Axel Bohmann, Julia Müller, Mirka Honkanen,
Miriam Neuhausen

A Large-scale Diachronic Analysis of the English Passive Alternation

Abstract: We present the first large-scale, multivariate study analysing the development of the passive alternation in 19th- and 20th-century American English. Based on 2,318,251 tokens of the BE- and the GET-passive, extracted from the Corpus of Historical American English, we explore the strength and stability of several reported constraints on the GET-passive, such as informality, subject responsibility, adversativity, and non-neutrality. Additionally, our analysis includes a range of syntactic predictors. The results indicate a persistent association of the GET-passive with informal contexts, but weakening of most other constraints. A particularly strong effect size is observed for the semantic group of the passivized verb, developed by clustering over a word-vector representation of all verbs. This finding indicates strong lexical-semantic conditioning of the passive alternation. We discuss several challenges in the big-data approach we use and develop a sketch of future research in this direction.

1 Introduction

The past two centuries have seen an increase in the use of GET to form passive sentences in the English language, especially in American English (Hundt 2001), and an even more dramatic drop in the frequency of the traditional BE-passive (Mair/Leech 2006). This development has been referred to as "one of the most active grammatical changes taking place in English" (Weiner/Labov 1983: 43), but its precise motivations are still not fully understood. Mair/Leech (2006: 332) explain the increase in GET at the expense of BE as part of the wider trend of "colloquialization", whereby writing adopts features of spoken language.

The two passive variants BE and GET are not always interchangeable (Xiao/McEnery/Qian 2006). In the rich previous literature, scholars have suggested differences in the semantics of the two passives, such as implications of adversativity (Chappell 1980), informality (Biber/Conrad/Leech 2003: 112), and agentivity (Toyota 2008: 157) for the GET-construction, as well as lexically conditioned preferences (Rühlemann 2007). Schwarz (2015; 2017) speculates that the restrictions on the GET-passive might be weakening over time, as part of its increasing "grammaticalization" (Hundt 2001; Hopper/Traugott 2003). Schwarz does not,

https://doi.org/10.1515/9783111017433-003

however, find evidence of this, nor does she, in fact, detect clear semantic differences between the two passives; subsequently, she encourages research to "focus on finding the factors that encourage the choice of GET" (2015: 166).

Due to the relative infrequency of passivizing GET and the difficulty of distinguishing it from other formally similar constructions, most previous research on the passive alternation has relied on close readings of constructed examples (e.g., Hatcher 1949; Chappell 1980) or considered bivariate distributional patterns in various digitized corpora (e.g., Collins 1996; Xiao/McEnery/Qian 2006; Coto Villalibre 2015). However, only few studies on the passive have adopted a multivariate statistical analysis, which is able to isolate and quantitatively compare the influence of competing factors. Given the semantic and stylistic nuances of the alternation, such a study on a large scale is needed to tease apart the various constraints on the choice of passive auxiliary.

In this paper, we track changes in passive auxiliary choice in a large corpus of written American English over nearly two centuries, investigating the influence of various semantic, textual, and syntactic factors on this variable. We employ automated sentiment analysis, distributional semantics, and a mixed-effects regression model to provide the first preliminary answers to our research questions:

1. How has the use of the BE- and GET-passive constructions with different lexical verbs changed between 1830–2009?
2. Are the alleged connotations of the GET-passive (informality, subject responsibility, adversativity/non-neutrality) empirically verifiable and historically stable?

2 The English Passive Alternation

The English language has two competing passive constructions. The canonical BE + past participle (hereafter, "the BE-passive"; (1)) varies with the newer variant with GET as the auxiliary verb ("the GET-passive"; (2)). In both cases, the affected patient acts as the syntactic subject, while the agent may be included in a *by*-prepositional phrase (PP) or omitted altogether.

(1) The burglar was arrested (by the police).

(2) The burglar got arrested (by the police).

The two constructions are, however, not always interchangeable, and a number of distinct syntactic, semantic, and pragmatic constraints have been suggested for

each. The BE-passive has been found to be much more frequent, to be favoured particularly in written genres, and more likely to occur with an overt agent in the form of a *by*-phrase (Xiao/McEnery/Qian 2006).

The GET-passive tends to emerge in informal contexts "with meanings connected with speaker attitude, judgment, and affective posture" (Carter/McCarthy 1999: 51). This is a frequently supposed semantic-pragmatic characteristic of the GET-passive (Biber/Conrad/Leech 2003; Xiao/McEnery/Qian 2006). The construction is further associated with subject responsibility, which refers to the idea that the passive subject is somehow responsible for the situation being brought about on themselves (Chappell 1980; Coto Villalibre 2015; Toyota 2008). Furthermore, it is often claimed that the situations the GET-passive encodes tend to have either adversative (Rühlemann 2007; Chappell 1980; Carter/McCarthy 1999; Toyota 2008) or more generally non-neutral (fortunate or unfortunate) consequences for the subject (e.g., Hatcher 1949; Fleisher 2006). By contrast, Coto Villalibre's (2015) findings indicate that the majority of GET-passives are semantically neutral. Consequently, he suggests that GET-passives are either now converging with the neutral BE-passive or adversativity was solely a "contextual feature" in the first place and not a defining property of the construction itself (Coto Villalibre 2015: 24). Furthermore, the GET-passive encodes only dynamic situations as opposed to the BE-passive (e.g., Xiao/McEnery/Qian 2006), which can be either dynamic or stative (Toyota 2008: 149). The semantic difference between the stative and dynamic variants can serve to avoid potential misinterpretations. The semantic ambiguity of (3), for example, is absent in the GET-variant (4) (Quirk et al. 1985: 162).

(3) The chair was broken.

(4) The chair got broken.

Notably, not all instances of BE/GET + past participle carry a true passive meaning. There are formally similar structures where the participle "has both adjectival and verbal properties" (5), or is fully stative and adjectival (6) (Quirk et al. 1985: 169–170). Moreover, GET + past participle may also represent the "middle voice", where the subject is "both the controller of the action and affected by it" (7) (Hundt 2001: 51; Croft 1991: 248).

(5) Wordsworth said he got so maddened by the sight of it that he threw up the job. (COHA fiction 1970).

(6) The board can get very excited about building, and there's a lot of energy around it. (COHA news 2007).

(7) The right has also had trouble getting organized for next spring's presidential elections. (COHA fiction 1917).

The distinction between central passives as in (1) and (2) and such more peripheral cases as in (5)–(7) has led scholars to postulate a "passive gradient" (Quirk et al. 1985: 167).

Given the fine-grained semantic-stylistic differences between the GET- and the BE-passive as well as the fact that they co-exist with formally identical non-passive constructions, circumscribing the variable context for this alternation is no easy task. Trying to ensure the referential equivalence of any attested form with its competing variant, most empirical studies to date have qualitatively examined individual instances (e.g., Collins 1996; Xiao/McEnery/Qian 2006; Rühlemann 2007; Coto Villalibre 2015; Schwarz 2015; 2017). Even though this practice may be effective in guaranteeing accountability, it entails two major problems.

Firstly, formal tests for passive centrality often rely on the (in)acceptability of constructed modifications to an attested candidate sentence. These include, for example, whether the auxiliary could be replaced by its competitor, or whether a corresponding active sentence can be formed. Yet, acceptability judgments in relation to the GET-passive have been subject to debate and change over even the past four decades. For example, Banks (1996: 127) claims utterance (8) to be "of doubtful acceptability", but a search in COHA yields 1,352[1] hits for the underlying structure GET + participle + *by*. For a diachronic study in particular, this means that a stable point of reference for such judgments is difficult to establish.

