UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Sassenhagen, J., et al. (2018). Frequency Sensitivity of Neural Responses to English Verb Argument Structure Violations. *Collabra: Psychology*, 4(1): 38. DOI: https://doi.org/10.1525/collabra.87

ORIGINAL RESEARCH REPORT

Frequency Sensitivity of Neural Responses to English Verb Argument Structure Violations

Jona Sassenhagen*, Ryan Blything†, Elena V. M. Lieven‡,§ and Ben Ambridge§

How are verb-argument structure preferences acquired? Children typically receive very little negative evidence, raising the question of how they come to understand the restrictions on grammatical constructions. Statistical learning theories propose stochastic patterns in the input contain sufficient clues. For example, if a verb is very common, but never observed in transitive constructions, this would indicate that transitive usage of that verb is illegal. Ambridge et al. (2008) have shown that in offline grammaticality judgements of intransitive verbs used in transitive constructions, low-frequency verbs elicit higher acceptability ratings than high-frequency verbs, as predicted if relative frequency is a cue during statistical learning. Here, we investigate if the same pattern also emerges in on-line processing of English sentences. EEG was recorded while healthy adults listened to sentences featuring transitive uses of semantically matched verb pairs of differing frequencies. We replicate the finding of higher acceptabilities of transitive uses of low- vs. high-frequency intransitive verbs. Event-Related Potentials indicate a similar result: early electrophysiological signals distinguish between misuse of high- vs low-frequency verbs. This indicates online processing shows a similar sensitivity to frequency as off-line judgements, consistent with a parser that reflects an original acquisition of grammatical constructions via statistical cues. However, the nature of the observed neural responses was not of the expected, or an easily interpretable, form, motivating further work into neural correlates of online processing of syntactic constructions.

Keywords: Verb-argument structure overgeneralization; Event Related Potentials; Statistical-learning; P600; Left anterior negativity (LAN)

Introduction

Acquisition of verb-argument structure preferences absent negative evidence

How do children learn not to say "*The magician disappeared the rabbit"? Instances of a broad class of statistical learning models – supervised learners – require an abundance of *negative examples*, i.e., explicit signals that a certain choice is illegal (Hastie, Tibshirani, & Friedman, 2009). But children learn the grammar of their target language(s) without negative evidence (Lieven, 1994) – e.g., without being told that *disappear*, unlike *remove* or *hide*, does not license a direct object. One suggestion for what information in the environment children are picking up on is that of baseline frequency (Braine & Brooks, 1995; for a similar perspective, see Goldberg, 2003). If a verb like *disappear* is encountered very frequently, but never with a direct object, children could note that if *disappear* allows direct objects, amongst the many usages of *disappear* encountered, some should have been transitive; and from that, infer that *disappear* does *not* allow direct objects. Thus, children would be able to supplant the need for negative evidence via frequency-weighted appraisal of absences. As explained by Pullum (2013), this inference can be summarized in terms of conditional probabilities; $P(O|V) = \frac{P(O \cap V)}{P(V)}$, i.e., the conditional probability of observing any direct object O following a given verb V increases with the number of joint observations of V and any O, and falls with the absolute number of observations of V.[1]

This proposal – the so-called *entrenchment* hypothesis – entails a crucial prediction: that the acceptability of constructions interacts with word frequency. For example, if speakers derive their knowledge about which verbs are transitive vs. intransitive from the relationship between the frequency of observing the verb at all vs. observing it in transitive constructions, then their confidence in judging a given intransitive verb's transitive usage as acceptable should be higher if the verb is encountered less often. Ambridge et al. (2008) have indeed shown that to be the case: in their study, transitive uses of low-frequency verbs were judged as more acceptable than those of high-frequency verbs.

* Department of Cognitive Neuropsychology, University of Frankfurt, DE

† University of Bristol, UK

‡ University of Manchester, UK

§ University of Liverpool, ESRC International Centre for Language and Communicative Development, UK

Corresponding author: Jona Sassenhagen (jona.sassenhagen@gmail.com)

Testing entrenchment: brain correlates

Many competing accounts to (e.g., Pinker, 1979) and criticism of (e.g., Yang, 2011) such statistical learning models exist. Here, we do not attempt a balanced review of the literature (see, e.g., Ambridge, Pine, Rowland, Chang, & Bidgood, 2013; Lidz & Gagliardi, 2015), but simply focus on tests of the prediction derived from the entrenchment hypothesis discussed above.[2] Acceptability ratings collected by Ambridge et al. (2008) indicate that off-line acceptability judgements are at least compatible with the entrenchment account. However, what, if any, are the on-line, incremental correlates of speakers' brains' processing these constructions? Online measures have confirmed language is processed incrementally (Bornkessel & Schlesewsky, 2006; Friederici, Mecklinger, Spencer, Steinhauer, & Donchin, 2001; Rayner & Clifton Jr, 2009), and global judgements of acceptability do not always directly mirror local processing at points of divergence.

Event-related potentials/ERPs (Luck, 2005), i.e., aggregated fast brain responses to temporally localised events, have established themselves as a premier tool for the study of online neural correlates of language processing (Bornkessel & Schlesewsky, 2006; Friederici, 2002). Previous research has yielded a series of ERP components associated with specific dimensions of language- and, more specifically, syntactic processing (Bornkessel & Schlesewsky, 2006; Friederici, 2002). These include the Left-Anterior Negativity/LAN and the associated Early Left-Anterior Negativity (Friederici, 2002; Friederici, Hahne, & Mecklinger, 1996) associated with, e.g., incorrect case marking of arguments; and the P600 (Osterhout & Holcomb, 1992), associated with broad classes of syntactic processes, including error monitoring (Meerendonk, Kolk, Chwilla, & Vissers, 2009) and integration (Bornkessel-Schlesewsky & Schlesewsky, 2008). However, such functional interpretations are routinely put into question (Coulson et al., 1998a; Sassenhagen & Bornkessel-Schlesewsky, 2015; Sassenhagen, Schlesewsky, & Bornkessel-Schlesewsky, 2014; Steinhauer & Drury, 2012). Absent a clear understanding of the functional roles – or even of an unambiguous measure for the identification of – these components, it is dangerous to conduct 'reverse inference' (Poldrack, 2011) of the form that observing component *A* indicates cognitive process *B*; instead, it should be preferred to simply consider ERP components as upper temporal bounds for the time point where an experimental manipulation is reflected in brain activity.

In this study, we aimed to conduct an initial mapping of the online correlates of entrenchment for the case of verb transitivity. We predicted 1. that offline behavioral ratings would, in conceptual replication of Ambridge et al. (2008), show increased acceptance of transitive uses of intransitive verbs for low- over high-frequency items; 2. that ERPs should show sensitivity – perhaps in the form of an attenuated P600 or LAN – already at the earliest position where the transitivity violation occurs, i.e., the position of the direct object. For this purpose, an auditory ERP experiment analogous to Ambridge et al. (2008) was implemented.

