particular construction: for instance, after *off*, the use of *of* is deemed non-standard and prescribed against in formal writing (Vartiainen & Höglund 2020), while after *out*, *of* is favored in formal written language (Estling 1999). We do not speculate on the social correlates or diachronic trajectory of the *of* variant in the *based off* construction, instead grouping together *off* and *off of* variants.

In fact, there are even more combinations of prepositions possible in the construction under study. Both *on* and *off* appear sporadically in our corpus preceding *from*:

(8) (a) And how much of ANS2 ***builds on from*** prior knowledge from ANS1? (r/UCDavis)

(b) His test[s] are very straight forward, heavily ***based off from*** his lectures. (r/UniversityOfHouston)

As far as we are aware, this combination of prepositions has not been previously remarked upon, and it is quite rare – we group the 34 such sentences in our data under the broader umbrellas of *on* and *off*.

3 Methods

3.1 The variable and the data sources

We examine variation and change in the prepositional complements of seven verbs (*base*, *build*, *capitalize*, *feed*, *profit*, *survive*, and *thrive*) in two corpora. These verbs were selected through manual inspection of tokens of verbs appearing with both *on* and *off* in a small sample corpus of posts from the social media website Reddit (Chang et al. 2020; Baumgartner 2019), described in more detail below. Verbs were selected primarily for practical purposes – for example, verbs that yielded too many irrelevant tokens (such as *live*, whose combination with *off* and *on* often expresses a location, like *Many students* ***live on*** *Fifth Street*) were not included. This list of verbs is intended to be representative, not exhaustive: our goal is to show that the shift is occurring in at least this handful of verbs.

Of these seven verbs, none is attested in the *Oxford English Dictionary* (OED) with an *off* complement, but five are attested with *on* complements, confirming that they traditionally take *on* in the standard language: *base*, *build*, *capitalize*, *feed*, and *thrive*. Of the remaining two, *survive* is not shown combining with any prepositions in the OED, but is attested with *on* in Google Ngrams (Michel et al. 2011) and our data. *Profit* is perhaps the outlier among our verbs: the OED lists it as combining with other prepositions, *by*, *of*, and *from*, whose usage rates exceed those of *on* in Google Ngrams at most time points. However, *profit on* is attested fairly robustly in Google Ngrams from 1800, and appears in our data, as in (5). Accordingly, we include it in our data (and return to its special status in Section 5).

Because our studied variable is infrequent (other than with *base*), we prioritized large data sets. Our data come from two sources, chosen for their size, their ease of use, and their ability to provide real- and apparent-time data. The first is a corpus of posts from Reddit (Chang et al. 2020; Baumgartner 2019), a news and discussion website divided into topic-specific ‘subreddits’, such as ‘r/linguistics’, a forum for discussion of topics and questions related to linguistics, and ‘r/Legomarket’, a forum where users coordinate buying, selling, and swapping LEGO products. Within a subreddit, discussions are grouped into threads: for instance, r/linguistics contains discussion threads devoted to specific academic articles and weekly Q&A threads where users are encouraged to ask and answer linguistics-related questions. Our Reddit data ranges from 2009 to 2018, though data before 2012 is sparse. The Reddit corpus contains over 7 billion utterances – that is, post submissions and comments (Baumgartner et al. 2020).

For this study, we selected posts from subreddits comprising three rough 'age cohorts': college, pregnancy, and young parent. Our college cohort data set includes posts from the r/college subreddit and subreddits from individual colleges (Ding 2018). The other cohorts comprise posts from r/BabyBumps (a pregnancy-related forum) and r/Parenting, respectively. These age cohorts are intended to show the presence of change in apparent time: we presume that participants in college-related forums tend to be younger than those in pregnancy forums, who in turn tend to be somewhat younger than participants in parenting discussions. These subreddits included 19.4 million utterances (of which 13 million were on college subreddits) with a total of 993 million words (547 million from college subreddits).[3]

Our second source of data is the Corpus of Contemporary American English (COCA, Davies, 2008–), which includes approximately one billion words from 1990–2019 in eight genres across formal written language (academic texts, newspapers, magazines, fiction), online written language (websites, blogs), television and movie subtitles, and spoken language (unscripted conversations from television and radio programs).

Our COCA data uses the magazines, newspapers, and spoken language genres; each includes approximately 125 million words. Two web-based genres (web and

[3] These numbers count punctuation marks as separate words and thus overstate the size of the corpus slightly.

blog posts) were excluded because all of their texts are coded as being from 2012, and thus cannot be used to show shifts in time. The other genres were excluded after preliminary searches showed very little use of the *off* variant.

Of our two data sources, Reddit plays the primary role. It has important advantages: it is very large and is written in more informal language, meaning that it contains many tokens of our verb–preposition constructions (three times as many as COCA; see Section 3.2 for precise counts). The division into subreddits also allows us to sample from (presumed) different demographics (see also CH-Wang & Jurgens 2021).

Reddit also has downsides as a source of sociolinguistic data. Its text is not lemmatized, so we cannot restrict our searches to verbal forms only, increasing the false positive rate (although we took measures to mitigate this; see Section 3.2). The geographical distribution of the Reddit data is also difficult to determine: Reddit draws users from around the world, and the college subreddits include colleges from outside the US (Ding 2018). In addition, the Reddit data falls within a narrow window of time, primarily 2012–2018.

These limitations of Reddit lead us to caution in using it to study variation in American English. Accordingly, we conduct a parallel study in COCA. Although COCA also does not contain sociodemographic information, its texts are all American English and its data is generally high-quality. COCA also has part-of-speech tagging, which in theory allows us to target verbal forms (however, the tagger sometimes misclassifies nouns like *building* as verbs; see Section 3.2). In addition, the greater time scale of COCA (stretching back to 1990) allows us to observe variation in the use of *off* for longer, and before use of *based off (of)* began to become salient

– around 2000, according to Janda (2020) and Curzan (2013).

At the same time, COCA has disadvantages compared to Reddit. Much of its text, even in the genres chosen, is more formal and edited. While we do look for genre differences in COCA, there is no expected demographic difference (and thus, no apparent-time interpretation) between the genres.

Thus, our main, more interesting findings are in the Reddit data. The COCA data is interpreted primarily as a sanity check on the Reddit results: since the two data sets produce qualitatively similar results, we conclude that the Reddit data is a broadly accurate representation of contemporary North American English usage.

3.2 Data extraction

We searched for various constructions including one of our seven verbs followed by the preposition *on* or *off*. These two components could be adjacent or separated by a nominal phrase[4] and/or one or more adverbs. In order to distinguish verbal constructions (e.g. *was based on it*, *will profit off it*) from non-verbal constructions (e.g. *a class based on it*, *make a profit off it*), we also tracked instances in which the verb was preceded by an auxiliary like forms of *be*, again with possible adverbs intervening. This allowed us to control for part of speech when modeling (see Section 3.3).

Some of the verbs being studied (*build*, *feed*, and *survive*) are optionally transitive; that is, they can take an overt object in addition to the prepositional object

[4] Nominal phrases could be composed of a stand-alone *pronoun* or a sequence of words centered around a noun, where optional components are in parentheses: *(article/determiner/possessive pronoun) (numeral) (adjective(s)) noun*.

(e.g. *The university has* ***built its reputation on*** *being nontraditional*); however, they most reliably show the desired variation when intransitive. The verb *build* shows *on/off* variation only in its metaphorical meaning, which the OED defines as 'to establish, develop, or construct (something abstract, such as a system of thought or belief, a reputation, a relationship, etc.)' (e.g. *The new law is* ***built on*** *solid legal principles*). However, transitive or passive uses of *build on/off* more often involve physical construction, meaning that false-positive sentences like *The first buildings were* ***built on*** *campus in 1812* are very common, especially on the college subreddits. In contrast, intransitive *build on/off*, as exemplified by (2) above, is exclusively metaphorical. Similarly, intransitive *feed* shows variation in preposition whether metaphorical (as in (4) above) or literal (e.g. *Some birds* ***feed off*** *insects*), while transitive *feed* includes many more irrelevant examples showing no variation: ***feeding my baby on*** *the couch*, ***off*** *my plate*, and so on. Likewise, *survive* can take an object in the meaning desired (e.g. *I* ***survived my pregnancy on*** *plain pasta*), but such cases have higher rates of false positives because the prepositional phrase can be part of the object (like *The king* ***survived the attempt on*** *his life*). To limit ourselves to a consistent construction that yields the most reliable data, we exclude tokens with an intervening object (indicative of a transitive verb) for all verbs except *base* (which is only ever used transitively).