(8) Mary got shot by John.

Secondly, in more practical terms, qualitative analysis of all tokens is costly and has therefore often been restricted to hundreds or thousands of cases. All the same, the passive alternation shows a number of characteristics that call for a larger quantitative approach. For instance, the distribution of the two variants is heavily imbalanced, with the BE-passive outnumbering the GET-passive by ratios between 10:1 (Xiao/McEnery/Qian 2006) and 100:1 (this study). This means that, for several thousand tokens considered, the analysis may only be able to provide insight into a handful of GET-passive cases. Such a limitation is unfortunate,

1 This is the figure the COHA online interface gave us at the time of writing the article. However, it does not seem to be stable either across time or the different search options of COHA. Search on 3 Nov 2022 yields 941 hits for the same query ("GET _v?n by") in the list display and 1,757 in the chart display.

particularly in this case, where numerous constraints at various levels (semantic-pragmatic, stylistic, syntactic) have been put forward. In order to effectively consider their relative importance, a multivariate analysis on a large scale is required. Below, we detail the methodological steps and initial findings of such a study.

3 Data and Methods

3.1 Data Collection

Our investigation relies on a large diachronic corpus, the Corpus of Historical American English (COHA; Davies 2010), which contains ca. 400 million words from U.S. newspapers and magazines as well as fiction and non-fiction books from the 19th and 20th centuries. For the time period from 1830 to 2009, we investigated 398.1 million lemmatized and part-of-speech-tagged tokens. We downloaded the data and wrote a Python (Python Software Foundation 2019) script to automatically extract all instances of lemma BE/GET + past participle, as tagged in COHA. We also included intervening adverbs and negators.

An accountable study of any alternation requires careful definition of the variable context (Poplack/Tagliamonte 1989). Ideally, only cases where both variants can be used interchangeably should be considered, and those that permit only one of the variants under discussion should be excluded. Similarly, all extracted forms should be genuine instances of the variable under investigation. The nature of the passive alternation as well as the number of observations in our study impose several difficulties in this regard. As there is no direct annotation for passive voice constructions in COHA, the automated search has to rely on the lemma/part-of-speech tagging in the corpus. In addition to tagging errors, constructions on the "passive gradient" (Quirk et al. 1985: 167; Collins 1996) exist where both GET and BE may be used, but where the choice between the two entails a semantic difference, as that between (9) and (10).

(9) Jerry was excited now. (COHA fiction 2017)

(10) Poor old Judge Richmond got excited and had another stroke. (COHA fiction 1921)

Both examples express a similar state of affairs – a person in the role of grammatical subject being in a state of excitement – but are differentiated by a greater

focus on the state in (9) and on the change of state in (10). These examples demonstrate that the definition of the variable context is not just a problem of automation, but of linguistic interpretation. Arguments could be found both to include or to exclude cases like this from an analysis of the passive alternation. Moreover, as mentioned above, the semantics of both constructions are not completely stable in diachrony. COHA contains many tokens of the BE-passive with clearly dynamic readings, especially from the 19th century. The tendency towards functional association of BE with stative and GET with dynamic situations develops only gradually over the course of the period covered by our data and is far from categorical even in the late 20th century. As such, it does not provide a categorical distinction that can be taken as underlying the analysis of all our tokens.

We recognize these problems as ongoing challenges for our project. We are in the process of addressing them through a combination of manual coding and writing a supervised classifier to separate true cases of variation from those that entail a semantic difference. For the present analysis, we rely on a fuzzy definition of the variable context and only exclude a handful of participle types that categorically do not participate in the alternation or are erroneously tagged: *married, engaged, betrothed, rid, wet, medicaid,* (*over*)*tired*, (*un*)*dressed*, and *clothed*. Further, we only retain participles that occur at least once with each auxiliary and at least ten times in total. After these exclusions, we are left with 2,318,251 observations to model, 2,292,328 of which occur with BE and 25,923 with GET.

3.2 Mixed-effects Logistic Regression

We run a mixed-effects logistic regression model, which quantifies the influence of various predictor variables on the likelihood of GET. Our model considers a number of hypotheses suggested in the extant literature. In addition to established predictors, we add two new variables – verb sentiment score and verb semantic group – to better operationalize previous claims in an empirical framework. The following predictors are included, ordered by the hypothesized constraints they operationalize:

Diachronic Change

– **Year** (continuous): For each observation, we include the year in which it is attested. Given the well-documented increase of the GET-passive and decline of the BE-passive, we expect this predictor to correlate positively with the likelihood of the former being selected. Since linguistic changes are

often accompanied by a levelling of constraints, we also consider an interaction term between year and each of the other predictors.

(In)formality

- **Genre** (fiction, magazines, newspapers, non-fiction books): This information is extracted as part of the COHA file meta-data. We hypothesize that fiction writing as the least formal genre is the most favourable towards GET-passives, while we expect the opposite for non-fiction books.
- **F-measure** (continuous): This predictor is based on the proportions of words from different word-classes, on the assumption that more formal texts contain more nouns, adjectives, prepositions, and articles, whereas more "contextual" texts feature more pronouns, verbs, adverbs, and interjections (Heylighen/Dewaele 2002: 8). A higher F-value indicates increased formality and is consequently expected to favour BE-passives.

Subject Responsibility

- **Subject animacy** (inanimate, animate, body part, unknown): Givón/Yang (1994: 120) suggest that the GET-passive disfavours inanimate referents if "the vestment of purpose, control and responsibility in the surface subject are necessary ingredients of the GET-passive". To test this, we classify the nearest noun or pronoun to the left of the passive (supposedly usually the clause subject) as inanimate (11), animate (12), or body part (13), assuming that the latter may meronymically signify animate entities. The "unknown" category comprises polysemous and mis-tagged items, collective nouns, place names, cases where the preceding NP could not be retrieved automatically as well as items that were not coded due to their low frequency. We hand-coded the most common 28,610 NP types, thus covering 1,803,838 instances or 77.81% of all subject tokens in the data. We checked the accuracy of our coding on a sample of 600 items; extrapolating the results to the hand-coded part of the whole data set gives an estimated accuracy of 87% for the animacy coding when the unknown cases are excluded.

(11) A green salad ($3.73) of beautiful baby mixed greens gets coated in a tart, fruity, unbalanced dressing that suggested sour pineapple juice. (COHA news 1992)

(12) A bad man gets found out sooner or later. (COHA fiction 1914)

(13) Throats got cleared and feet were shifted. (COHA fiction 1979)

– **Agent PP with *by*** (present, absent): If an instance of *by* is found to the immediate right of the participle, this is coded as an agent PP. The presence of an additional constituent encoding the agent of the situation is assumed to at least partially mitigate subject responsibility and thus expected to favour BE. According to Xiao/McEnery/Qian (2006), GET-passives occur even less frequently than BE-passives with an overtly expressed agent. We spot-checked a random sample of 200 instances with *by* and found 96.5% to be genuine agent PPs.

Adversativity/Non-neutrality

– **Main verb negative emotion** (continuous): For each verb participle in our data, we extract its sentiment value from SentiWordNet (Baccianella/Esuli/Sebastiani 2010), a sentiment dictionary containing a positivity (14), a negativity (15), and an objectivity score (16) for over 100,000 words. The main verb negative emotion value is simply the negativity score as found in SentiWordNet. If the sentence is negated, the same score is used with its sign reversed. According to both the adversativity and the more general non-neutrality hypothesis, higher scores for this predictor are expected to favour GET.