Specifically, we presented participants with intransitive verbs in transitive and intransitive context; i.e., intransitive contexts were ungrammatical. To test for entrenchment, we employed both high- and low-frequency verbs; the entrenchment hypothesis predicts ERP effects accompanying the interaction between grammaticality and frequency.

Methods

Stimulus Construction

A factorial 2 × 2 design was laid out with the factors Grammaticality (transitive vs. intransitive uses of intransitive verbs; **T/I**), and verb frequency (high vs. low frequency members of semantically matched verb pairs; **HF/LF**).

Verbs were selected based on meeting two criteria. First, to control for semantic properties (and thus provide a fair test of entrenchment), each verb was part of a pair of verbs with similar semantics (e.g., laugh/giggle). Second, each member of a verb pair differed in corpus frequency (according to Zipf SUBTLEX-UK frequency scores; Van Heuven, Mandera, Keuleers, & Brysbaert, 2014), see **Table 2**. Mean syllable counts (1.75) were equal for

Table 1: Verb pairs, with SUBTLEX frequency measures (Zipf scores), percent transitive out of all non-periphrastic occurrences, as well as the difference between low- and high-frequency scores per pair.

HF verb	Zipf score	% transitive	LF verb	Zipf score	% transitive	HF > LF
laugh	4.77	.4	giggle	3.55	.2	1.22
fall	5.01	.0	tumble	4.50	1.2	0.51
disappear	4.21	.0	vanish	3.25	.0	0.96
smile	4.71	9.3	grin	3.55	1.1	1.16

Table 2: Example sentences. Critical positions for EEG analysis are shown in italics.

	Condition	Example
(1)	**T/HF**	*On Wednesday, Bob laughed *the* girl in the kitchen.
(2)	**I/HF**	On Wednesday, Bob laughed *in* the kitchen.
(3)	**T/LF**	*On Wednesday, Bob giggled *the* girl in the kitchen.
(4)	**I/LF**	On Wednesday, Bob giggled *in* the kitchen.

both groups. For all verbs, intransitive occurrences were vanishingly rare (modal transitive counts = 0%; Bidgood, 2016). Although these criteria restricted our stimuli to eight verbs, they were necessary in order to be consistent with designs used in previous behavoral studies (e.g., Ambridge et al., 2008).

Then, English sentences were constructed following the form PP1 NP1 V (NP2) PP2. Verbs were always intransitive, so that all transitive constructions – where an NP was placed directly after the verb – were ungrammatical. Examples are shown in (1–4).

The critical position is the determiner for **T**ransitive sentences, and the second preposition (in) for **I**ntransitive ones (indicated in **Table 2** by *italics*). At this position, the transitivity violation became apparent for **T** sentences, while no such violation happened on **I** sentences. Importantly, these two items differ strongly in their lexical content and their syntactic implications. For this reason, **I** and **T** sentences can not be directly compared with on-line methods at this position, as this contrast would be highly confounded by lexical material (see e.g., Steinhauer & Drury, 2012). Specifically, it would contrast a preposition (*in*) with a determiner (*the*). Prepositions and determiners – and these words in particular – differ in multiple dimensions; for example, they license rather different continuations. Thus, the main contrast of grammaticality cannot be naively taken to be the cause behind any observed differences in the independent variables.

The experimental hypothesis, instead, referred instead to the interaction between the factors **T/I** and **HF/LF**. Specifically, the contrast between ungrammatical **T** and grammatical **I** sentences – the ungrammaticality effect – should be more pronounced for **LF** than for **HF** items.

10 sentences were constructed for each verb, resulting in 160 sentences total, with 40 per condition. Each verb was paired with 10 NP1s (one-syllable common English male names; *Bob, Scott …*), five initial PP1s (*On Monday, On Tuesday, … On Friday.*), and two sentence-final PP2s semantically matched to the verb pairs (i.e., *disappeared* was paired with *as if by magic* or *at the picnic*). To ensure as little variability as possible at critical positions, NP2 was always the same: *the girl*. Sentences were matched across all four conditions so that each combination of PP1 and NP1 occurred in all four conditions, and within verb pairs, selections of PP2s were matched.

A fifth verb pair was included in the design – *stay/wait* – but excluded from further analysis, because according to SUBTLEX-UK scores, frequencies of these words are actually nearly identical (5.37 vs. 5.39). No filler items were included; all sentences had essentially the same shape. On one hand, this highly repetitive design potentially isolates the critical manipulation, while attenuating other factors. On the other hand, this presentation form is very unlike ordinary language, and of most sentence processing experiments. However, previous studies have demonstrated that in many cases, highly repetitive lexical items (Renoult & Debruille, 2011) and syntactic constructions (Sassenhagen & Bornkessel-Schlesewsky, 2015; Sassenhagen et al., 2014) still induce what is often take to be the canonical correlates of, e.g., lexical and syntactic processing (N400 and P600).

Spoken sentences were recorded by a male native speaker of English, with natural prosody. To avoid acoustic cues on ungrammatical sentences, a cross-splicing technique was employed. For each set of sentences with shared lexical material (e.g., *On Wednesday, Bob … the girl in the kitchen.*), a suitable transitive verb was selected (e.g., *On Wednesday, Bob amused the girl in the kitchen.*). For experimental sentences, the transitive verb was replaced by a recording of the critical verb in the same context. For *I* sentences, NP2 was removed from the recording. Audio manipulations were conducted in Audacity (2.1.1; Audacity Team, 2015).

Experimental and EEG setup

Sentences were presented over loudspeakers via E-Prime 1.0. Presentation order was pseudo-randomised on each run, while ensuring sentences featuring the same verb or its matched pair never directly followed each other. On each trial, an asterisk appeared on a computer screen and the audio file started playing. 800 msec after sentence offset, a question mark appeared, prompting participants to press a button to indicate the grammaticality of the preceding sentence (yes or no).[3] Following the button press, the question mark was replaced by a feedback screen indicating the percentage of correct answers in order to ensure participant's attentiveness. 1000 msec later, the next trial was started. Trials were presented in blocks of 10, with a short break after each block. Including electrode preparation, each session lasted approximately 90 minutes.

While participants performed the task, EEG was recorded via a Biosemi Active-Two system featuring 64 electrodes positioned according to the 10–20 system. Two additional electrodes (CMS & DRL) featured as ground and online reference; four further electrodes were used to record horizontal and vertical EOG. An online bandpass from .16–100 Hz was applied, and data sampled at 1000 Hz.

20 undergraduate students (psychology, University of Manchester) participated in the experiment, receiving course credit. All were right-handed, monolingual English speakers, and consented to the experimental procedures after they had been sufficiently informed about them. The study was approved by the University of Manchester's ethics committee.