Data from Reddit was retrieved from the Pushshift.io Reddit Corpus (Baumgartner 2019) through ConvoKit (Chang et al. 2020). We used a Python script to search for the sequences described in the previous paragraph. The Reddit Corpus is not lemmatized, so we conducted a string-based search using lists of forms according to their parts of speech in CELEX (Baayen, Piepenbrock & Gulikers 1995).

Thus, for example, hits for the verb *survive* included the words *survive*, *survived*, *survives*, and *surviving*.

This type of search naturally yields false positives. To investigate how many, we looked at a sample of 100 sentences for a number of configurations based on verb form, presence of a direct object (for *base*), and presence of an auxiliary (if a given category had fewer than 100 sentences, we looked at all of them). This sample revealed several frequent undesired prepositional phrases, which we filtered out of our data: *on/off campus* (with up to three words intervening to account for phrases like *on the main campus*, extremely common in college subreddits), *on X's own* (with one word between the preposition and *own*, most common with *survive* and *thrive*), and *on demand/a schedule/a routine* (commonly used to discuss feeding practices in the pregnancy and parenting subreddits).

Configurations that still yielded rates of false positives above 13% in the sample were removed as well. These included:

- *bases* not followed by an object (often nominal: *the **bases on** the baseball field*)
- *building* (often nominal: *a **building on** campus*)
- *built* (often passive: *a community **built on** respect*, *housing **built on** the quad*)
- most forms of *feed* without auxiliaries (often nominal: *a **feed on** YouTube*) or with passive auxiliaries (reliably passive: *the students were **fed on** junk food*)
- *profit or profits* (often nominal: *make a **profit off** his image*) unless preceded by an auxiliary (e.g. *the school doesn't **profit off** of certain classes*)

While the remaining data does still have a small proportion of false positives, we do not think they substantially skew the results. In fact, the results are quite robust to the presence of false positives: earlier versions of the data set, with fewer tokens removed, yielded very similar results.

In COCA, each word is tagged with its lemma and part of speech. Our COCA search included forms of our seven verbs tagged as verbs. While we expected that COCA would have fewer false positives, this turned out to not always be true, and we removed tokens with *on X's own* and several configurations that had false positive rates above 15%. As with the Reddit corpus, some of these had nouns misclassified as verbs (*building*, *profits*). The tagger classifies *fed* as either a past participle or a past-tense form; the former were removed, as they were more often passive. The tagger was not so accurate with *built*, so all sentences with this form were removed, whether it was tagged as a participle or past tense. Sentences with *survived*, *surviving*, and *thriving* were also removed due to false positives stemming from locational prepositional phrases (e.g. *public education is* ***thriving on*** *the West Coast*). Ironically, this seems to be an issue specific to COCA because its texts are *more diverse* than our Reddit data, where many of the locational prepositional phrases for *survive* and *thrive* involved mentions of campus and were thus easy to filter out.

Token counts by corpus and lemma are provided in Table 1.

	base	*build*	*capitalize*	*feed*	*profit*	*survive*	*thrive*	**total**
Reddit	133 675	3497	803	533	242	1498	1252	**141 500**
COCA	41 744	1648	1770	1644	150	355	976	**48 287**

Table 1: Token counts by corpus and lemma.

3.3 Statistical analysis

The data was analyzed using logistic regression in R (R Core Team 2023). For each corpus, we fitted three regressions involving different subsets of verbs and factors to account for the differences between *base* and other verbs: first of all, *base* dwarfs the other verbs in frequency, comprising 94% of tokens for Reddit and 86% for COCA. Second, as described in Section 3.2, *base* is transitive (appearing either with an object or as a passive), while the others are intransitive in our data set. This difference in syntactic construction makes it difficult to compare *base* with the other verbs.

Our dependent variable is preposition, coded as a binary between *on* (marked as 0) and *off* (marked as 1). The sequence *off of* is classified as *off* and is quite common: *off of* constitutes 46% of all *off* tokens in Reddit and 24% of *off* tokens in COCA.

The regressions were fitted using using R's *buildmer* package (Voeten 2023) through forward stepwise comparison; factors were only included in the model if they significantly improved its fit and improved its Akaike Information Criterion (AIC), which penalizes model complexity (additional model factors). Factors that improved the model but were problematic due to sparse data were removed.

The first regression for each corpus looks only at *base*; the second looks at the remaining verbs, which are always intransitive in our data (as discussed in Section 3.2, examples of these verbs followed by direct objects or in the passive were filtered out). The third regression compares the verbal passive construction *be based* (that is, *based* preceded by a passive auxiliary) to the other intransitive verbs. This filtering increases parallels between *base* and the other verbs by removing two configurations in which *base* regularly appears but the other verbs do not: transitive uses in which *base* is separated from its preposition by an overt object (like *the professor* ***based the textbook on*** *his lectures*) and adjectival passives that are not fully verbal in structure (like *a textbook* ***based on*** *the professor's lectures*).

The models included the following key factors:

- Year (centered around the median year with substantial data, 2015 for Reddit and 2005 for COCA)
- Source/genre: college (baseline) vs. pregnancy vs. parenting for Reddit, magazine (baseline) vs. news vs. spoken for COCA
- Verb: not used in first regression (*base* only), sum-coded in second regression (all verbs other than *base*), and dummy-coded in third regression (all verbs, with *base* as baseline)

The most frequent source/genre was chosen as the baseline: about 70% of the Reddit tokens are from college subreddits, while COCA is more balanced: only about 39% of its tokens are from magazines, while 36% are from newspapers. Two-way interaction terms between these three factors were considered as candidates for the models.

A number of morphosyntactic factors were also considered. These factors differed slightly according to the properties of the model, as follows. The regressions for *base* had a candidate factor comparing passives with intervening adverbs to passives with no interveners on the one hand and to actives with intervening objects and, optionally, adverbs on the other. As shown in Table 2, the factors of voice and presence of an intervener are confounded, in that active sentences must have interveners;[5] accordingly, these two factors were combined into a single factor to avoid an unbalanced combination of factors in the models.

	active	passive
no intervener	—	*based on*
intervener	*based it (mostly) on*	*based entirely on*

Table 2: Interaction of voice and presence of intervener for *base*.

The other two regressions for each corpus, which included only verbs without direct objects, had a candidate factor marking the presence of an intervening adverb, again to test for an effect of verb–preposition adjacency.

The regression for verbs other than *base* also included a factor comparing uninflected verb forms (*capitalize*) to third-person singular (*capitalizes*), past/participle (*capitalized*), and progressive (*capitalizing*) forms; this factor was excluded from the regression with all verbs (*base* included) because the form of *base* tokens in this regression was uniformly *based.* Finally, all of the regressions had a candidate factor marking whether the object of the preposition was definite (that is, beginning with *the*).

We did not have any *a priori* hypotheses regarding the effect of these morphosyntactic factors, which we chose because they involve properties of the verb or the relationship between verb and preposition, or, in the case of definiteness of the prepositional object, because this has been shown to have an effect in previous

[5] The corpora do contain a small number of tokens coded as active without interveners. Some of these involve extraction of the object (e.g. *He presented papers that he* ***based on*** *his research*), while many are typos where the last letter of *based* is omitted. All such tokens were removed from our data.

instances of non-phonological variation (Bresnan et al. 2007; Grafmiller & Szmrecsanyi 2018). Although we offer potential *post hoc* explanations of their effects, their primary purpose is to account for potential morphosyntactic confounds to the extrinsic and lexical factors that are the primary object of our study.

Output for all models can be found in the Appendix.