(14) To-day my eyes will be gladdened by the consummation of my great achievement. (COHA fiction 1859, positivity score: 0.875)

(15) Fred has a horror of being henpecked. (COHA non-fiction 1953, negativity score: 1)

(16) Continue past the bronze statue of the angel to the paved road that is flanked by the fourteen stations of the cross. (COHA fiction 2002, objectivity score: 1)

– **Main verb positive emotion** (continuous): The calculation of the score is analogous to the above. Whereas the adversativity hypothesis expects no effect of positive emotion scores, or potentially one disfavouring GET, the non-neutrality hypothesis suggests that higher scores for this predictor correlate positively with the likelihood of selecting GET. Positive and negative emotion scores are not, as one might expect, highly correlated ($\kappa = 1.98$) and can therefore be used as predictors in one model.
– **Main verb semantic group** (21 levels): Beyond emotion scores on a simple linear continuum, we test whether different verb groups show distinct selectional preferences in relation to BE and GET. Rühlemann (2007: 122) demonstrates this for individual verbs, arguing that "grammar and lexis can be

shown to a large extent to merge into one another". Here, we attempt to find systematicity beyond isolated items by clustering the 1,800 main verb participle types in our data into natural groups. This is done on the basis of a distributional semantic model for the entire corpus and a subsequent cluster analysis of the relevant participles. For the distributional semantic model, we rely on the word2vec algorithm (Mikolov et al. 2013) as implemented in Python's Gensim module (Řehůřek/Sojka 2010). The output is a representation of each participle in a 100-dimensional vector space based on its co-occurrence with other words in the corpus. A model-based cluster analysis with Gaussian mixture models (Fraley/Raftery/Scrucca 2019) is performed to find the appropriate number of clusters and establish each participle's cluster membership.

Space does not permit us to introduce each verb group here; please consider the appendix for suggested labels and the ten verbs most strongly associated with each cluster. In general, the clusters are characterized by a high degree of semantic coherence, but overall frequency plays a role as well, such that there are several clusters whose main feature is that their members occur only a handful of times in the entire corpus.

Since Gensim's word vector representation is based on single words, we could not treat different phrasal and prepositional, as well as polysemous, verbs separately.

Finally, we code each observation for a number of morpho-syntactic features. Some of these have been remarked on in the literature but are not immediately connected to any of the hypotheses above. Others are included for more exploratory reasons and based on the general assumption that any element making the construction more complex (such as a negator or intervening adverb) tends to disfavour GET.

Morpho-syntactic Constraints

- **Negator** (presence, absence)
- **Form of auxiliary verb** (present (16), preterite (13), perfect (17), infinitive (18), *-ing* (19))
- **Intervening adverb** (presence, absence): Carter/McCarthy (1999: 53) find that "no adverbials occur in medial position between *get* and the main verb past participle".
- **Complementation with a *to*-infinitive** (presence, absence): Xiao/McEnery/Qian (2006: 112) claim that only the BE-passive allows for an infinitival complement. We automatically code for each observation whether the verb is immediately followed by *to* and an infinitival verb.

(17) Donnie was the one who had gotten nicked by a stray bullet in Donkey Creek, earned himself bragging rights if nothing else. (COHA fiction 2005)

(18) And as it is worth fighting for, the insurance companies here will do all they can to get compensated for their losses. (COHA magazine 1883)

(19) At any rate, those things are getting said nowadays; he'll have to hear them sooner or later. (COHA fiction 1889)

Continuous variables are standardized, i.e., their z-scores are used to allow for better comparisons of predictors on different scales. All correlations between numeric predictors are $r < 0.12$. Therefore, they can all be entered into the model. Generalized variance inflation factors are < 10 for year and *to*-complementation, and < 5 for all other predictors.

Categorical variables are treatment-coded, with the following baseline levels: the largest semantic verb cluster 13, no following *by*-PP, no *to*-complementation, unknown animacy (i.e., infrequent nouns), no negation, no adverb, and infinitive for the form of the auxiliary.

A logistic mixed-effects model with verb lemma as a random effect and random slopes for year was fitted in Julia (Bezanson et al. 2017). A random-effects Principal Component Analysis, as advocated by Bates et al. (2018), indicates that the model is not overparameterized and both the random intercept and slope are justified by the data.

4 Results

We present the results of our model with effects plots created using the *ggeffects* package (Lüdecke 2018) in R (RStudio Team 2020). These plots visualize predictions for main effects and significant interactions with their respective confidence intervals. Please refer to the appendix for the full set of model coefficients.

One of the strongest predictors in our model is, as expected, the publication year of the text. Fig. 1 shows how the likelihood of the GET-passive has increased over the past 200 years.

Many of the other predictors show statistically significant interaction with year; these will be discussed below along with other main effects that do not participate in significant interactions. The results are grouped around the relevant hypotheses.

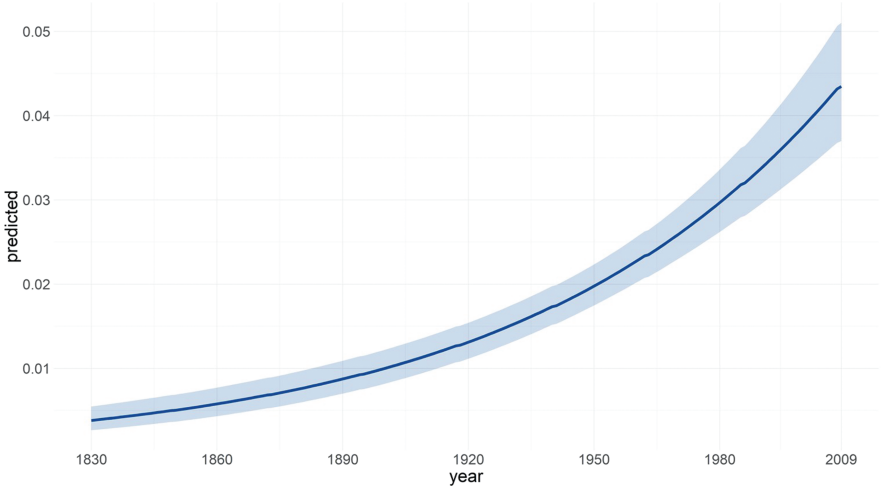

Fig. 1: Predicted likelihood of GET for year as a main effect.

4.1 (In)formality

The GET-passive has become more likely in each of the four genres over time. Notably, this factor interacts with year (Fig. 2) so that the relative increase is greater in magazines and particularly in newspaper writing, which was the genre in which the GET-passive was the least likely in the early 19th century. The likelihood of GET remains the lowest in the most formal genre, non-fiction books, which supports the hypothesis about the informal nature of the GET-passive. Hundt/Mair (1999: 236) suggest that genres differ with regard to "openness to innovation" and "to external socio-cultural influences", finding journalistic writing more "agile" and academic writing more "uptight" and "prone to retain conservative forms". This might explain why the most noticeable increase in the GET-passive is found in newspapers and magazines.

We expected a higher F-value, indicative of formality, to increase the likelihood of BE being selected as the passive auxiliary. Our data confirm this. Furthermore, the likelihood of selecting GET increases most steeply over time for less formal texts with a lower F-value.