Behavioral analysis

The dependent variable for the analysis of acceptability judgements was the accuracy of judgements. A judgement was deemed correct if the participant had labelled a trial as acceptable if it was intransitive, or inacceptable if it was transitive, otherwise as incorrect. For visualisation purposes, for all four conditions, scores were averaged within subjects, means and 95% confidence intervals calculated, and plotted (Waskom et al., 2018).

To investigate if acceptability judgements were affected by the frequency manipulation, a hierarchical bayesian regression model was fit to the response accuracies. (Ambridge et al., 2008 had originally employed an ANOVA,

but since then, best-practices recommendations have begun emphasising the need to account for both stimulus and item random effects, as well as for direct modelling binary choices; see e.g., Jaeger, 2008). The model included the fixed effects Frequency (**HF/LF**), Grammaticality (**T/I**), and the interaction; as random effects, participant and verb pair were included. The model was built in the Python package Bambi (Yarkoni & Westfall, 2016), with default priors, and a logit link function (as the dependent variable is binary). Although it would have improved power (Cohen, 1983), it was decided not to include frequency as a continuous predictor 1. because no assumptions could be made about the specific shape of the frequency effect (which is unlikely to be linear), and 2. because it would complicate the control of semantics via the pairing of verbs, and 3. to keep analysis of EEG and behavioral data aligned, which additionally would have been infeasible to conduct with the mixed-model approach required to account for frequency as a continuous factor.

EEG analysis

Preprocessing

EEG analysis was conducted in MNE-Python (Gramfort et al., 2013). Data was downsampled to 200 Hz, and subjected to ICA decomposition (Jung et al., 2000). Artefactual components – blinks and horizontal eye movements – were identified via the semi-automatic Corrmap procedure (Viola et al., 2009), and removed from the data. Then, datasets were re-referenced to linked mastoids, leaving 61 channels.

Epochs were extracted around critical words, i.e., the first word after the verb (indicated in italics in **Table 1**). Recall that no direct contrast between **I** and **T** conditions is possible, because they differ in lexical and syntactic status. Instead, the interaction effects are of interest. Epochs consisted of the 300 msec preceding up to the 900 msec following the critical words.

Detection, interpolation and removal of artefactual channels and epochs was conducted via the fully automated Autorej tool (Jas, Engemann, Bekhti, Raimondo, & Gramfort, 2017). No epochs were rejected for incorrect answers, because we attempted to study correlates of certain syntactic constructions, regardless of conscious, explicit judgements (Osterhout & Mobley, 1995). Datasets with fewer than 75% trials remaining in any of the conditions (after fully automatic removal of artefactual data via Autorej) were rejected completely, leading to the exclusion of 4 data sets. Thus, 16 data sets remained for further analysis, with on average 38 (30–40) trials. Because these rejections were based on EEG-internal criteria, participants were not excluded from the behavioral analysis if their EEG data was rejected in this process. This means that EEG analysis and analysis of behavioral data do not refer to exactly the same sample.

Trials were averaged within conditions, resulting in one Event-Related Potential per condition per subjects, and a pre-stimulus baseline was subtracted. A Savitzky-Golay-filter, the default filter for evoked potentials implemented in MNE-Python, was applied for smoothing the waveforms.

Statistical Inference and Visualisation

For the visualisation of results, electrodes were grouped by Regions of Interest (Anterior/Posterior vs. Left/Midline/Right), 1 Standard Error of the mean was calculated, and across-subject grand-averages plotted (see **Figure 2**).

For statistical inference, for each dataset, two contrasts were calculated by subtraction and averaging: first, Grammaticality – all **T** vs. all **I**. This contrast was investigated to ensure that participants showed responses of on-line, incremental detection of syntactic violations at the expected position. However, note again that positions differed in their lexical content, entailing that fine-grained interpretations is not licensed, as they may result not from structural differences, but from the difference in lexical material. Second, the Grammaticality × Frequency interaction: **(T/LF − I/LF) − (T/HF − I/HF)**. This contrast contained the difference in the Grammaticality effect for high- vs. low-frequency verbs, i.e., the key contrast of this experiment. Frequency was not treated as a continuous factor to 1. not impair the pairing of semantically matched verbs, 2. enable a permutation-based approach to statistical inference (mixed-model estimation within the massively univariate framework would require a prohibitive number of models to be fitted, with results not straight-forwardly interpretable).

Both contrasts were separately subjected to a cluster-based permutation test for statistical thresholding (Maris & Oostenveld, 2007). These belong to the class of massively univariate tests, where, in absence of a motivation for testing a specific window, every individual time/sensor coordinate is subjected to a test, and the aggregated tests are subjected to correction for multiple comparisons. Cluster-based permutation tests exploit correlations across time and space to conduct massively univariate investigations while retaining sufficient power. We selected Threshold-Free Cluster Enhancement/TFCE (Mensen & Khatami, 2013) (as implemented in MNE-Python), because it minimizes researcher degrees of freedom on virtue of not having crucial parameters to tune, and because it allows voxel-level inference. Specifically, for both contrasts, first, a surrogate distribution under the null hypothesis was constructed. For this, over 1000 permutations, difference waveforms were randomly flipped in sign, averaged, and in the resulting grand average ERP, cluster-enhanced scores were calculated as laid out in Mensen & Khatami (2013). Then, the cluster-enhanced scores of the original data were collected, and compared against the surrogate values. Data points in the extreme tails of the surrogate distribution – corresponding to p values < .05, corrected for multiple tests – were marked. The resulting statistical significance masks were plotted over the grand average visualised as heatmaps (see **Figure 3**).

The main effect of Grammaticality was investigated as a manipulation check (as a lack of a Grammaticality effect would be highly surprising); a late positivity was expected (Osterhout & Holcomb, 1992). Post-hoc, after having seen the data, it was decided to provide an accessible summary of the pattern of results for the interaction effect. For this purpose, a spatial filter was created by averaging the Violation vs. Control difference waves in

the 600–800 msec time window across time and across subject. This resulted in a vector, with one number per channel, corresponding to the topographical pattern of the grammaticality effect. This was done in order to summarize the interaction effect. For each participant, for each condition, the time window from 200–400 msec post onset was selected and averaged across all time points. Then, the dot product between this vector and the grammaticality effect topographical pattern was calculated, resulting in a single number per participant per condition. This number corresponds to the strength of the Grammaticality pattern throughout the 200–400 msec time window, for all four conditions. The purpose of this linear reduction was to summarize the pattern of effects without having to manually decide on, e.g., an electrode to summarize the data at (Parra, Spence, Gerson, & Sajda, 2005). Means and confidence intervals were calculated, and the results plotted analogous to the behavioral results. Remember this was done exclusively for visualisation purposes, and bears no inferential value (Vul, Harris, Winkielman, & Pashler, 2009).