4 Results

We test the following hypotheses on the Reddit data:

1. A real-time shift toward *off*: the proportion of *off* is increasing year-by-year.
2. An apparent-time shift toward *off*: the proportion of *off* hits in the college-age cohort will be higher than that of the pregnancy-age cohort, which will in turn be higher than that of the parent-age cohort.

We find both of these hypotheses to be confirmed in the Reddit data set, which we present first. Afterward, we replicate the real-time trend in the COCA data set, as well as many of the comparisons between *base* and the other verbs. As described in Section 3.1, the COCA data set is smaller but has better tagging and metadata. Replication of the Reddit results in COCA strengthens our confidence that the Reddit data is giving us a real signal despite its shortcomings (e.g. dialectal heterogeneity).

4.1 Reddit

Figure 2 shows the rate of use of *off* (as opposed to *on*), aggregated across all seven verbs studied (*base, build, capitalize, feed, profit, survive, thrive*), over nine years of real time in the Reddit corpus. Though *off* is the minority variant, it shows a steady, linear rise from 7% to 10% over the decade. The effect of year is significant ($p \leq .004$) in all three regressions (*base* alone, all other verbs, all verbs combined; see Tables 4–5 in the Appendix).

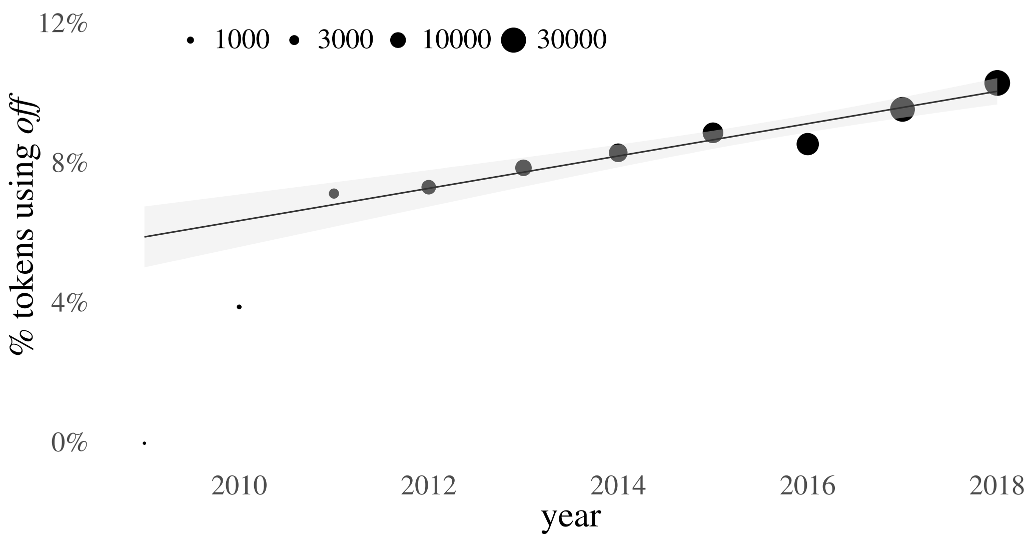


Figure 2: Rate of use of *off* in real time, aggregated over all seven verbs studied, Reddit data.

Figure 3 adds an apparent-time perspective to Figure 2 by plotting the college, pregnancy, and parenting cohorts separately. We see a neat cohort effect: college posters have the highest rate of *off*, parenting posters have the lowest, and pregnancy posters are in the middle. In all three Reddit regression models (Tables 4–6 in the Appendix), the differences between the cohorts are generally significant ($p \leq .04$).[6] Figure 3 suggests that the difference between the college and pregnancy cohorts is equivalent to about five years of real time (about two percentage

[6] The effect of pregnancy in the regression containing intransitive verbs has $p = .04$; all others

points, approximately half the rise shown by the aggregated data over the decade studied), and the difference between the pregnancy and parenting cohorts is similar, if somewhat larger.

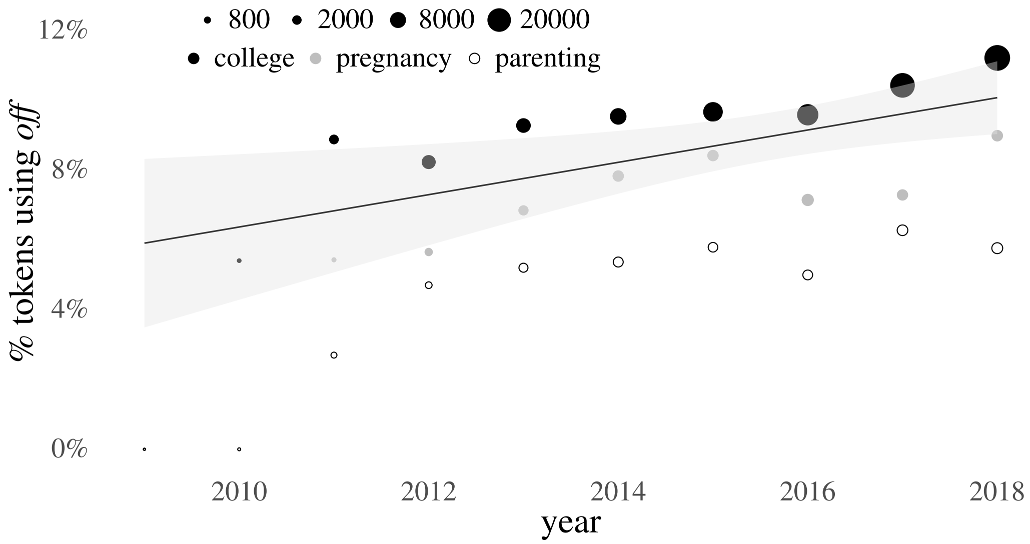


Figure 3: Rate of use of *off* in real time, aggregated over all seven verbs studied, by subreddit type, Reddit data.

The apparent-time effect can also be derived from the regression models. Table 3 shows the coefficients for year and subreddit type for the models based on the Reddit data in Tables 4–6 in the Appendix. The coefficient for year represents the estimated yearly change in use of *off*, while the coefficients for pregnancy and parenting compare those respective cohorts with the baseline, college (in the models, these coefficients are negative, since *off* is used less frequently in these subreddits

have $p < .001$. The estimated marginal means (cf. Lenth 2023) between pregnancy and the other two cohorts in the two regressions containing non-*base* verbs are also not significant, likely due to the presence of an interaction term between verb and subreddit type.

than in college subreddits). Dividing the subreddit type coefficient by the year coefficient thus gives an estimate of the apparent-time effect of subreddit types. Indeed, the first two models yield plausible results, indicating that posters in pregnancy and parenting subreddits are 5–7 and 10–15 years older than posters in college subreddits, respectively. The estimated apparent-time effects of the model comparing passive *based* to the other verbs has a much larger estimated effect (15 and 33 years, respectively); however, this is likely due to the substantially lower coefficient size for year, which in turn may reflect the fact that much of the weight of year in this model is caught up in its interaction with verb. Indeed, removing this interaction term makes the estimated apparent-time effects smaller, though still larger than in the other two models (10 years for pregnancy and 23 for parenting).

model	year	pregnancy	parenting	$\frac{\text{pregnancy}}{\text{year}}$	$\frac{\text{parenting}}{\text{year}}$
base	0.054	0.362	0.818	6.70	15.15
intransitive verbs	0.072	0.394	0.755	5.47	10.49
all verbs					
full model	0.029	0.423	0.962	14.59	33.17
without verb * year	0.041	0.417	0.954	10.04	22.98

Table 3: Absolute value of coefficients for year and subreddit type for Reddit models, with their quotients (interpretable as apparent-time differences).

Finally, Figure 4 shows verb-by-verb data for verbs other than *base*. Since 94% of tokens are *base*, the real- and apparent-time patterns for this verb are largely captured by the aggregated patterns in Figure 3, and including them in Figure 4 would lead to an issue in depicting the differences in scale. At the baseline of 2015, the verb *capitalize* has a significantly lower rate of *off* than *base* ($\beta = -1.55$, $p < .001$), while all other verbs have significantly higher rates of *off* than *base*

($p < .001$), as shown in Table 6 in the Appendix. For *profit*, in particular, the rate of *off* is very high – nearly at ceiling.