4.2 Subject Responsibility

As can be seen in Fig. 3, the significance of *by*-PP as a predictor has decreased over time so that in the early 19th century, its presence disfavours GET even

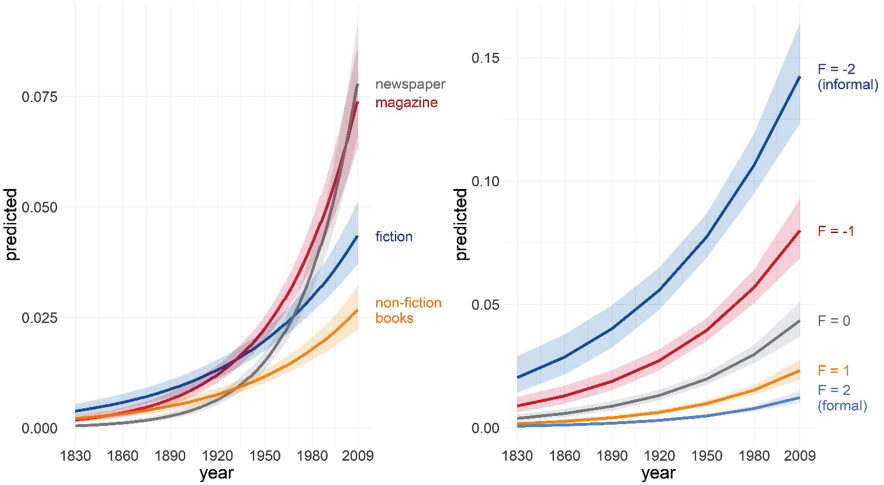

Fig. 2: Predicted likelihood of GET for the interactions between year and genre (left) and z-scored F-measure (right). Higher F-measures indicate higher levels of formality.

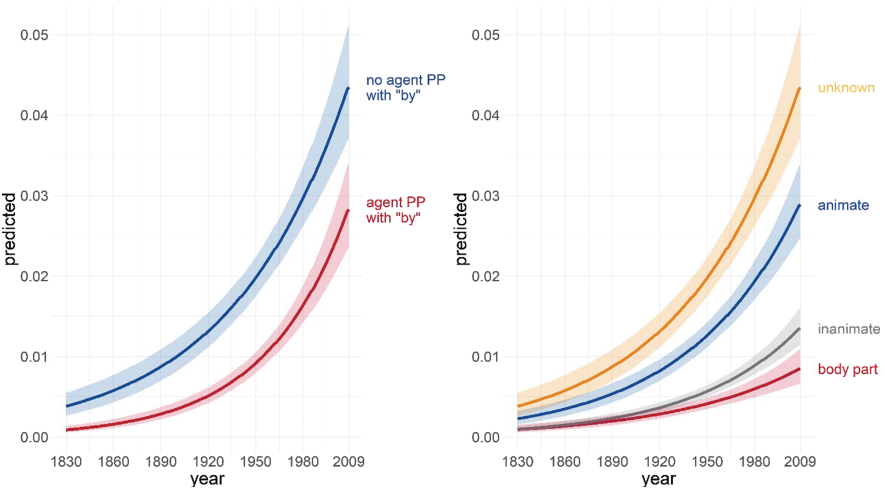

Fig. 3: Predicted likelihood of GET for the interactions between year and *by*-PP (left) and animacy (right).

more strongly than it does by the end of the 20th century. This finding can be related to an overt agent being less compatible with the concept of subject responsibility, lending some support to the hypothesis that the GET-passive may be used to encode situations where the grammatical subject has some agency over the situation. However, the GET-passive appears to be losing this

semantic nuance gradually and to be developing in the direction of a more general passive form.

Going in line with this, our model suggests animate subjects to rather co-occur with GET and inanimate subjects with BE. Unexpectedly, body-part subjects show a strong association with the BE-passive; this category, however, covers only 2.7% of the data. The only significant interaction is between inanimate subjects and year; GET becomes relatively more likely to occur with inanimate subjects over time. The significance of subject responsibility seems to be lessening.

4.3 Adversativity/Non-neutrality

A higher sentiment score in either direction – positive or negative – makes the occurrence of GET as the passive auxiliary slightly more likely (Fig. 4). While statistically significant, however, the effect size is quite negligible. Considering the large size of the corpus, the very small coefficients (–0.016 for a negative score; –0.023 for a positive one), and the rather large confidence intervals, we do not think the data offer genuine support to the adversativity or non-neutrality hypotheses.

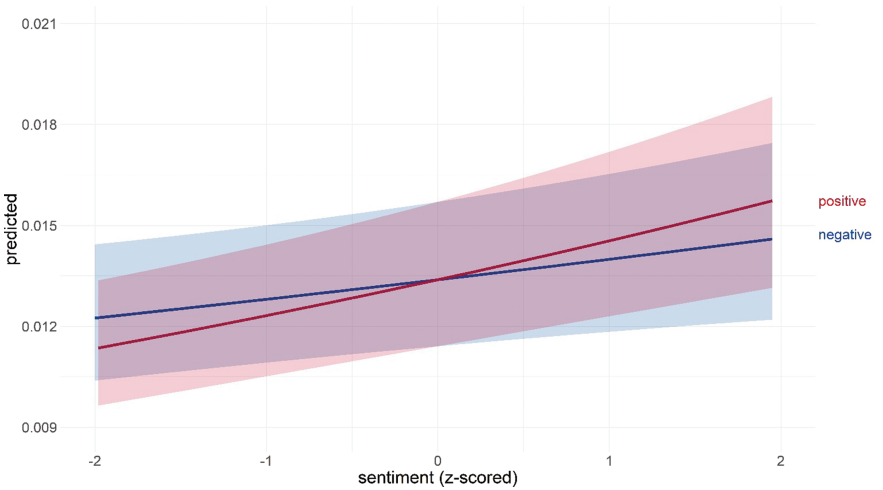

Fig. 4: Predicted likelihood of GET for positive and negative sentiment as a main effect.

Verb semantics, however, does seem to play a role in the choice of the passive auxiliary. Semantic clustering sheds more detailed light on this. The dot-and-whiskers plot in Fig. 5 visualizes the coefficients and their confidence intervals

for those 17 clusters that show no significant interaction with year. The groups that differ significantly from the largest, baseline cluster 13 are marked with asterisks.

The clusters that strongly and steadily predict BE contain more formal verbs encoding deontic (*obliged*, *allowed*) and mental actions (*defined*, *considered*), and changes in quality/quantity (*increased*, *postponed*). Clusters showing steady relative preference for GET encode physical motion (*stomped*, *whacked*), and concrete actions, for instance, in the culinary context (*salted*, *baked*). Additionally, most of the low-frequency verb clusters are located towards the GET-end of the spectrum. Some of them contain primarily negative verbs – for example *short-changed*, *gypped*, and *guillotined* in cluster 17 – which could be seen as tentative support to the adversativity hypothesis. However, we are inclined to think that the main defining feature of these clusters is the low frequency of their members. The relatively higher likelihood of attesting GET with infrequent verbs shows that the construction is by no means restricted to entrenched combinations like "got killed".

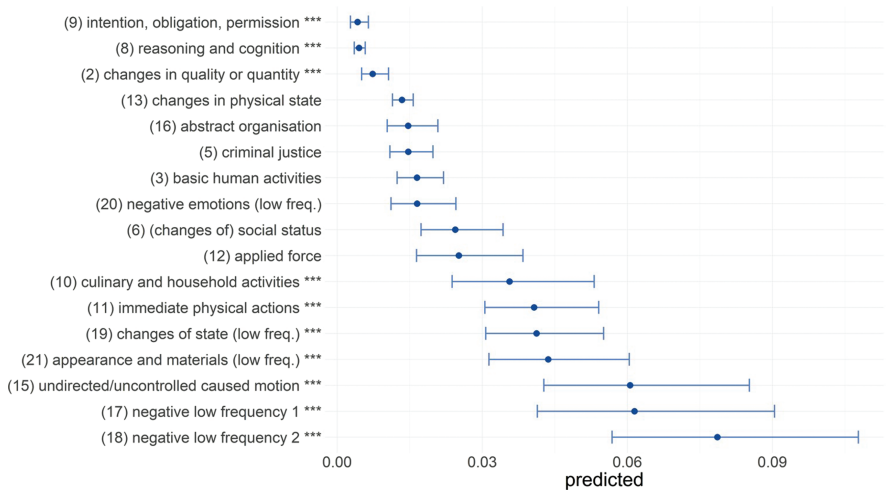

Fig. 5: Predicted likelihood of GET for cluster assignment as a main effect.