Finally, we investigated to what extent the crucial interaction effect changes over time. For this, we binned each participants' trials into quintiles by experiment time (i.e., first fifth of trials, second fifth ...), averaged trials from this bin by condition, calculated the interaction effect as above, and extracted the pattern strength as described above. 95% confidence intervals over participants were calculated. The linear correlation between quintile and interaction effect pattern strength was calculated for each dataset, and a rank-sum test applied to investigate if the correlations deviated significantly from zero. The purpose of this analysis – motivated by comments of our anonymous reviewers – was to investigate if the highly repetitive nature of the stimulus material was underlying the pattern of results; for example, if it were observed that the effects occurred only in the later time bins, strategic processing effects could be assumed to underlie the results. Conversely, if a negative time trend could be observed, the repetitive nature of the stimuli might suppress any potential real effects.

Results

Behavioral results

Rating accuracies were near perfect for all conditions with the exception of low frequency violations, which were rated to be acceptable in >6.5% of cases; error rates were: Control, High Freq.: 0.93%, Low Freq. 1.32%. Violation, High Freq.: 1.58%, Low Freq.: 6.58%. See **Figure 1**, right. Pointing towards the statistical reliability of these findings, Bayesian modelling (summarized in **Table 3**) did not indicate a main effect of Frequency, nor one of Grammaticality, but the Credible Interval for the coefficient for the Frequency × Grammaticality interaction exceeded zero – although only weakly (mean: 1.283, SD: .623). This is in agreement with Ambridge et al. (2008), who report a similar Grammaticality × Frequency interaction in a graded acceptability task.

EEG results

For the main effect of Grammaticality (**T/I**), ERPs (**Figure 2**) prominently showed a late component consisting of a parietal positivity and a frontal negativity, peaking between 600–800 msec. Cluster-based permutation testing with

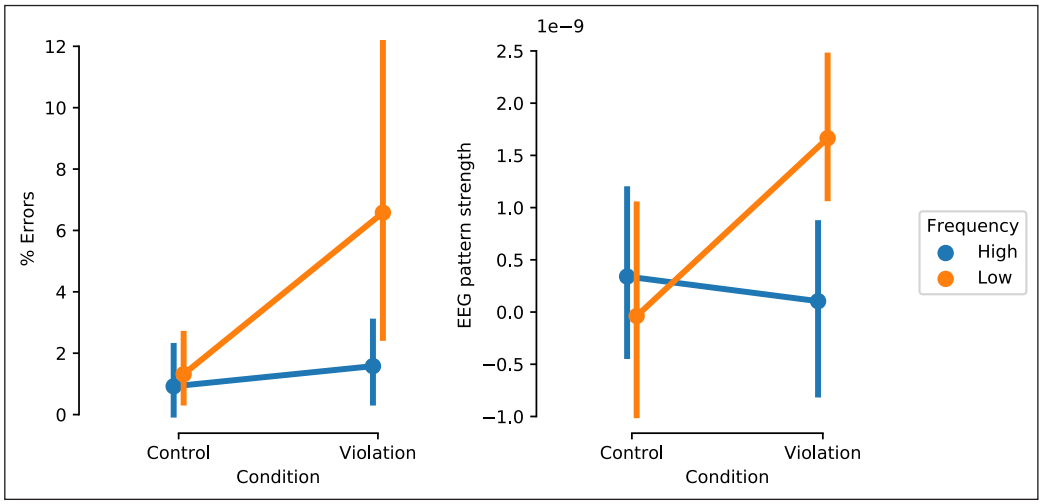

Figure 1: Left: Mean response accuracy per condition. Error bars reflect 95% bootstrapped confidence intervals across subjects. Right: as before, but EEG pattern strength (i.e., the averaged occurrence of the P600 effect throughout the time window where the interaction effect is significant; see text for details) per condition. Note that Violation and Control conditions should not be directly compared to each other (see text).

Table 3: Response accuracy modelling results.

	Mean	SD	95% CI upper	95% CI lower
Frequency	0.375	0.514	–0.571	1.418
Grammaticality	0.575	0.552	–0.397	1.621
Frequency × Grammaticality	1.283	0.623	0.027	2.448

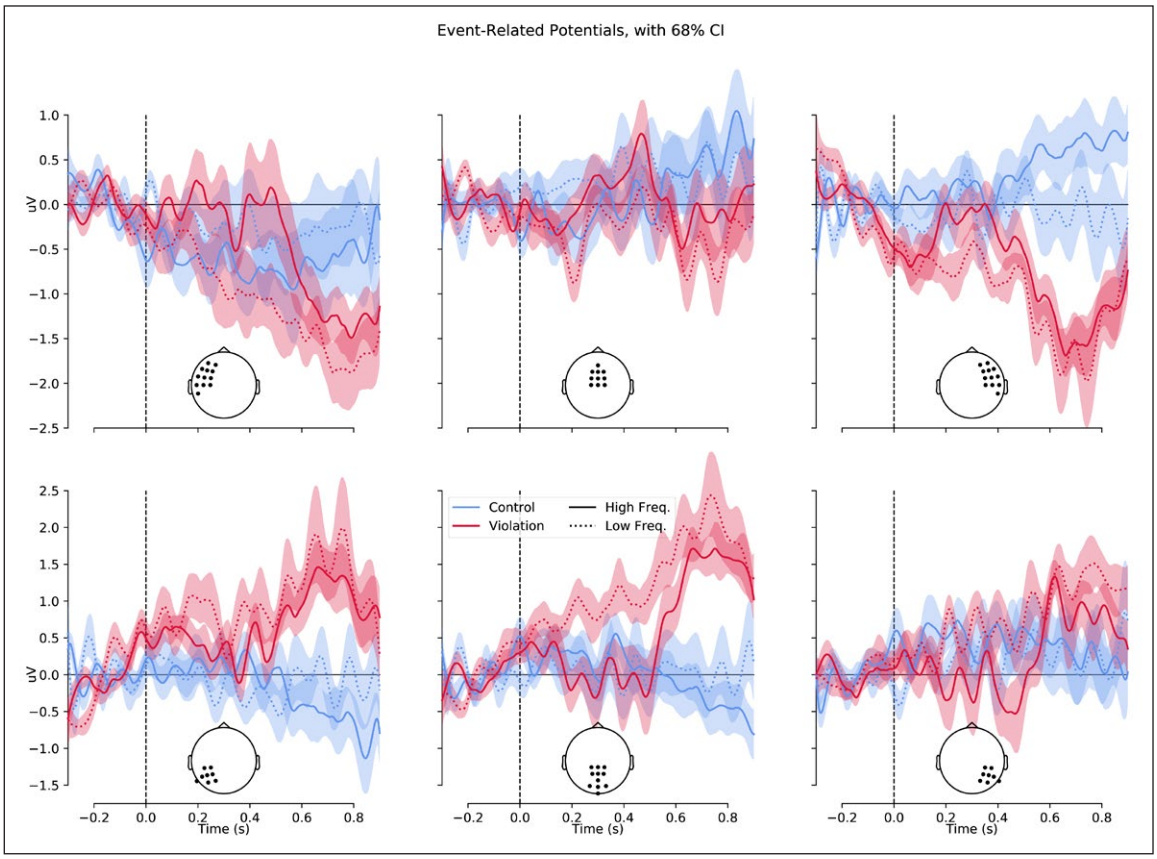

Figure 2: Grand average ERPs per condition, grouped by 6 Regions of Interest, with 68% Confidence Intervals.