In addition, most of the verbs show a similar pattern of real- and apparent-time effects as *base*, though often with more noise: real-time increase, with college posters leading parenting and pregnancy posters. The one main exception is *feed*, where the aggregate real-time line in Figure 4 trends *downward*. Indeed, in the model comparing *base* with other verbs (Table 6 in the Appendix), *feed* is the only verb with a negative interaction with year (though it is not significant). The interaction term ($-.061$) is greater in absolute value than the main effect of year (.029), meaning that the model suggests that use of *off* is decreasing for *feed* year-by-year (not just increasing more slowly than *base*). This downward trend seems to be concentrated in a larger number of tokens of *feed on* than expected in the last couple of years in parenting and pregnancy forums. While we have no explanation for this distribution, we note that these forums include frequent discussion of babies' feeding habits. Many of the false positives (including the common *feed on demand*, see Section 3.2) have been successfully filtered out, but some remain. The relatively small number of tokens and issue with specialized vocabulary mean that this verb's results should be taken with a grain of salt. There is one verb with a significant interaction term with year: the rate of *off* increased significantly more quickly for *survive* than *base*.

Figure 4 also shows that the differences between the subreddit types vary somewhat between the verbs; these are the significant interactions between verb and subreddit type in the models comparing *base* with the other verbs (Table 6 in the Appendix) and the other verbs with each other (Table 5 in the Appendix). In particular, *build* and *thrive* take *off* significantly less in parenting posts than college posts, while *capitalize*, *feed*, and *survive* take *off* significantly more in parenting than college subreddits (for *survive*, the difference is only significant in the model comparing *base* to other verbs). Although these interactions are difficult to interpret, we can speculate on their sources. First, *capitalize* may be a floor effect: the verb is rare and almost always takes *on*, so a small number of tokens of *capitalize off* in parenting posts (6 out of 118) could lead to a higher rate than expected. As explained in the previous paragraph, *feed* has a somewhat anomalous distribution that may lead to unexpected effects. In Figure 4, we see that *build* is disproportionately frequent in college posts, especially in later years when *off* is more frequent in general; the same verb has a steady rate from year to year in parenting posts. This temporal bias towards *build off* in college posts may be driving the significant interaction between *build* and parenting (which does not show up in pregnancy posts because the verb is too rare there). Finally, *survive* and *thrive* make a curious pair: although they are quite similar in meaning, their interaction terms go in opposite directions. The former may be related to the verb's interaction with year described above: the use of *off* increases significantly more rapidly, and in Figure 4, this steeper slope is concentrated in college posts specifically. On the other hand, *thrive* seems to have a similar low use of *off* in parenting forums, but here the verb is *less* frequent in college posts, and thus has a more sporadic distribution.

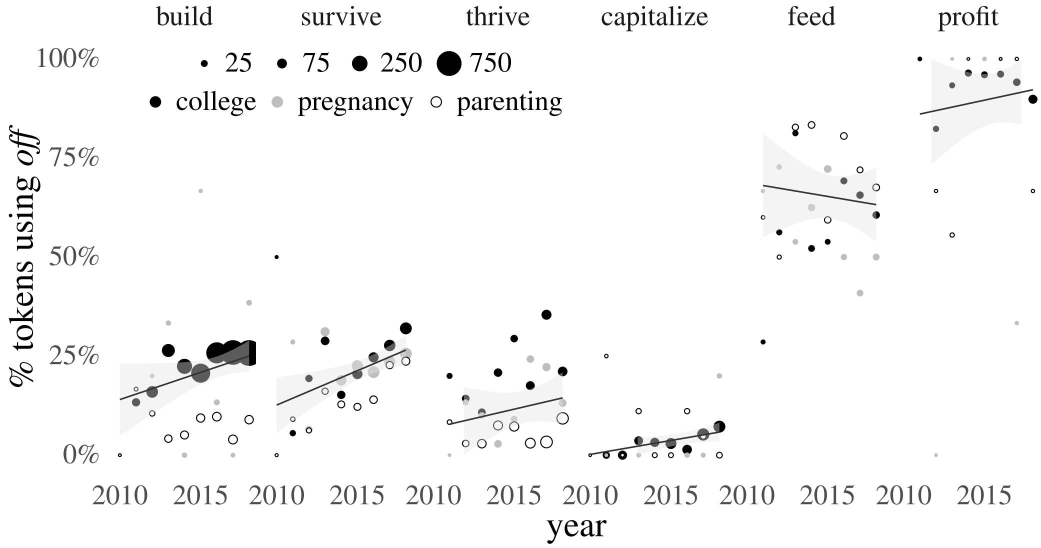


Figure 4: Rate of use of *off* in real time for verbs other than *base*, split by verb (ordered by frequency in the studied corpus) and by subreddit type, Reddit data.

4.2 COCA

Figure 5 shows the rate of *off* (as opposed to *on*), with or without a following preposition, across all verbs in real time from 1990 to 2019 in COCA. Compared to Reddit, *off* is much less common in COCA: even in 2019, the rate of *off* only reaches about 3%, compared to 10% in Reddit. However, there is a clear trend upwards, as *off* appeared well below 1% of the time in 1990. The effect of year is significant ($p < .001$) in all three regressions (shown in Tables 7–9 in the Appendix).

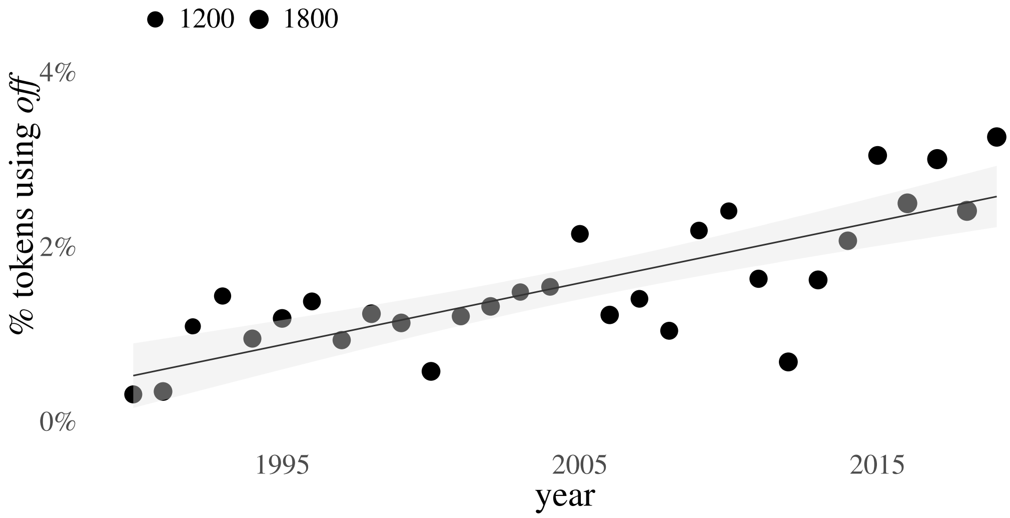


Figure 5: Rate of use of *off* in real time, aggregated over all seven verbs studied, COCA data.

Figure 6 splits the data according to text type: magazines, newspapers, and spoken language. Here we do not see the same stark pattern as in Reddit: the three text types are intermingled and seem quite similar on visual inspection. The statistical models do detect significant differences. In the model limited to *base* (Table 7 in the Appendix), *off* appears in spoken language more often than in magazines ($p < .001$), but there is no significant difference between magazines and newspapers; comparison of estimated marginal means finds a significant difference between spoken language and newspapers ($p < .001$). The difference is equivalent to 5.6 years of real time, given that the coefficient for spoken language is 5.6 times greater than the year coefficient; however, there is no reason to suspect that this corresponds to any apparent-time difference, especially because the difference is not consistent year-over-year as it is with subreddit types in the Reddit data. In the model including verbs other than *base* (Table 8 in the Appendix), both newspapers and spoken language have higher use of *off* than magazines ($p < .001$ for both); in fact, *off* appears *more often* in newspapers than spoken language, though the difference is not significant. However, as we will see below, this effect seems to be located in specific verbs; this model does not include an interaction term between verb and text type because the categorical patterning of *survive* in spoken language (76 tokens, all with *on*) throws off the confidence intervals for all of the spoken-language interaction terms.