Fig. 6 shows the remaining four clusters that display diachronic developments different from the general trend. We see the cluster containing verbs of goal-directed/controlled motion (*thrown*, *carried*) becoming drastically more likely to combine with GET, and verbs of public communication (*printed*, *published*) rising relatively more steeply as well, while the likelihood of GET has not increased much with time for verbs that can be expected to be often used statively: clusters

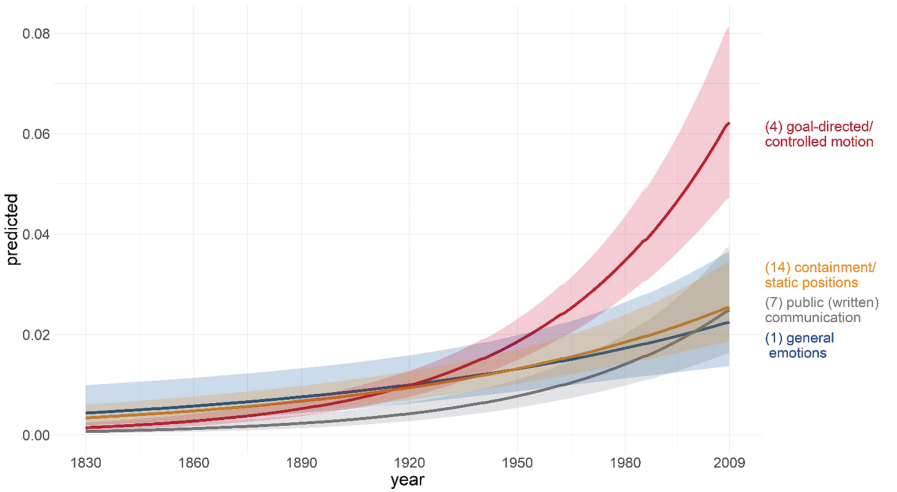

Fig. 6: Predicted likelihood of GET for the interaction between year and cluster assignment.

containing socially accepted emotions (*annoyed*, *excited*) and containment/static positions (*covered*, *stacked*).

4.4 Morpho-syntactic Constraints

The morpho-syntactic predictors are included primarily to explore the supposition that as the GET-passive grammaticalizes (Hopper/Traugott 2003), it becomes increasingly available for different and more complex sentence structures as well. We do see a trend in this direction regarding the presence of a *by*-PP, a negator, and/or a *to*-complement (Fig. 7), each of which disfavours GET significantly more strongly in the early 19th century than in the late 20th and early 21st century; in the newest data, in fact, the model predicts a slightly higher likelihood for GET in negated sentences. The presence of an intervening adverb, however, now lowers the likelihood of GET even more than before.

The form of the auxiliary was included as a variable rather for exploratory reasons with no specific hypothesis attached to it. The results, while statistically significant, are not easy to interpret or explain (Fig. 7). Perfect forms lag clearly behind in the general trend towards more use of GET over time, and while one still finds the highest likelihood of GET with *-ing*-forms, this preference has become less strong over time. The avoidance of GET with perfect and preterite may have to do with a need to differentiate the GET-passive from possessive (HAVE) *got*.

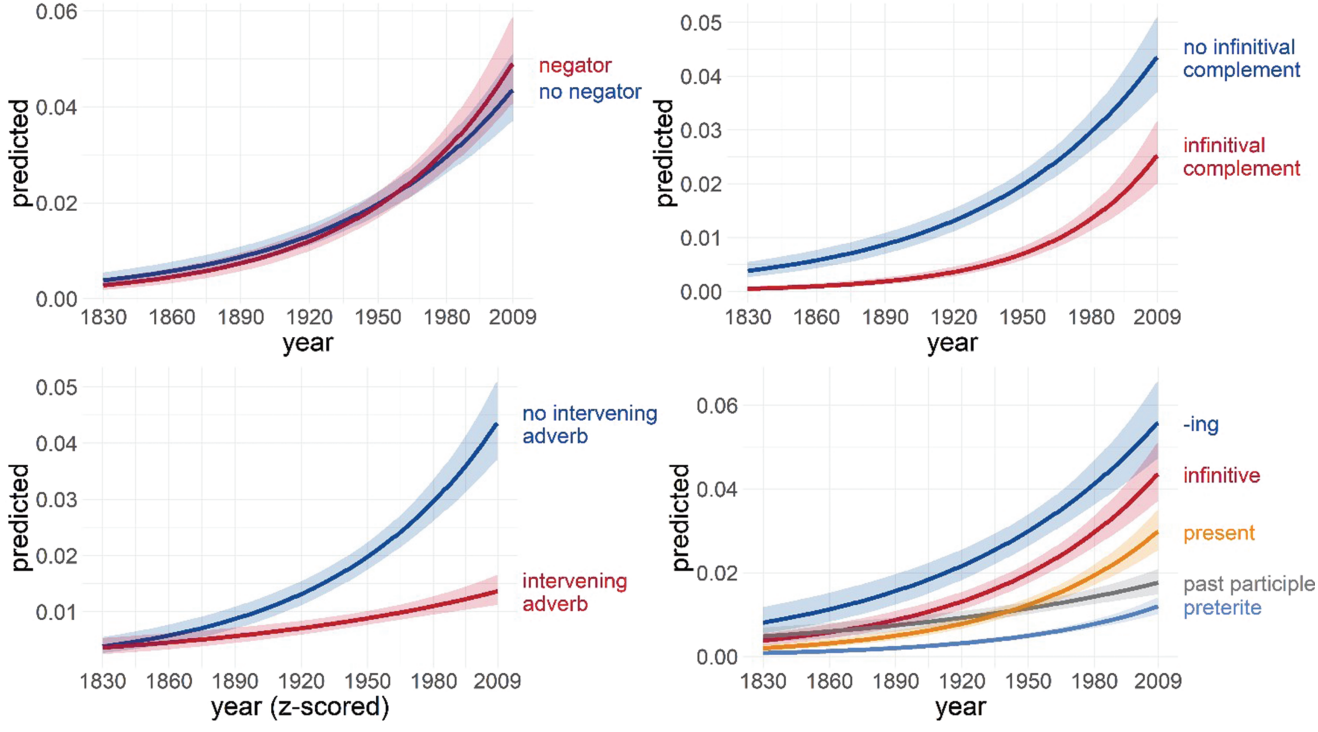

Fig. 7: Predicted likelihood of GET for the interactions between year and negator (top left), infinitival complement (top right), intervening adverb (bottom left), and tense of auxiliary verb (bottom right).

5 Discussion

The first thing to note is the number of main effects and interactions that emerge as significant. Before interpreting any of these in detail, the results of the statistical analysis speak to the complexity of linguistic conditioning in the case of the passive alternation. Investigations of isolated predictors based on raw frequency counts or individual bivariate tests are not ideally suited to do justice to this level of complexity. Our study is the first to our knowledge that grounds its analysis of the English passive alternation in a fully accountable multivariate design, which we maintain is necessary to disentangle the influence of competing conditioning factors.