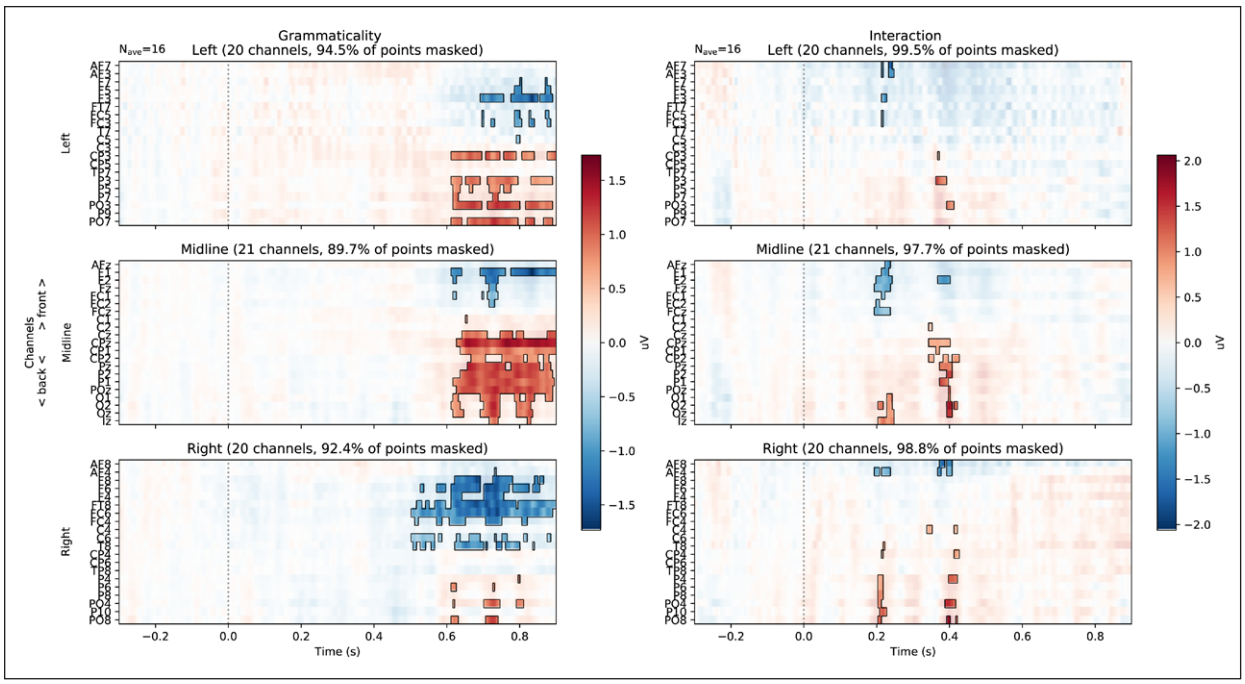

Figure 3: ERPimages (grand averaged activity over time for each channel, plotted as heatmaps) for both contrast, masked for statistical significance. Channels are grouped by hemisphere, and sorted from back to front.

TFCE (**Figure 3**) indicated the statistical significance of this effect ($p < .05$) – although note again that this effect is hard to interpret due to the divergent lexical material. The interaction effect (e.g., the difference between transitive uses of high- vs. low-frequency verbs) exhibited a similar pattern exclusively for low-frequency violations in an earlier time window (approx. 200–400 msec). While much less extensively distributed across time and space, TFCE also indicated this contrast to be statistically significant ($p < .05$).

Visualising the form of the interaction effect by quantifying the strength of the late-window violation

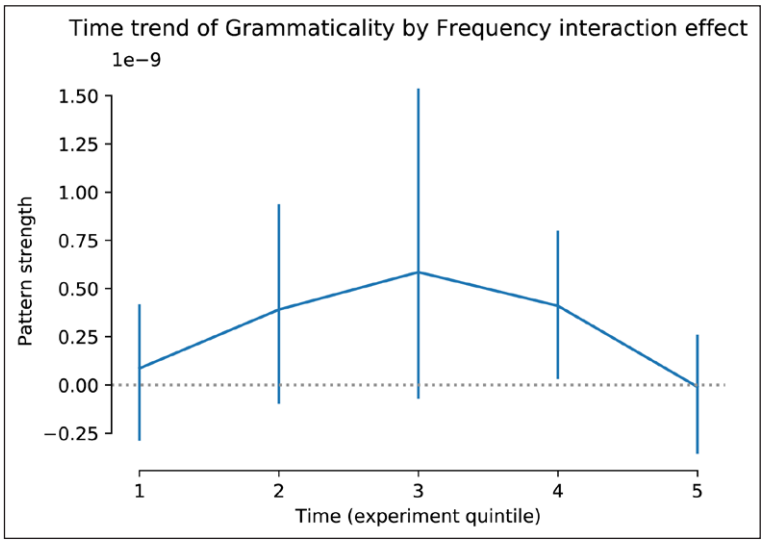

Figure 4: Change of Grammaticality × Frequency interaction effect strength (summarized as in Figure 1) over the time course of the experiment, i.e., binned by experimental quintile. Error bars show 95% confidence intervals across participants. There is no clear time trend; the average correlation is. 1 ($p = 0.41$).

effect indicated that a similar pattern as in the late time window was also observed in the early time window in contrast between low frequency violations (where it was stronger) and high frequency violations (see **Figure 1**).

There was no clear time trend (see **Figure 4**). While the interaction effect was nominally positive in the first four time bins (and slightly below zero in the last). The (again, nominally) strongest effect occurred in the middle bin. The correlation between time bin quintile and the strength of the interaction effect did not significantly diverge from zero ($r = .1$, $p = 0.41$).