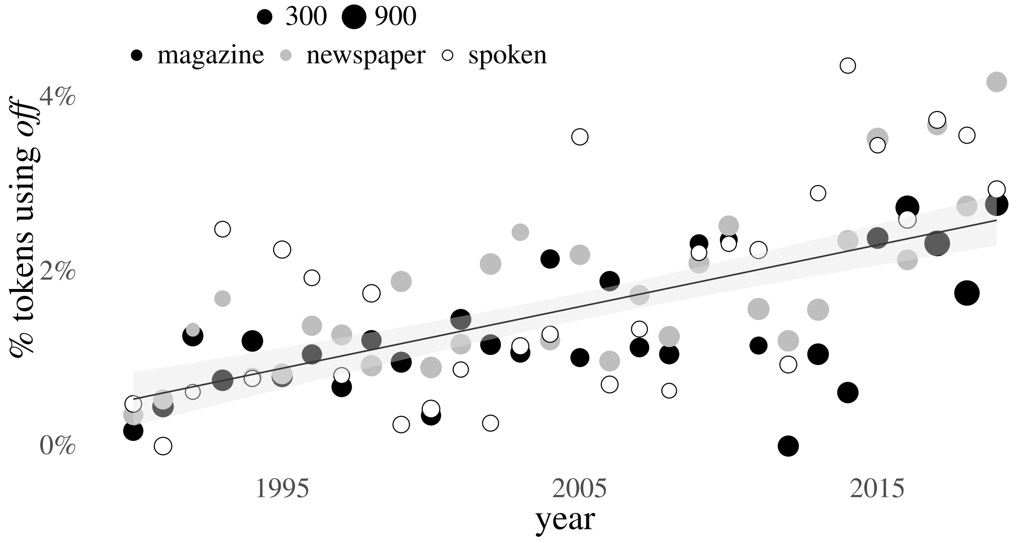


Figure 6: Rate of use of *off* in real time, aggregated over all seven verbs studied, by text type, COCA data.

Finally, Figure 7 shows verb-by-verb data for verbs other than *base*. We see that *capitalize* and *survive* very rarely, if ever, take *off*, while *profit* almost always takes *off*. Meanwhile, *feed* shows a stark genre difference in its use of *off*: magazines have a low rate of *feed off*, while newspapers and especially spoken language show much higher rates. From reading the graph, we would expect that the differences in text type should be concentrated in their interaction with verb, and this is what we see in the regression comparing passive *be based* to the other verbs (Table 9 in the Appendix). According to this model, *off* is used more often in spoken language than magazines, though not quite significantly so ($\beta = .75$, $p = .054$), while there is almost no difference between newspapers and magazines ($\beta = .02$, $p = .960$). However, looking at the interaction term, *feed off* appears significantly and substantially more often in newspapers than in magazines ($\beta = 1.54$, $p = .001$). Inspection of the relevant cases reveals no obvious pattern explaining this effect. The verb *profit* has significant interactions as well: *profit off* appears significantly *less* often in newspapers ($\beta = -2.11$, $p = .023$) and spoken language ($\beta = -2.38$, $p = .008$) than in magazines, in which we find 42 tokens of *profit off* and only two of *profit on*. Finally, *capitalize off* is more common in spoken language than magazines, though the effect does not reach significance ($\beta = 2.06$, $p = .071$).

The verb-by-verb results in COCA are qualitatively similar to those of Reddit: *feed* has a somewhat higher rate of *off* than most of the verbs, suggesting that the high rate of *off* for *feed* is not due to idiosyncrasies of the data source. Likewise, *profit* appears almost entirely with *off*. Since COCA has a much lower rate of *off* in general, the very low rate of *off* for *capitalize* does not stand out from that of the other verbs; its difference from *be based* is not significant. The other verbs appear with *off* significantly more than *be based* ($p \leq .04$).

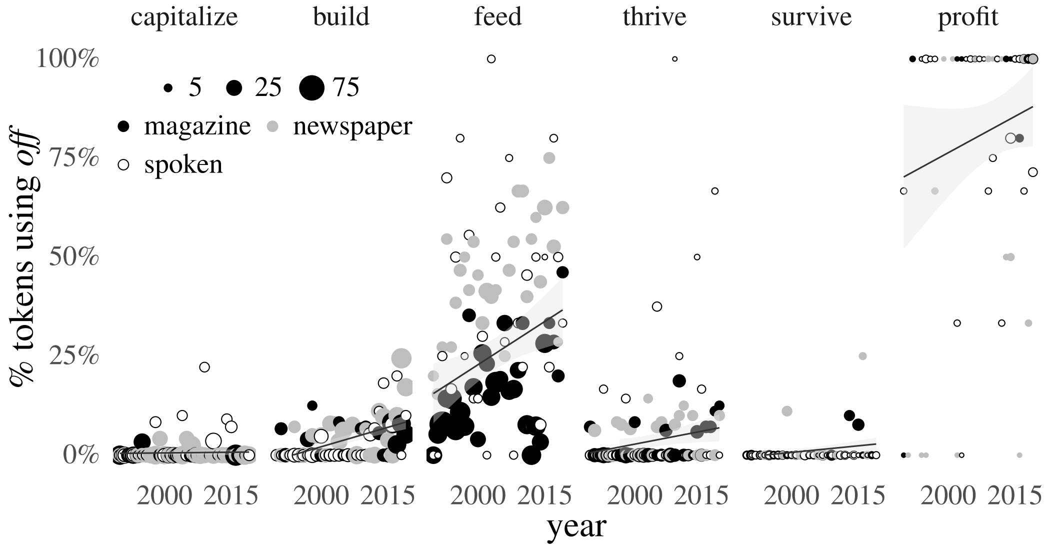


Figure 7: Rate of use of *off* in real time for verbs other than *base*, split by verb (ordered by frequency in the studied corpus) and by text type, COCA data.

4.3 Morphosyntactic factors

In the previous sections, we discussed extrinsic and lexical factors affecting the choice of preposition: real time, subreddit/genre, and verb. Some of our regressions also showed significant morphosyntactic effects: verb form, presence of material intervening between verb and preposition, and whether the prepositional object is definite.

The regression with *base* on Reddit data (Table 4 in the Appendix) showed that active uses of this verb are significantly more likely to take *off* than passive uses, and the difference is very large ($\beta = 1.96$, $p < .001$; for comparison, the effect size of real time is $\beta = .05$ per year). Other morphosyntactic effects did not substantially improve the model and were not added. The *base* model for COCA (Table 7 in the

Appendix) does not include this factor, because it is categorical in the COCA data: *off* never occurs even once in passive uses of *base* with an adverbial intervener, whereas in active uses and in passive uses without an intervener, *off* is merely very rare.

The Reddit regression with the intransitive verbs (that is, all except *base*, Table 5 in the Appendix) shows an effect of verb form: *off* occurs significantly more often with progressive forms (e.g. *surviving*) than uninflected forms (e.g. *survive*; $\beta =$ .56, $p <$ .001). Verb form does not improve the COCA model (Table 8 in the Appendix) and is not added to it.

Another detectable syntactic effect in the regressions including verbs other than *base* is that *off* is used less often when an adverbial intervenes between verb and preposition (e.g. *survives mostly on*) than when verb and preposition are adjacent. This factor is significant in the Reddit model (in Table 6 in the Appendix) comparing *base* to other verbs ($\beta = -.30$, $p < .001$), though it is not added to the Reddit model comparing the non-*base* verbs to one another (Table 5 in the Appendix). It is significant in both of these models for the COCA data (Tables 8 and 9 in the Appendix).

Finally, the two COCA models including verbs other than *base* have one more significant syntactic factor: *off* is used more often when its complement is definite (starts with *the*). This factor is not added to the Reddit models.

4.4 Summary

Our findings can be summarized as follows:

- The *off* variant is steadily increasing in real time, in both corpora. The effect is strongest for *base*, which is the most frequent verb, but is present for others too.