As regards the specific hypotheses about such conditioning, our results offer partial confirmation for many of these but also suggest a need for further qualification. Unsurprisingly, our data corroborate the general rise of the GET-passive, a process that picks up speed in the latter half of the period covered by our data, i.e., during the 20th century. This change is accompanied by an encroachment of GET into initially disfavouring contexts, as seen in the weakening of morpho-syntactic constraints. The fact that the presence of an intervening adverb shows an effect in a different direction, emerging as a constraint only as time progresses, is an interesting reversal of this general trend in need of further analysis in the future. One explanation might be that many of these sentences are not true passives at all, but that the adverb pre-modifies an adjectival usage of the participle, as for example in (20). However, initial qualitative analysis of a subset of the data does not offer conclusive confirmation of this hypothesis; see for example (21).

(20) If insurance companies cannot go directly to the client where such recourse is morally justified, they will simply step up their rates. (COHA news 1928)

(21) The silence of midnight was almost constantly interrupted by the howling of wild beasts. (COHA magazine 1835)

The innovative GET-variant maintains its informal stylistic profile throughout the time period covered in this study, as seen in the effects plots for both genre and F-measure. The divergent paths of different genres over time speak to general developments in the register ecology of written English (Hundt/Mair 1999) rather than calling the informality hypothesis into question in general. We do not see a weakening of the formality constraint; on the contrary, lower F-measure values (indicating less formality) become more strongly associated with GET over time.

The results are less conclusive concerning the related hypotheses of subject responsibility and adversativity/non-neutrality. The former constraint is found to operate in our data, but to be weakening over time, as seen in the interactions of year with both *by*-PPs and inanimate subjects. This trend is indicative of an expansion of the GET-passive into more general passive contexts. Currently, our operationalization of subject animacy achieves accuracy in about two-thirds of all cases. With sufficient data, such a level of precision is able to offer meaningful results, but nonetheless, more work needs to be dedicated to improving classification accuracy for subject animacy.

While the effects of verb sentiment scores reach statistical significance, their magnitudes are among the smallest in our model. This might either weaken the non-neutrality hypothesis, or be due to the operationalization of sentiment. Specifically, we simply derived verb sentiment scores from an available sentiment lexicon. Missing entries for many lemmas in our data, in addition to insensitivity to the wider sentential context, render the sentiment scores somewhat suspect, and may partially explain their very small effect sizes. We already consider the presence of a negator, thus improving the accuracy of the sentiment scores, but we are convinced that the more appropriate level of sentiment analysis is the entire sentence embedding the given passive instance. Yet, in small-scale comparisons of manually and automatically scored sentiments we observed that automated sentiment scores on the sentence-level suffer from accuracy issues. The older data in particular present challenges to scoring algorithms, most of which have been trained on contemporary language. All in all, compared to other conditioning effects, we see little reason to attach strong meaning to adversativity or non-neutrality as an explanatory variable.

Perhaps our most important finding is the importance of verb semantic cluster. This predictor's effect size outshines that of all others, lending strong support to Rühlemann's (2007) account of the GET-passive as largely lexically conditioned. However, the fact that cluster membership holds such explanatory power, even after accounting for individual verb lemmas with random intercepts, indicates that the behaviour of isolated lexemes is not the appropriate level of generalization. Instead, groups of related verbs that can be identified through corpus-based, quantitative methods show similar passivation profiles. We see our key contribution in sharpening the focus on the relevance of this constraint and in outlining a principled method for operationalizing it quantitatively by means of a distributional word vector model.

Given that our key finding concerns the influence of verb cluster membership, critical attention to this variable is particularly warranted. As with any clustering solution, it is difficult to justify our choice in absolute terms. In order to maximize accountability, we opted for model-based clustering, which has the advantage of

identifying the optimal number of clusters in a mathematically principled way. However, a hierarchical, density-based, or any other cluster solution could also have been chosen, yielding potentially different results. We do not see a choice of method immune to criticism. However, in experimenting with various techniques and parameter settings, we noted largely convergent patterns in how individual verbs are grouped together. The difference, then, is more one of detail than of substance. Should different clustering solutions produce highly divergent results, this should be taken as evidence that the underlying vector representation of the words is not reliable, a problem we do not face in this analysis.

The behaviour of the various morpho-syntactic predictors indicates the expansion of GET to more general contexts. *By*-PPs, negators, and *to*-complements used to disfavour GET more strongly in the older data than now. The usage of different tense and aspectual forms does not paint a clear picture, other than the perfect and the preterite being the least GET-friendly forms. This behaviour may be explained by a need to differentiate the GET-passive from possessive (HAVE) *got*. Furthermore, the presence of adverbial premodification strongly mediates the diachronic rise of GET, such that the increase is much more pronounced in bare cases without an intervening adverb. Our tentative working hypothesis is that the presence of an adverb strongly correlates with adjectival uses of the past participle, i.e., marginal instances on the passive gradient. It is possible that these do not constitute valid variable contexts, an issue that will be addressed in future work.

To boil the discussion down to its essence, our large-scale quantitative approach marks a radical divergence from previous research on the GET-passive. Both the potential insights and the problems entailed by this approach are considerable. It is our position that the former more than justify the general choice of method and the latter can, and will, be addressed more comprehensively in future research.

6 Outlook

The present paper contains preliminary findings from our investigation into the factors influencing the passive alternation in American English. More precisely, we investigate change in the choice between BE and GET as the passive auxiliary over the past two centuries, drawing on data provided by COHA. To our knowledge, this big data approach is the first study embedding a large range of variables into a multivariate framework and is therefore able to account for the multitude of predictors and their interactions. Our method allows for an in-

depth analysis of the passive construction and offers potential for similar studies of other syntactic and semantic phenomena, such as changes in the use of the progressive.

To summarize the results, we observed a general rise of GET both in absolute numbers and as a competitor to BE throughout the observed time period (1830–2009). Our informality hypothesis was confirmed: GET is more likely in less formal texts, and this tendency has actually strengthened over time. We have further shown that, unlike previously assumed, the constraint of subject responsibility has weakened over time and GET has expanded its reach to more general passive contexts. This is suggested by the decline of the inhibiting effect of both *by*-PPs and inanimate subjects over time. The behaviour of the morpho-syntactic constraints also lends support to this interpretation. In terms of adversativity/non-neutrality, our findings were less conclusive, but in general, we did not find strong support for the significance of these suggested semantic characteristics of the GET-passive. Furthermore, our results corroborate our supposition that the grouping of verbs into semantic clusters is more revealing than an analysis based on isolated lexemes, such as verb sentiment. The highly effective and methodologically innovative process of semantic clustering supports the notion of lexical conditioning of the GET-passive, fitting with current usage-based approaches that see grammar and meaning as tightly integrated.

Several challenges remain to be addressed. Most importantly, we have adopted a very general notion of the variable context, focusing on all cases of BE/GET + past participle, as tagged in COHA, with only very few exclusions. This fuzzy context comprises individual forms that cannot be counted as 'true passives' and lack full referential equivalence to a competing variant. Whether it makes sense to treat, for example, GET *excited* and BE *excited* as directly competing variants is questionable. We therefore recognize that further pruning of the data is desirable in order to keep the analysis more accountable. Currently, we are working on semi-automated as well as fully qualitatively guided ways of narrowing down the variable context to central passives, excluding adjectival uses. The semi-automated measures will include, among others, how often an assumed participle is used with a copula, adjectives, or *very* and other degree elements in COHA, as well as whether it allows *un*-prefixation. The extent to which the results reported here are corroborated by these analyses will be an important touchstone for our method.