Discussion

To investigate reflections of entrenchment resulting from statistical learning of syntactic constructions during online language processing, we conducted an experiment resembling Ambridge et al. (2008), but measuring EEG while participants listened to spoken sentences. Behavioral results indicate that rating transitive uses of intransitive verbs is sensitive to verb corpus frequency, with up to 6.5% of low-frequency intransitive verbs rated as acceptable (supported by the Grammaticality × Frequency interaction effect). ERPs indicated a similar pattern. In addition to a P600-like response to the transitivity violation, the Grammaticality × Frequency interaction induced an early ERP difference. Post-hoc attempts to visualise the nature of this effect indicate that it can be understood as a P600-like pattern (albeit much earlier; 200–400 msec), stronger for low- than for high-frequency verbs.

A recent meta-analysis of 19 offline-grammaticality-judgment datasets (Ambridge, Barak, Wonnacott, Bannard, & Sala, 2018) found strong evidence for the existence of an entrenchment effect on verb-argument-structure overgeneralization errors. That is, even after controlling for verb semantics and frequency in particular constructions, the overall frequency of a particular verb was shown to influence participants' judgments, (such that, for example, as sentence such as "*Bob laughed the girl" is rated as less acceptable than "*Bob giggled the girl").

The aim of the present study was to investigate whether this well-established behavioral effect (also observed in the behavioral responses in this study) can be observed using an online EEG paradigm and, if so, whether the specific morphology of neural effects can further inform about the cognitive processes underlying this frequency sensitivity. The findings were somewhat equivocal. Although the EEG data did suggest that participants exhibit sensitivity to verb frequency when encountering argument-structure overgeneralizations, this effect was not easily interpretable in that it did not unambiguously appear as any specific well-known ERP component, and the more probable candidates did not unambiguously suggest any one interpretation.

Speculations on underlying neurocognition

While the observation of an interaction effect in the ERP was as predicted by the entrenchment account, its specific nature was not as expected. It did not appear in the form of a modulation of the P600, nor did it clearly reflect as an LAN. The effect was not left lateralized (otherwise, it would have mostly reflected in the top panel in **Figure 3**, and in the top left panel in **Figure 2**). It also appeared too early to be a modulation of the P600 (200–400 msec, rather than 600–800 msec).

Remember that one attempt to summarize the pattern of results is that an EEG pattern similar to the P600 marked the interaction contrast in a much earlier time window. As noted, the observed interaction effect was revealed by a massively univariate test with cluster-based permutation control for multiple tests, i.e., a procedure largely robust to experimenter degrees of freedom (as no parameters where tuned, e.g., no time windows or electrodes selected manually). Yet, the resulting pattern is hard to interpret 1. because the decision to summarize the data was made post-hoc, having seen the data, 2. because the directionality of an effect cannot reliably be made based on the data alone – e.g., perhaps the effect is a parietal positivity for high frequency verb violations, or an anterior negativity with

a scalp topography similar to, but an underlying neural substrate very different from the P600 effect, 3. because the violation and the control ERPs cannot be directly compared due to differences in lexical items.

However, if taken at face value, if the late positivity is understood to be an index of syntactic error detection, then arguably, the observed pattern points in the wrong direction; low-frequency violations show a stronger pattern than high-frequency verbs, although participants were more committed to categorizing the latter as ungrammatical. However, it has been questioned to what degree the P600 is an index of syntactic violations in themselves. For example, it has been suggested to reflect processing costs during syntactic integration (Kaan, Harris, Gibson, & Holcomb, 2000), and index the integration of new referents into the ongoing discourse (Burkhardt, 2007). I.e., perhaps the parser, when encountering an NP following an intransitive low-frequency verb, is initially willing to open a new grammatical or discourse slot, but not for high-frequency verbs, where parsing is simply interrupted (with the later P600-like effect reflecting an attempt to repair the broken parse). However, all such interpretations are highly speculative, especially as long as the functional interpretation of the late positivity is debated.

Given 1. the topographical similarity between the early (200–400 msec) and late (>600 msec) effects, and 2. the timing of the early effect, it could be speculated to be an instance of the P300 component (Sutton, Braren, Zubin, & John, 1965). The P300 (for reviews, see Nieuwenhuis, Aston-Jones, & Cohen, 2005; Polich, 2007) is negatively correlated with the probability of a stimulus, i.e., particularly surprising constellations would be expected to induce a P300. It is not unequivocally clear how this would fit with our findings. Supposedly, high-frequency violations should be the least predictable/probable condition, and thus elicit a P300. Instead, low-frequency violations show a more positive EEG in this time window. This could, again, be taken, while corroborating the general idea that word frequency influences grammaticality, as indicating that this influence goes in the opposite direction of that suggested by the entrenchment hypothesis. However, more recent interpretations of the P300 (Nieuwenhuis et al., 2005; O'Connell, Dockree, & Kelly, 2012) generally argue the P300 does not simply mark probability, but indexes decision making; the correlation with probability is indirect. This effectively discourages simply taking the P300 as a marker of, e.g., which condition out of a set is more surprising. Similarly, the decision making interpretation of the P300 does not strongly constrain the possible interpretations of our results. Both a less predictable and a more predictable construction could be argued to require a decision (e.g., a decision to commit to an interpretation, or to revise an interpretation).

Note also that – as has already been hinted at above – it has been suggested the P600 shares its neural substrate with the P300 (Coulson et al., 1998b; Sassenhagen et al., 2014), so to some extent, a P300-based and a P600-based interpretation might resemble each other strongly.

Limitations

A premier limitation of this study is the small sample size. Only 20 subjects could be recorded, of which 4 had to be dropped, leaving an uncomfortably low sample size of 16. This means all estimations are highly imprecise; it is possible that the major neural consequences of the tested manipulation were not captured in a representative manner. (However, of course the false positive rate is unaffected by low sample size.)

The stimulus set employed was highly repetitive, and contained no fillers. We see no sensible path by which this could inflate effects resulting from this manipulation (e.g., if processing becomes more automatic and lexical items are repeated constantly, if at all, frequency effects should decrease, not increase). However, it is possible to have attenuated effects; potentially, it might have obscured important neural correlates of frequency-sensitive processing. Our investigation of the time trend of effects did, if at all, support this later interpretation; the smallest effect was observed for the latest trials in the experiment. I.e., the effect did not emerge over the time course of the experiment (perhaps as participants incrementally build up processing strategies).

A much larger item base would also have allowed an initial attempt to map the dose-dependence of the Grammaticality × Frequency interaction. This curve is likely non-linear. I.e., it is likely that effects "bottom out" in the higher range of verb frequency; presumably, there is little difference between a verb within the 99th and one in the 98th percentile of frequency. As is, we have only tracked the difference for a small group of semantically matched pairs which categorically differ in frequency (note that *tumble*, in the low frequency group, is actually more common than *disappear*, in the high frequency group, according to SUBTLEX scores). It is also possible that the SUBTLEX-UK corpus, while highly regarded and validated, is not the appropriate measure for this analysis; perhaps a corpus more strongly slanted towards younger ages could more accurately model preferences.