- The Reddit corpus shows an apparent-time increase of the *off* variant as well. While there is some difference in the rate of *off* between different text types in COCA, we do not see an analogous steady, consistent gap.
- Different verbs are at different points in the change toward *off*. In both corpora, *profit* and *feed* take higher rates of *off* than others, with *profit* almost categorically taking *off*.
- The use of *off* is also influenced by internal morphosyntactic factors. Most consistently, *off* is less common when the verb and preposition are separated by an adverb.

5 Discussion

The previous section confirmed that all seven verbs studied here take *off* complements. The general trend, with few exceptions, is that all verbs are changing toward use of *off* in both real and, where data is available, apparent time.

5.1 Main findings

There are two questions we want to address here. The first is: why is this change happening? The second is: why is it happening in these verbal constructions specifically? After all, *off* is not replacing *on* across the English language in general: not when *on* is used with its core physical meaning, nor when it is used in other metaphorical ways, such as *airing on television* or *kept on file*.

To answer the first question, we turn to a suggestion from Janda (2020). In explaining the rise of *based off (of)*, Janda proposes:

> [I]f one derives something from a source, then a crucial pathway between them leads from the source to the derivative; something takes off from – or is taken off (of) – the source and travels – or is brought – to/as the derivative. [. . .] 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. (597)

In other words, *off* suggests extraction, while *on* suggests foundation. Perhaps, then, the shift from *on* to *off* is a change from a metaphor of foundation to one of extraction. That, then, suggests an answer to the second question: the verbs undergoing this change are those that are compatible with this 'extraction' meaning.

Somewhat speculatively, we observe that there may be a correlation between the strength of a verb's association with these meanings of extraction and/or foundation and its likelihood of change. We found in both corpora that *profit* had the highest rates of *off* by far: even going back to the 1990s in the COCA data, *profit off* well exceeds 50% *off* usage (19/25 tokens in that decade). Data from the Google Books Ngram Corpus from 2019, plotted in Figure 8, likewise show that *profit off* started gaining ground in the 1990s. This suggests that *profit* was an earlier shifter than the other verbs, not categorically different from them. Indeed, the lexical semantics of *profit* are particularly well-suited for a metaphor of extraction.

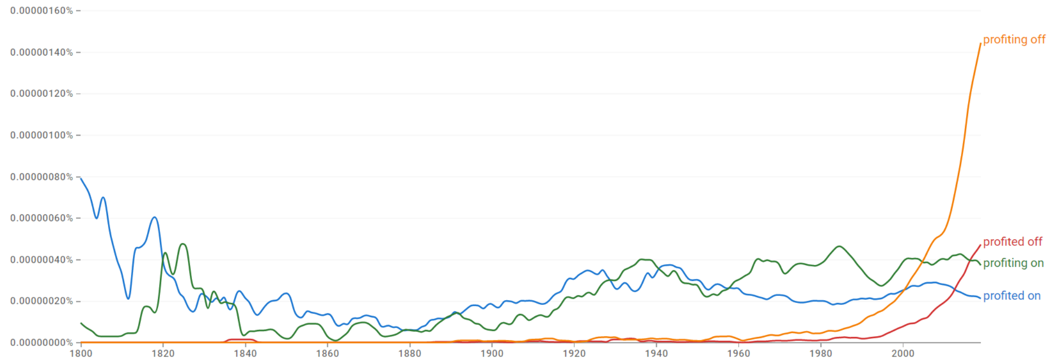


Figure 8: Rates of *on* and *off* with forms of *profit* in Google Book Ngram data from 2019.

By contrast, *base* shows lower rates of *off* than any other verb except *capitalize*, a pattern that holds in both corpora. Again, the lexical semantics of *base* are particularly compatible with a metaphor of foundation, perhaps leading it to have resisted shifting longer than the others. (On the other hand, the even lower rate of *off* with *capitalize* may be a reflection of its belonging to a higher register.)

We note also that *base* is by far the most frequent verb in both data sets. The resistance of highly frequent forms to change is well known from work on morphological changes such as analogical leveling (e.g. Hooper 1976). Still, despite its overall low rate of *off*, *base* is making up for lost time: in COCA, at least, it is changing toward *off* faster than all the other verbs in our study (though this difference is not significant for two verbs).

All told, the picture here suggests a change that has started in different verbs at different times, and is progressing for different verbs at different rates. This is reminiscent of lexical diffusion in phonology: a change starts in one environment (in this case, potentially *profit*) and then gradually expands to others (Wang 1969).[7]

[7] This means that the change studied here shows a different pattern than the syntactic changes

The lexically specific nature of this change raises questions about whether any social evaluation associated with the change is similarly lexically specific. Behrens (2014), as cited in Section 1, notes that *based off* is salient enough that some speakers notice it and prescriptively judge it as incorrect; references to and expressions of negative evaluations of the shift can also be found on usage guides (e.g. Merriam-Webster n.d.) and blog posts and comments (e.g. Jerz 2014). It is likely that *base* has attracted overt comment due to its high frequency and its rapid rate of change. But given that the other six verbs studied here are also changing in the same direction, do they share the same social evaluation? The question of whether the social evaluation of a variant extends to all environments in which that variant surfaces, or whether social evaluation may interact with internal (linguistic) constraints on variation, is longstanding in sociolinguistics, going back at least to Weiner & Labov (1983). This could be examined in follow-up work on this change.

studied by Kroch (1989), in which a change is initiated at the same time in all contexts in which it occurs, and progresses at the same rate in all of them (the "Constant Rate Effect"). However, we think that the change toward *off* is not a counterexample to the Constant Rate Effect: it is a different type of change. Constant Rate-type changes reflect "a single underlying change in grammar" that can be seen simultaneously in multiple environments (Kroch 1989: p. 199). But in the same way that regular Neogrammarian changes can coexist with lexically diffused changes in phonology (Labov 1981), we believe Constant Rate–type changes can coexist with lexically diffused morphological changes like the one we document here. No single "abstract grammatical option" (to use Kroch's [1994] terminology) underlies the choice between *on* and *off*; instead, the choice of word, we suggest, is diffusing lexically from one context to another, and, as such, the quantitative patterns of change can diverge across contexts.

5.2 Text sources and apparent time

Both the Reddit and the COCA corpus are divided up into three different text sources. In the former, examples were drawn from subreddits themed around college, pregnancy, and parenting. We use this as a proxy for apparent time, under the assumption that posters on college subreddits are younger (and thus more advanced in the change) than posters on pregnancy subreddits, who are in turn younger than people posting about parenting. This assumption was borne out: qualitatively, Figure 3 shows a fairly consistent difference between the three subreddit types that looks equivalent to the difference of a few years; this impression is largely confirmed quantitatively as well by the models, as shown in Table 3.

The COCA corpus included texts from magazines, newspapers, and spoken language. This difference, in contrast, is not expected to correspond to apparent time, since it reflects differences in the mode of production rather than the demographics of people producing the texts. Indeed, Figure 6 presents a stark contrast to the Reddit data in Figure 3: the rates of *off* among the different text types are quite intermingled and close together, and certainly do not show the lockstep pattern of the Reddit data. The COCA models generally show that *off* is more common in spoken language than magazines – which is understandable, given that the latter likely go through more editing. However, the lack of a sharp cohort effect in COCA further reinforces the apparent-time interpretation of the Reddit data.

5.3 Morphosyntactic factors

The regressions showed evidence of a number of internal grammatical factors influencing the use of *off* – in particular, *off* is used less frequently when an adverb or adverbial phrase intervenes between the verb and the preposition (like *survive almost entirely on*). This effect is generally driven by verbs other than *base*, and seems to split into two: first, interveners are much less frequent, so low-frequency verbs like *capitalize* never appear with *off* and interveners. Second, verbs with higher rates of *off*, especially *feed*, have lower rates of *off* with interveners. There is no obvious explanation for this latter effect. If the shift in prepositions is lexically driven – that is, mediated by each individual lexical item – then intervening material between verb and preposition may make the ties between the two weaker and cause reversion to a default preposition – which, at least for now, is more likely to be *on*. A similar explanation may explain the voice difference for *base*: in the Reddit data, *off* is more frequent in active forms like *based it (mostly) off* than in passive forms with interveners like *was based mostly off*. For *base*, the passive form is much more common than the active, and the latter may be treated by speakers as having different lexical properties including a higher rate of *off*, untethered as it is from the extremely frequent construction *based on/off*.