Beyond the factors discussed in this paper, we are planning on extending the investigation to other sources of variation in the passive. One such factor potentially contributing to variation is dynamicity, which would be highly interesting to consider, but difficult to operationalize and extract automatically. We have only begun to scratch the surface of the potential offered by the big

data approach considering a wide range of variables. The results so far are encouraging and call for further investigations, such as comparing the two constructions across varieties of English.

7 Appendix

Tab. 1: Top ten verbs per cluster.

	Cluster	Verbs
1	general emotions	*annoyed, excited, embarrassed, puzzled, confused, alarmed, shocked, irritated, perplexed, frightened*
2	changes in quality or quantity	*increased, reduced, postponed, delayed, diminished, improved, accelerated, removed, eliminated, changed*
3	basic human activities	*lived, sat, stood, talked, spent, waited, told, said, loved, listened*
4	goal-directed/controlled motion	*thrown, carried, swept, driven, flung, brought, poured, hurled, dragged, put*
5	criminal justice	*murdered, hanged, convicted, slain, jailed, accused, sentenced, imprisoned, raped, fined*
6	(changes of) social status	*excommunicated, naturalized, proscribed, flogged, impeached, beheaded, reprimanded, reunited, martyred, paroled*
7	public (written) communication	*printed, published, issued, written, signed, circulated, mailed, quoted, delivered, copied*
8	reasoning and cognition	*defined, considered, exercised, imposed, characterized, represented, criticized, known, dealt, deemed*
9	intention, obligation, permission	*obliged, allowed, forced, prepared, let, tempted, needed, accustomed, invited, refused*
10	culinary and household activities	*fried, salted, stewed, broiled, baked, boiled, buttered, pickled, roasted, canned*
11	immediate physical actions	*bumped, tripped, bounced, beat, hammered, scrambled, hopped, spun, pounded, rattled*
12	applied force	*cocked, clutched, tugged, tightened, braced, bent, stroked, straightened, tilted, rubbed*
13	changes in physical state	*withered, distorted, bruised, starved, purified, shed, smothered, chilled, spoiled, shattered*
14	containment/static positions	*lined, covered, stacked, littered, packed, loaded, crowded, surrounded, crammed, filled*
15	undirected/uncontrolled caused motion	*throwed, whacked, plopped, bailed, blowed, bowled, chucked, stomped, shooed, gobbled*
16	abstract organisation	*targeted, coordinated, subsidized, monitored, programmed, evaluated, publicized, oriented, mapped, channeled*

Tab. 1 (continued)

	Cluster	Verbs
17	negative low frequency 1	*bushwhacked, zonked, jobbed, short-changed, gypped, guillotined, propositioned, jugged, articled, resupplied*
18	negative low frequency 2	*bulldozed, bluffed, cloned, sensitized, sidelined, circumcised, electrocuted, trashed, shacked, bedeviled*
19	changes of state (low freq.)	*overheated, winded, overdone, readjusted, freshened, unhooked, flurried, dehydrated, primed, unmade*
20	negative emotions (low freq.)	*incensed, infuriated, humiliated, embittered, thwarted, enraged, angered, intimidated, harassed, cowed*
21	appearance and materials (low freq.)	*rumpled, matted, creased, freckled, discolored, patched, knitted, glazed, laced, rusted*

Tab. 2: Model coefficient estimates.

term	estimate	standard error	z-value	p-value
(Intercept)	−6.29927	0.0825846	−76.28	<1e-99
year_z	1.00038	0.06697	14.94	<1e-49
genre: mag	−0.0770036	0.0338057	−2.28	0.0227
genre: news	−0.70799	0.0714953	−9.90	<1e-22
genre: nf	−0.554269	0.0423025	−13.10	<1e-38
fMeasure_z	−0.744673	0.0161895	−46.00	<1e-99
animacy_cat: animate	0.352907	0.0234325	15.06	<1e-50
animacy_cat: body part	−0.711412	0.0566091	−12.57	<1e-35
animacy_cat: inanimate	−0.466123	0.0267487	−17.43	<1e-67
by: by	−0.471775	0.0280702	−16.81	<1e-62
participleNegative_z	0.0450047	0.016817	2.68	0.0074
participlePositive_z	0.0840761	0.0165777	5.07	<1e-6
ClusterAssignment: 1	−0.5899	0.220049	−2.68	0.0073
ClusterAssignment: 2	−0.916412	0.171836	−5.33	<1e-7
ClusterAssignment: 3	−0.0941738	0.125678	−0.75	0.4537
ClusterAssignment: 4	−0.585104	0.104404	−5.60	<1e-7
ClusterAssignment: 5	−0.212574	0.131247	−1.62	0.1053
ClusterAssignment: 6	0.305356	0.15939	1.92	0.0554
ClusterAssignment: 7	−1.43674	0.195456	−7.35	<1e-12
ClusterAssignment: 8	−1.4051	0.1066	−13.18	<1e-38
ClusterAssignment: 9	−1.47553	0.201912	−7.31	<1e-12
ClusterAssignment: 10	0.695088	0.193572	3.59	0.0003
ClusterAssignment: 11	0.833613	0.129664	6.43	<1e-9
ClusterAssignment: 12	0.335724	0.203223	1.65	0.0985
ClusterAssignment: 14	−0.646327	0.132854	−4.86	<1e-5
ClusterAssignment: 15	1.25195	0.165797	7.55	<1e-13

Tab. 2 (continued)

term	estimate	standard error	z-value	p-value
ClusterAssignment: 16	−0.213618	0.162823	−1.31	0.1895
ClusterAssignment: 17	1.26807	0.192014	6.60	<1e-10
ClusterAssignment: 18	1.53212	0.157929	9.70	<1e-21
ClusterAssignment: 19	0.846663	0.135071	6.27	<1e-9
ClusterAssignment: 20	−0.0925117	0.188688	−0.49	0.6239
ClusterAssignment: 21	0.906087	0.155566	5.82	<1e-8
aux_tense: ing	0.503215	0.0343079	14.67	<1e-47
aux_tense: perf	−0.364226	0.0270991	−13.44	<1e-40
aux_tense: pres	−0.522158	0.0257872	−20.25	<1e-90
aux_tense: pret	−1.43234	0.0280256	−51.11	<1e-99
adverb: rr	−0.317307	0.0187147	−16.95	<1e-63
to_compl: toComp	−0.64692	0.053865	−12.01	<1e-32
negator: neg	−0.0451412	0.0269028	−1.68	0.0934
year_z & genre: mag	0.374067	0.0293471	12.75	<1e-36
year_z & genre: news	0.777004	0.0560803	13.86	<1e-42
year_z & genre: nf	0.0285464	0.03764	0.76	0.4482
year_z & fMeasure_z	0.0563708	0.0141633	3.98	<1e-4
year_z & animacy: animate	0.027297	0.0211088	1.29	0.1960
year_z & animacy: body part	−0.0789024	0.0509244	−1.55	0.1213
year_z & animacy: inanimate	0.052503	0.0241587	2.17	0.0298
year_z & by: by	0.145539	0.0239594	6.07	<1e-8
year_z & participleNegative_z	−0.0164403	0.0143878	−1.14	0.2532
year_z & participlePositive_z	−0.0229869	0.0146535	−1.57	0.1167
year_z & ClusterAssignment: 1	−0.2683	0.128219	−2.09	0.0364
year_z & ClusterAssignment: 2	−0.028102	0.107379	−0.26	0.7935
year_z & ClusterAssignment: 3	−0.0540112	0.0819367	−0.66	0.5098
year_z & ClusterAssignment: 4	0.349616	0.0713911	4.90	<1e-6
year_z & ClusterAssignment: 5	0.155091	0.0909364	1.71	0.0881
year_z & ClusterAssignment: 6	0.147213	0.107106	1.37	0.1693
year_z & ClusterAssignment: 7	0.287328	0.125378	2.29	0.0219
year_z & ClusterAssignment: 8	0.101418	0.0662312	1.53	0.1257
year_z & ClusterAssignment: 9	0.121376	0.116831	1.04	0.2988
year_z & ClusterAssignment: 10	−0.119907	0.124978	−0.96	0.3373
year_z & ClusterAssignment: 11	−0.0560703	0.0839791	−0.67	0.5043
year_z & ClusterAssignment: 12	−0.254271	0.13054	−1.95	0.0514
year_z & ClusterAssignment: 14	−0.160681	0.0802212	−2.00	0.0452
year_z & ClusterAssignment: 15	0.00868104	0.128198	0.07	0.9460
year_z & ClusterAssignment: 16	0.00729065	0.110859	0.07	0.9476
year_z & ClusterAssignment: 17	0.261172	0.150832	1.73	0.0834
year_z & ClusterAssignment: 18	−0.0999417	0.112238	−0.89	0.3732
year_z & ClusterAssignment: 19	−0.166637	0.0982637	−1.70	0.0899
year_z & ClusterAssignment: 20	0.00322216	0.1208	0.03	0.9787
year_z & ClusterAssignment: 21	−0.200298	0.10481	−1.91	0.0560

Tab. 2 (continued)

term	estimate	standard error	z-value	p-value
year_z & aux_tense: ing	−0.141531	0.0308343	−4.59	<1e-5
year_z & aux_tense: perf	−0.33251	0.0256249	−12.98	<1e-37
year_z & aux_tense: pres	0.0749081	0.0234108	3.20	0.0014
year_z & aux_tense: pret	0.060927	0.0256607	2.37	0.0176
year_z & adverb: rr	−0.162972	0.0176527	−9.23	<1e-19
year_z & to_compl: toComp	0.213035	0.044595	4.78	<1e-5
year_z & negator: neg	0.0627364	0.0239071	2.62	0.0087

References

Baccianella, Stefano/Esuli, Andrea/Sebastiani, Fabrizio (2010): "SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*. 17–23. May 2010, Valetta. 2200–2204.

Banks, David (1986): "Getting by with Get." In: *La Linguistique* 22, 125–130.

Bates, Douglas/Kliegl, Reinhold/Vasishth, Shravan/Baayen, Harald (2018): "Parsimonious Mixed Models." arXiv:1506.04967 [stat.ME], 1–21.

Bezanson, Jeff/Edelman, Alan/Karpinski, Stefan/Shah, Viral B. (2017): "Julia: A fresh approach to numerical computing." In: *SIAM Review* 59, 65–98.

Biber, Douglas/Conrad, Susan/Leech, Geoffrey (2003): *Student Grammar of Spoken and Written English*. 2nd edn. Harlow: Longman.

Carter, Ronald/McCarthy, Melissa (1999): "The English get-passive in spoken discourse: Description and implications for an interpersonal grammar." In: *English Language and Linguistics* 3, 41–58.

Chappell, Hilary (1980): "Is the get-passive adversative?" In: *Papers in Linguistics* 13, 411–452.

Collins, Peter C. (1996): "Get-passives in English." In: *World Englishes* 15, 43–56.

Coto Villalibre, Eduardo (2015): "Is the get-passive really that adversative?" In: *Miscelánea: A Journal of English and American Studies* 51, 13–30.

Croft, William (1991): *Syntactic Categories and Grammatical Relations. The Cognitive Organization of Information*. Chicago: The University of Chicago Press.

Davies, Mark (2010-): *Corpus of Historical American English (COHA)*. Online at: https://www.english-corpora.org/coha/ <03.11.2021>.

Fleisher, Nicholas (2006): "The origin of passive get." In: *English Language and Linguistics* 10, 225–252.

Fraley, Chris/Raftery, Adrian E./Scrucca, Luca (2019): *Mclust: Gaussian mixture modelling for model-based clustering, classification, and density estimation*. Online at: https://cran.r-project.org/web/packages/mclust/index.html <12.05.2022>.

Givón, Thomas/Yang, Phil (1994): "The rise of the English GET-passive." In: Barbara Fox/Paul J. Hopper (Eds.): *Voice. Form and Function*. Amsterdam: John Benjamins, 119–150.

Hatcher, Anna G. (1949): "To get/be invited." In: *Modern Language Notes* 64, 433–446.
Heylighen, Francis/Dewaele, Jean-Marc (2002): "Variation in the contextuality of language: An empirical measure." In: *Foundations of Science* 7, 293–340.
Hopper, Paul J./Traugott, Elizabeth C. (2003): *Grammaticalization*. 2nd edn. Cambridge: Cambridge University Press.
Hundt, Marianne (2001): "What corpora tell us about the grammaticalisation of voice in get-constructions." In: *Studies in Language* 25, 49–88.
Hundt, Marianne/Mair, Christian (1999): "'Agile' and 'uptight' genres: The corpus-based approach to language change in progress." In: *International Journal of Corpus Linguistics* 4, 221–242.
Lüdecke, Daniel (2018): "ggeffects: Tidy data frames of marginal effects from regression models." In: *Journal of Open Source Software* 3, 772.
Mair, Christian/Leech, Geoffrey (2006): "Current changes in English syntax." In: Bas Aarts/April McMahon (Eds.): *The Handbook of English Linguistics*. Malden: Blackwell, 318–342.
Mikolov, Tomas/Chen, Kai/Corrado, Gred/Dean, Jeffrey (2013): "Efficient estimation of word representations in vector space." arXiv:1301.3781 [cs.CL], 1–12.
Poplack, Shana/Tagliamonte, Sali A. (1989): "There's no tense like the present: Verbal -s inflection in early Black English." In: *Language Variation and Change* 1, 47–84.
Python Software Foundation (2019): *Python 2.7.10*.
Quirk, Randolph/Greenbaum, Sydney/Leech, Geoffrey/Svartvik, Jan (1985): *A Comprehensive Grammar of the English Language*. London: Longman.
Řehůřek, Radim/Sojka, Petr (2010): "Software framework for topic modelling with large corpora." In: *Proceedings of Workshop on New Challenges for NLP Frameworks, LREC 2010*, 17–23 May 2010, Valletta, 46–50.
RStudio Team (2020): *RStudio. Integrated Development for R*. Boston.
Rühlemann, Christoph (2007): "Lexical grammar: The GET-passive as a case in point." In: *ICAME Journal* 31, 111–127.
Schwarz, Sarah (2015): "Passive voice in American soap opera dialogue." In: *Studia Neophilologica* 87, 152–170.
Schwarz, Sarah (2017): "'Like getting nibbled to death by a duck': Grammaticalization of the get-passive in the TIME Magazine Corpus." In: *English World-Wide* 38, 305–335.
Toyota, Junichi (2008): *Diachronic Change in the English Passive*. Basingstoke: Palgrave Macmillan.
Weiner, Judith E./Labov, William (1983): "Constraints on the agentless passive." In: *Journal of Linguistics* 19, 29–58.
Xiao, Richard Z./McEnery, Tony/Qian, Yufang (2006): "Passive constructions in English and Chinese: A corpus-based contrastive study." In: *Languages in Contrast* 6, 109–149.