We also note that the present analysis was not pre-registered. Different analysis choices could have been made, many of which would have been defensible. We chose a conservative, exploratory method for the analysis of ERPs here – cluster-based permutation tests – but it is possible that another, equally well justified approach could have led to different conclusions. This entails the need for a pre-registered, high-powered replication, in part to validate the results, in part to more precisely track the nature of the frequency sensitivity of verb subcategorization violation processing – i.e., the shape of the dose-dependence of the frequency effect should be mapped by exploring a broad range of verbs, spread across the frequency range, while still keeping track of, e.g., semantic and phonological differences.

Conclusions

We provide initial evidence that online processing of syntactic constructions is already sensitive to word frequency. This adds to the evidence on reflections of statistical learning even in the adult parser. Further research should explore

the precise neurocognitive form of this sensitivity, on larger, more variable samples of items and stimuli.

Data Accessibility Statements

All analyses were conducted with custom Python scripts, using the iPython platform (Pérez & Granger, 2007). Data and the underlying Jupyter notebooks – containing reproducible code for all analyses – are made available on github.[4] This repository also links to the data required for reproducing the analyses.

Notes

[1] Pullum (2013) in fact provides an explanation in terms of Bayes' Rule, but we think expressing it in terms of conditional probabilities is somewhat more general.

[2] Neither do we attempt to distinguish entrenchment from a similar proposal, preemption (e.g., Goldberg, 2003), under which what is relevant is not overall verb frequency, but frequency in particular constructions (see Ambridge, Barak, Wonnacott, Bannard & Sala, 2018, for an attempt to distinguish the two).

[3] Note that in Ambridge et al. (2008), participants conducted a more fine-grained graded estimation task. Here, a simplified version was chosen in order to reduce the complexity of the task for participants, who had already undergone preparation for EEG measurements.

[4] github.com/jona-sassenhagen/sassenhagen_blything_lieven_ambridge_collabra.

Additional Files

The additional files for this article can be found as follows:

- **Fig 1: Dependent variables, summarized, split by condition.** Shows response accuracies and a (nonindependent!) estimate of ERP effect size for the Grammaticality X Frequency contrast. DOI: https://doi.org/10.1525/collabra.87.s1
- **Fig 2: Event-related potentials, separated by Region of Interest.** Shows ERP time courses for each condition, plus a confidence interval, aggregating over six channel groups. DOI: https://doi.org/10.1525/collabra.87.s2
- **Fig 3: ERP "Image" for critical contrasts.** Shows ERP time courses for each channel for Grammaticality main effect and Grammaticality X Frequency interaction effect, with masking for statistical significance, split into three channel groups. DOI: https://doi.org/10.1525/collabra.87.s3
- **Fig 4: Time course of interaction effect.** Shows the Grammaticality X Frequency interaction for 5 time bins (i.e., early vs. late in the experiment). DOI: https://doi.org/10.1525/collabra.87.s4
- **Table S1.** Table of Test Sentences. DOI: https://doi.org/10.1525/collabra.87.s5

Ethics and Consent

We confirm that we have read the Journal's position involved in ethical publication and affirm that this manuscript is consistent with those guidelines.

Acknowledgements

Elena Lieven and Ben Ambridge are Professors in the ESRC International Centre for Language and Communicative Development (LuCiD). The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. This work was supported by an Economic and Social Research Council Doctoral Training Centre +3 (Ph.D.) award to Ryan Blything. http://www.esrc.ac.uk/. Account Code: ES/J500094/1. JS was supported by ERC grant 617891 to Christian J. Fiebach.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

JS wrote the manuscript and analysed the data. RB designed the stimuli and conducted the experiment. RB, EVML and BA designed the experiment.

References

Ambridge, B., Barak, L., Wonnacott, E., Bannard, C., & **Sala, G.** (2018). Effects of both preemption and entrenchment in the retreat from verb overgeneralization errors: Four reanalyses, an extended replication, and a meta-analytic synthesis. *Collabra: Psychology*, *4*(1). DOI: https://doi.org/10.1525/collabra.133

Ambridge, B., Pine, J. M., Rowland, C. F., Chang, F., & **Bidgood, A.** (2013). The retreat from overgeneralization in child language acquisition: Word learning, morphology, and verb argument structure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(1), 47–62. DOI: https://doi.org/10.1002/wcs.1207

Ambridge, B., Pine, J. M., Rowland, C. F., & **Young, C. R.** (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*, 87–129. DOI: https://doi.org/10.1016/j.cognition.2006.12.015

Bidgood, A. (2016). *The retreat from overgeneralisation errors: A multiple-paradigm approach* (PhD thesis). University of Liverpool.

Bornkessel, I. D., & **Schlesewsky, M.** (2006). The extended argument dependency model: A neurocognitive approach to sentence comprehension across languages. *Psychological Review*, *113*(4), 787–821. DOI: https://doi.org/10.1037/0033-295X.113.4.787

Bornkessel-Schlesewsky, I., & **Schlesewsky, M.** (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews*, *59*(1), 55–73. DOI: https://doi.org/10.1016/j.brainresrev.2008.05.003

Braine, M. D. S., & **Brooks, P.** (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In: Tomasello, M., & Merriman, W. E. (Eds.), *Beyond names for things: Young children's acquisition of verbs*, 353–376. Hillsdale, NJ: Erlbaum.

Burkhardt, P. (2007). The P600 reflects cost of new information in discourse memory. *Neuroreport*,

18(17), 1851–1854. DOI: https://doi.org/10.1097/WNR.0b013e3282f1a999

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253. DOI: https://doi.org/10.1177/014662168300700301

Coulson, S., King, J. W., & **Kutas, M.** (1998a). ERPs and domain specificity: Beating a straw horse. *Language and Cognitive Processes*, *13*(6), 653–672. DOI: https://doi.org/10.1080/016909698386410

Coulson, S., King, J. W., & **Kutas, M.** (1998b). Expect the Unexpected: Event-related Brain Response to Morphosyntactic Violations. *Language and Cognitive Processes*, *13*(1), 21–58. DOI: https://doi.org/10.1080/016909698386582

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Science*, *6*(2), 78–84. DOI: https://doi.org/10.1016/S1364-6613(00)01839-8

Friederici, A. D., Hahne, A., & **Mecklinger, A. A.** (1996). Temporal structure of syntactic parsing: Early and late event-related brain potential effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1219–1248. DOI: https://doi.org/10.1037/0278-7393.22.5.1219

Friederici, A. D., Mecklinger, A. A., Spencer, K. M., Steinhauer, K., & **Donchin, E. E.** (2001). Syntactic parsing preferences and their on-line revisions: A spatio-temporal analysis of event-related brain potentials. *Cognitive Brain Research*, *11*(2), 305–323. DOI: https://doi.org/10.1016/S0926-6410(00)00065-3