6 Conclusion

The purpose of this study was twofold. First, we aimed to capture an in-progress shift in the prepositional complement of verbs like *base* from *on* to *off*. While this change has been previously documented for both *base* and other verbs, this is the first study that systematically investigates the change in progress across multiple verbs, text types, and morphosyntactic contexts. Our results are clear: *off* is used across many different verbs, and its use is increasing in both real and apparent time. Moreover, this change shows no sign of stopping and looks to be picking up other verbs in its path as well: examples of *off* can be found even with verbs like *depend* and *rely*, which the authors of this paper, who generally accept the tokens in our corpora and likely produce *off* fairly regularly as well, find crashingly bad.[8] Thus, our study lays important groundwork for future study of this linguistic variable, including both sociolinguistic factors that we did not study systematically (region, gender, etc.) and a closer look at internal linguistic factors, including those we studied and those we did not. For example, one direction for future research is to investigate geographical patterning of this variation in a large data set with geographical metadata, such as a Twitter-based corpus.

The second main purpose of our study was methodological. Reddit represents an enormous body of informal text that could serve as a valuable resource for socio- and other linguistic research. However, its users are anonymous and we typically have no demographic information about them (though see Flesch 2019). Thus, we use this study as a proof of concept for the efficacy of the Reddit corpus. Its results are qualitatively similar to those of a more cultivated corpus, COCA: even though

[8] Examples, taken from outside the subreddits studied in this paper, are shown below.

(i) Edit: forgot to mention on the TooGoodToGo app you can get a bunch of bagels for $3.99 **depending off** the bagel shop. There are also many other cool findings, so you should check it out. (r/AskNYC)

(ii) They **relied off of** my written statement more than anything, because I have issues talking about any of it, so the statement will be important if you have issues talking about it as well. (r/Veterans)

The comment history of the users who made these posts suggests that they are native English speakers in the United States.

some of the tokens in the Reddit corpus were undoubtedly made by non-native speakers, it is reliable enough to display broad trends. Moreover, the organization of Reddit into subreddits, which are often very specific, allows us to approximate its users' demographic properties – in this case, age. This analytical move was successful: the effects of subreddit were interpretable in terms of time and yielded sensible results well within the range of our expectations from the inferred demographic correlates of subreddit. Thus, we hope that this study will serve as inspiration for future use and exploration of Reddit as a source of sociolinguistic data, both general and demographically specific.

APPENDIX

In Tables 5 and 8, the factor representing the six verbs is sum-coded: the five factors compare each of the first five verbs to the grand mean (the mean of the means of the dependent variable – in this case, likelihood of *off* – for each verb). The estimate for the sixth verb, *profit*, is the negative sum of all five factors. Thus, in the tables below, we provide an estimate for *profit* (the negative sum of the five factors) but not a standard error or p value, since it is not represented by a separate factor in the model. Each of the five factors, in turn, includes some influence of *profit* in addition to the listed verb. Thus, the estimates for sum-coded factors are rather easier to interpret than their corresponding standard error and p values.

Terms are listed in the order in which they are added to the model, roughly corresponding with importance. Low p values are marked as follows: *** for $p <$ .001, ** for $.001 \leq p < .01$, * for $.01 \leq p < .05$, . for $.05 \leq p < .1$.

	β	SE	p
Intercept	–2.44	.07	<.001 ***
Voice and intervener (default: passive, intervener)			
active, intervener	1.96	.07	<.001 ***
passive, no intervener	–0.03	.07	.694
Subreddit type (default: college)			
pregnancy	–0.36	.03	<.001 ***
parenting	–0.82	.04	<.001 ***
Year	0.05	.01	<.001 ***

Table 4: Coefficients for model with all tokens of *base*, Reddit data.

	β	SE	p
Intercept	–0.75	.08	<.001 ***
Verb (compared to grand mean)			
build	–0.46	.08	<.001 ***
survive	–0.47	.10	<.001 ***
thrive	–0.52	.12	<.001 ***
capitalize	–2.66	.18	<.001 ***
feed	1.17	.15	<.001 ***
profit	2.93	—	—
Subreddit type (default: college)			
pregnancy	–0.39	.19	.039 *
parenting	–0.75	.14	<.001 ***
Verb form (default: base)			
third-person singular	–0.01	.07	.844
progressive	0.56	.12	<.001 ***
past	–0.07	.14	.639
Year	0.07	.01	<.001 ***
Verb * Subreddit type (compared to grand mean * college)			
build * pregnancy	0.29	.32	.365
survive * pregnancy	0.18	.22	.430
thrive * pregnancy	–0.24	.27	.378
capitalize * pregnancy	0.51	.64	.432
feed * pregnancy	0.11	.27	.674
profit * pregnancy	–0.85	—	—
build * parenting	–0.61	.21	.004 **
survive * parenting	0.23	.21	.269
thrive * parenting	–0.88	.22	<.001 ***
capitalize * parenting	1.09	.41	.007 **
feed * parenting	1.14	.23	<.001 ***
profit * parenting	–0.97	—	—

Table 5: Coefficients for model with intransitive tokens of verbs other than *base*, Reddit data.

	β	SE	p
Intercept	−1.86	.02	<.001 ***
Verb (default: base)			
build	0.66	.06	<.001 ***
survive	0.72	.11	<.001 ***
thrive	0.57	.13	<.001 ***
capitalize	−1.55	.24	<.001 ***
feed	2.34	.16	<.001 ***
profit	4.40	.27	<.001 ***
Subreddit type (default: college)			
pregnancy	−0.42	.07	<.001 ***
parenting	−0.96	.08	<.001 ***
Year	0.03	.01	.004 **
Intervener (default: no)			
yes	−0.30	.08	<.001 ***
Verb * Subreddit type (default: base * college)			
build * pregnancy	0.31	.32	.341
survive * pregnancy	0.28	.15	.064 .
thrive * pregnancy	−0.15	.24	.545
capitalize * pregnancy	0.56	.76	.464
feed * pregnancy	0.29	.24	.220
profit * pregnancy	−0.90	.72	.207
build * parenting	−0.42	.21	.050 *
survive * parenting	0.44	.20	.031 *
thrive * parenting	−0.69	.23	.003 **
capitalize * parenting	1.29	.48	.007 **
feed * parenting	1.49	.24	<.001 ***
profit * parenting	−0.71	.58	.221
Verb * Year (default: base)			
build * year	0.03	.03	.180
survive * year	0.07	.03	.044 *
thrive * year	0.09	.05	.055 .
capitalize * year	0.18	.10	.068 .
feed * year	−0.06	.05	.193
profit * year	0.01	.11	.920

Table 6: Coefficients for model with passive tokens of *base* and intransitive tokens of other verbs, Reddit data.

	β	SE	p
Intercept	−7.13	.24	<.001 ***
Year	0.17	.02	<.001 ***
Text type (default: magazine)			
newspaper	0.06	.25	.826
spoken	0.94	.22	<.001 ***

Table 7: Coefficients for model with all tokens of *base*, COCA data.

	β	SE	p
Intercept	−3.60	.16	<.001 ***
Verb (compared to grand mean)			
capitalize	−2.50	.28	<.001 ***
build	−1.01	.20	<.001 ***
feed	1.95	.15	<.001 ***
thrive	−0.59	.20	.004 **
survive	−1.98	.56	<.001 ***
profit	4.13	—	—
Text type (default: magazine)			
newspaper	1.23	.11	<.001 ***
spoken	0.95	.14	<.001 ***
Year	0.07	.02	<.001 ***
Prepositional object (default: indefinite)			
definite	0.51	.11	<.001 ***
Intervener (default: no)			
yes	−1.38	.41	.001 ***
Verb * Year (compared to grand mean)			
capitalize * year	−0.05	.03	.121
build * year	0.04	.02	.054 .
feed * year	−0.03	.02	.051 .
thrive * year	0.00	.02	.889
survive * year	0.05	.06	.393
profit * year	−0.01	—	—

Table 8: Coefficients for model with intransitive tokens of verbs other than *base*, COCA data.

	β	SE	p
Intercept	−6.67	.38	<.001 ***
Verb (default: base)			
capitalize	−0.03	1.07	.974
build	2.51	.48	<.001 ***
feed	4.89	.39	<.001 ***
thrive	2.79	.49	<.001 ***
survive	1.91	.94	.042 *
profit	9.49	.84	<.001 ***
Text type (default: magazine)			
newspaper	0.02	.43	.960
spoken	0.75	.39	.054 .
Year	0.16	.03	<.001 ***
Intervener (default: no)			
yes	−1.53	.42	<.001 ***
Prepositional object (default: indefinite)			
definite	0.48	.11	<.001 ***
Verb * Text type (default: base * magazine)			
capitalize * newspaper	0.96	1.23	.434
build * newspaper	0.77	.52	.140
feed * newspaper	1.54	.45	.001 ***
thrive * newspaper	0.35	.61	.565
survive * newspaper	0.36	1.10	.741
profit * newspaper	−2.11	.93	.023 *
capitalize * spoken	2.06	1.14	.071 .
build * spoken	−0.53	.56	.344
feed * spoken	0.30	.42	.487
thrive * spoken	0.60	.59	.311
survive * spoken	−12.76	259.81	.960
profit * spoken	−2.38	.90	.008 **
Verb * Year (default: base)			
capitalize * year	−0.14	.05	.002 **
build * year	−0.05	.03	.109
feed * year	−0.12	.03	<.001 ***
thrive * year	−0.09	.04	.012 *
survive * year	−0.04	.09	.658
profit * year	−0.12	.04	.002 **

Table 9: Coefficients for model with passive tokens of *base* and intransitive tokens of other verbs, COCA data.

REFERENCES

Baayen, R. H., R. Piepenbrock & L. Gulikers. 1995. *CELEX2 LDC96L14*. Web Download. Philadelphia: Linguistic Data Consortium.

Barratt, Leslie. 2006. What speakers don't notice: Language changes can sneak in. *TRANS. Internet-Zeitschrift für Kulturwissenschaften* 16.

Baumgartner, Jason. 2019. Reddit corpus. https://files.pushshift.io/reddit/ (16 July, 2024).

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire & Jeremy Blackburn. 2020. The Pushshift Reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 830–9.

Behrens, Susan J. 2014. *Understanding language use in the classroom: A linguistic guide for college educators*. Toronto: Multilingual Matters.

Behrens, Susan J. & Cindy Mercer. 2007. The style of which this is written: Neutralization of prepositions in English. *NADE Digest* 3(2), 47–58.

Bermel, Neil & Luděk Knittl. 2012. Morphosyntactic variation and syntactic constructions in Czech nominal declension: Corpus frequency and native-speaker judgements. *Russian Linguistics* 36(1), 91–119.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Brook, Marisa & Emily Blamire. 2023. Language play is language variation: Quantitative evidence and what it implies about language change. *Language* 99(3), 491–530.

Chang, Jonathan P., Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang & Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt & Stefan Ultes (eds.), *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 57–60. Association for Computational Linguistics.

Curzan, Anne. 2013. Based off of what? https://www.chronicle.com/blogs/linguafranca/based-off-of-what (6 August, 2024).

Davies, Mark. 2008–. The Corpus of Contemporary English (COCA). https://www.english-corpora.org/coca/ (16 July, 2024).

Ding, Karl. 2018. GitHub - karlding/college-subreddits. https://github.com/karlding/college-subreddits (10 June, 2021).

Estling, Maria. 1999. Going out (of) the window? *English Today* 15(3), 22–7.

Estling, Maria. 2000. Competition in the wastebasket: A study of constructions with *all*, *both* and *half*. In Christian Mair & Marianne Hundt (eds.), *Corpus linguistics and linguistic theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*, 103–16. Amsterdam & Atlanta: Rodopi.

Flesch, Marie. 2019. “That spelling tho”: A sociolinguistic study of the nonstandard form of *though* in a corpus of Reddit comments. *European Journal of Applied Linguistics* 7(2), 163–88.

Grafmiller, Jason & Benedikt Szmrecsanyi. 2018. Mapping out particle placement in Englishes around the world: A study in comparative sociolinguistic analysis. *Language Variation and Change* 30(3), 385–412.

Hooper, Joan B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In William M. Christie Jr. (ed.), *Current progress in historical linguistics: Proceedings of the Second International Conference on Historical Linguistics, Tucson, Arizona, 12–16 January 1976*, 96–105. Amsterdam: North Holland.

Iyeiri, Yoko, Michiko Yaguchi & Hiroko Okabe. 2004. *To be different from* or *to be different than* in present-day American English? *English Today* 20(3), 29–33.

Janda, Richard D. 2020. Perturbations, practices, predictions, and postludes in a bioheuristic historical linguistics. In Richard D. Janda, Brian D. Joseph & Barbara S. Vance (eds.), *The handbook of historical linguistics*, vol. II, chap. 24, 523–650. Hoboken, NJ: Wiley Blackwell.

Jerz, Dennis J. 2014. Does the phrase "based off of" make you shudder... or shrug? https://jerz.setonhill.edu/blog/2014/12/29/does-the-phrase-based-off-of-make-you-shudder-or-shrug/ (7 November, 2024).

Kiefer, Ferenc. 1985. The possessive in Hungarian: A problem for natural morphology. *Acta Linguistica Academiae Scientiarum Hungaricae* 35(1–2), 85–116.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1(3), 199–244.

Kroch, Anthony. 1994. Morphosyntactic variation. In *Proceedings of the 30th Annual Meeting of the Chicago Linguistics Society*, 180–201.

Labov, William. 1981. Resolving the Neogrammarian controversy. *Language* 57(2), 267–308.

Lenth, Russell V. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.9.0. https://CRAN.R-project.org/package=emmeans.

Mair, Christian. 2007. British English/American English grammar: Convergence in writing–divergence in speech? *Anglia* 125(1), 84–100.

Merriam-Webster. n.d. Is it 'based on' or 'based off'? https://www.merriam-webster.com/grammar/based-on-vs-based-off (7 November, 2024).

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak & Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331(6014), 176–82.

Nylund, Anastasia & Corinne Seals. 2010. "It's not that big (of) a deal": The sociolinguistic conditioning of inverted degree phrases in Washington, DC. *University of Pennsylvania Working Papers in Linguistics* 16(2), 133–40.

Oxford University Press. 2023. Base (v.3). In *Oxford English dictionary*. https://www.oed.com/dictionary/base_v3 (12 November, 2024).

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna. https://www.R-project.org/.

Schlüter, Julia. 2022. Language corpora and the teaching and learning of English as an international language. In Marcus Callies & Stefanie Hehner (eds.), *Pluricentric languages and language education: Pedagogical implications and innovative approaches to language teaching*, 166–89. London: Routledge.

Tabachnick, Guy. 2023. *Morphological dependencies*. New York University dissertation.

Traugott, Elizabeth Closs. 2004. A critique of Levinson's view of Q-and M-inferences in historical pragmatics. *Journal of Historical Pragmatics* 5(1), 1–26.

Vartiainen, Turo & Mikko Höglund. 2020. How to make new use of existing resources: Tracing the history and geographical variation of *off of*. *American Speech* 95(4), 408–40.

Voeten, Cesko C. 2023. *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. R package version 2.9. https://CRAN.R-project.org/package=buildmer.

CH-Wang, Sky & David Jurgens. 2021. Using sociolinguistic variables to reveal changing attitudes towards sexuality and gender. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia & Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9918–38.

Wang, William S-Y. 1969. Competing changes as a cause of residue. *Language* 45(1), 9–25.

Weiner, E. Judith & William Labov. 1983. Constraints on the agentless passive. *Journal of Linguistics* 19(1), 29–58.