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Science*, *7*(5), 219–224. DOI: https://doi.org/10.1016/S1364-6613(03)00080-9

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Hämäläinen, M., et al. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*. DOI: https://doi.org/10.3389/fnins.2013.00267

Hastie, T., Tibshirani, R., & **Friedman, J. H.** (2009). *The Elements of Statistical Learning*. Springer. DOI: https://doi.org/10.1007/978-0-387-84858-7

Jaeger, F. T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. DOI: https://doi.org/10.1016/j.jml.2007.11.007

Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & **Gramfort, A.** (2017). Autoreject: Automated artifact rejection for meg and eeg data. *NeuroImage*, *159*, 417–429. DOI: https://doi.org/10.1016/j.neuroimage.2017.06.030

Jung, T.-P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & **Sejnowski, T. J.** (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178. DOI: https://doi.org/10.1111/1469-8986.3720163

Kaan, E., Harris, A., Gibson, E., & **Holcomb, P.** (2000). The P600 as an index of syntactic integration difficulty. *15*(2), 159–201. DOI: https://doi.org/10.1080/016909600386084

Lidz, J., & **Gagliardi, A.** (2015). How nature meets nurture: Universal grammar and statistical learning. *Annu. Rev. Linguist.*, *1*(1), 333–353. DOI: https://doi.org/10.1146/annurev-linguist-030514-125236

Lieven, E. V. M. (1994). Crosslinguistic and crosscultural aspects of language addressed to children. In: Gallaway, C., & Richards, B. J. (Eds.). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511620690.005

Luck, S. J. (2005). *An introduction to the event-related potential technique*. The MIT Press.

Maris, E., & **Oostenveld, R.** (2007). Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods*, *164*(1), 177–190. DOI: https://doi.org/10.1016/j.jneumeth.2007.03.024

Mensen, A., & **Khatami, R.** (2013). Advanced eeg analysis using threshold-free cluster-enhancement and non-parametric statistics. *NeuroImage*, *67*(Supplement C), 111–118. DOI: https://doi.org/10.1016/j.neuroimage.2012.10.027

Nieuwenhuis, S., Aston-Jones, G., & **Cohen, J. D.** (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. *Psychological Bulletin*, *131*(4), 510–532. DOI: https://doi.org/10.1037/0033-2909.131.4.510

O'Connell, R. G., Dockree, P. M., & **Kelly, S. P.** (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*(12), 1729–1735. DOI: https://doi.org/10.1038/nn.3248

Osterhout, L., & **Holcomb, P. J.** (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. DOI: https://doi.org/10.1016/0749-596X(92)90039-Z

Osterhout, L., & **Mobley, L. A.** (1995). Event-Related Potentials Elicited by Failure to Agree. *Journal of Memory and Language*, *34*(6), 739–773. DOI: https://doi.org/10.1006/jmla.1995.1033

Parra, L. C., Spence, C. D., Gerson, A. D., & **Sajda, P.** (2005). Recipes for the linear analysis of EEG. *Neuroimage*, *28*(2), 326–341. DOI: https://doi.org/10.1016/j.neuroimage.2005.05.032

Pérez, F., & **Granger, B. E.** (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, *9*(3), 21–29. DOI: https://doi.org/10.1109/MCSE.2007.53

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283. DOI: https://doi.org/10.1016/0010-0277(79)90001-5

Poldrack, R. A. (2011). Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding. *Neuron*, *72*(5), 692–697. DOI: https://doi.org/10.1016/j.neuron.2011.11.001

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *118*(10), 2128–2148. DOI: https://doi.org/10.1016/j.clinph.2007.04.019

Pullum, G. K. (2013). The central question in comparative syntactic metatheory. *28*. DOI: https://doi.org/10.1111/mila.12029

Rayner, K., & **Clifton, C. Jr.** (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, *80*(1), 4–9. DOI: https://doi.org/10.1016/j.biopsycho.2008.05.002

Renoult, L., & **Debruille, J. B.** (2011). N400-like potentials and reaction times index semantic relations between highly repeated individual words. *Journal of Cognitive Neuroscience*, *23*(4), 905–922. DOI: https://doi.org/10.1162/jocn.2009.21410

Sassenhagen, J., & **Bornkessel-Schlesewsky, I.** (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex*, *66*, A3–A20. DOI: https://doi.org/10.1016/j.cortex.2014.12.019

Sassenhagen, J., Schlesewsky, M., & **Bornkessel-Schlesewsky, I.** (2014). The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language*, *137*, 29–39. DOI: https://doi.org/10.1016/j.bandl.2014.07.010

Steinhauer, K., & **Drury, J. E.** (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language*, *120*(2), 135–162. DOI: https://doi.org/10.1016/j.bandl.2011.07.001

Sutton, S., Braren, M., Zubin, J., & **John, E. R.** (1965). Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, *150*(700), 1187. DOI: https://doi.org/10.1126/science.150.3700.1187

van de Meerendonk, N., Kolk, H. H. J., Chwilla, D. J., & **Vissers, C. T. W. M.** (2009). Monitoring in Language Perception. *Language and Linguistics Compass*, *3*(5), 1211–1224. DOI: https://doi.org/10.1111/j.1749-818X.2009.00163.x

Van Heuven, W. J., Mandera, P., Keuleers, E., & **Brysbaert, M.** (2014). SUBTLEX-uk: A new and improved word frequency database for british english. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. DOI: https://doi.org/10.1080/17470218.2013.850521

Viola, F. C., Thorne, J., Edmonds, B., Schneider, T., Eichele, T., & **Debener, S.** (2009). Semi-automatic identification of independent components representing EEG artifact. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *120*(5), 868–877. DOI: https://doi.org/10.1016/j.clinph.2009.01.015

Vul, E., Harris, C., Winkielman, P., & **Pashler, H.** (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. DOI: https://doi.org/10.1111/j.1745-6924.2009.01125.x

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Qalieh, A., et al. (2018, July). Mwaskom/seaborn: V0.9.0 (july 2018). DOI: https://doi.org/10.5281/zenodo.1313201

Yang, C. (2011). Computational models of syntactic acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *3*(2), 205–213. DOI: https://doi.org/10.1002/wcs.1154

Yarkoni, T., & **Westfall, J.** (2016). Bambi: A simple interface for fitting bayesian mixed effects models.

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.87.pr

How to cite this article: Sassenhagen, J., Blything, R., Lieven, E. V. M., & Ambridge, B. (2018). Frequency Sensitivity of Neural Responses to English Verb Argument Structure Violations. *Collabra: Psychology*, 4(1): 38. DOI: https://doi.org/10.1525/collabra.87

Senior Editor: Rolf Zwaan

Editor: Fernanda Ferreira

Submitted: 11 March 2017 **Accepted:** 29 September 2018 **Published:** 30 October 2018

UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS