UNIVERSITY of CALIFORNIA PRESS | Collabra

Ambridge, B., et al. (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): 23. DOI: https://doi.org/10.1525/collabra.133

ORIGINAL RESEARCH REPORT

Effects of Both Preemption and Entrenchment in the Retreat from Verb Overgeneralization Errors: Four Reanalyses, an Extended Replication, and a Meta-Analytic Synthesis

Ben Ambridge[*,†], Libby Barak[‡], Elizabeth Wonnacott[§], Colin Bannard[*] and Giovanni Sala[‖]

How do speakers avoid producing verb overgeneralization errors such as **She covered paint onto the wall* or **She poured the cup with water?* Five previous papers have found seemingly contradictory results concerning the role of *statistical preemption* (competition from acceptable alternatives such as *She covered the wall with paint* or *She poured water into the cup*) and *entrenchment* (a mechanism sensitive to all uses of the relevant verb). Here, we use more appropriate measures of preemption and entrenchment (attraction measures based on the chi-square statistic, as opposed to using only the frequency of occurrence in favoured constructions) as well as more appropriate statistical analyses and, in one case, a larger corpus to reanalyse the data from these studies. We find that for errors of verb argument structure overgeneralization (as in the examples above), preemption/entrenchment effects are almost always observed in single-predictor models, but are rarely dissociable, due to collinearity. Fortunately, this problem is much less acute for errors of reversative *un-* prefixation (e.g., **unsqueeze; *uncome*), which could in principle be blocked by (a) non-reversative uses of the same verb root (e.g., *squeeze, come*; entrenchment), and/or (b) lexically-unrelated verbs with similar meanings to the relevant *un-* forms (e.g., *release, go*; preemption). Across a reanalysis of two previous studies of *un-* prefixation, and a new extended replication with adults, we find dissociable effects of both preemption and entrenchment. A meta-analytic synthesis revealed that, across the studies, both effects are reliable, though preemption appears to increase with age. We conclude that a successful account of the retreat from verb overgeneralization is likely to be one that yields preemption and entrenchment as effects that fall naturally out of the learner's attempts to communicate meaning, rather than one that treats these effects as mechanisms in their own right, and discuss current accounts that potentially meet this criterion. Finally, we set out some methodological recommendations that can be profitably applied not only to corpus-based experimental studies, but studies of child language acquisition in general.

Keywords: child language acquisition; verb argument structure; locative; dative; morphology; preemption; entrenchment; competition; semantics; discriminative learning; Rescorla-Wagner

Introduction

A defining characteristic of language is its productivity (Humboldt, 1836; Chomsky, 1957): Speakers do not simply maintain an inventory of rote-learned utterances, but rather form generalizations that allow them to produce novel utterances. For example, although there is considerable disagreement as to the precise nature of the processes and representations involved, there must be some mechanism that allows English-speaking children to take a verb that they have heard only in one construction and use it productively in another construction:[1]

Figure locative (caused motion)
[NP] [VERB] [NP] ([PP])
She sprayed paint (onto the wall).

Ground locative (causative)[2]
[NP] [VERB] ([NP] [PP])
She sprayed the wall (with paint).

* University of Liverpool, Liverpool, UK
† ESRC International Centre for Language and Communicative Development (LuCiD), SE
‡ Princeton University, Princeton, New Jersey, US
§ University College London, London, UK
‖ Osaka University, Osaka, JP
Corresponding author: Ben Ambridge (ben.ambridge@liverpool.ac.uk)

At the same time, adult generalization is often restricted such that some novel combinations of lexical items and constructions are deemed to be ungrammatical (or, at least, somewhat less than fully acceptable to most adult speakers).[3] In the course of development, children may not follow these restrictions, leading to overgeneralization errors, such as the use of cover in the figure locative (e.g., **I'm gonna cover a screen over me* [age 4;5]) or *pour* in the ground locative (e.g., **Mommy, I poured you…with water* [age 2;11]; examples from Bowerman, 1988; see also Braine, 1971; Baker, 1979; Pinker, 1989; Brooks & Tomasello, 1999; Bowerman, 1988).

The question of how learners restrict their generalizations lies at the heart of language acquisition research (Bowerman, 1988). There is widespread agreement that semantic factors play an important role (e.g., Ambridge et al., 2008; Fisher, Gleitman & Gleitman, 1991; Goldberg, 1995; Pinker, 1989). For example, the ground-locative construction is more acceptable for verbs like *cover* that are construed to cause a change of state (e.g., something becoming covered) than for verbs that do not, (e.g., *pour*) (Gropen et al., 1991; Bidgood et al., 2014). The analyses presented here confirm the importance of these semantic factors, but focus mainly on the statistical distribution of verbs. The reason for this focus is that the previous studies that we reanalyse here (all involving the first author), provided suggestive evidence for two statistical learning hypotheses – preemption and entrenchment – but struggled both to accurately operationalize and to dissociate these predictors.

The ***entrenchment*** hypothesis (e.g., Braine & Brooks, 1995) states that repeated presentation of a verb (e.g., *fill*), regardless of construction (e.g., *The tub filled up; She filled the bowl; She filled out the form*), causes the learner to make a probabilistic inference that the use of this verb in non-attested constructions is unacceptable (e.g., **She filled the water into the tub*). Intuitively, entrenchment can be understood as the inference that, once a verb has been witnessed thousands of times, any constructions in which it remains unattested – or attested only very infrequently – are probably less than fully grammatical for that verb (otherwise, given its overall frequency, it would presumably have appeared in these constructions by now; Hahn & Oaksford, 2008; Perfors, Tenenbaum, & Wonnacott, 2010; Perfors & Wonnacott, 2011). In support of this proposal, judgment studies (e.g., Theakston, 2004; Stefanowitch, 2008) have shown that the unacceptability of errors is positively correlated with verb frequency, since attested uses strengthen this "inference from absence". For example, **She filled water into the tube* is judged to be (even) worse than **She infused water into the tube*.

Statistical Preemption (Goldberg, 1995, 2006) is also a probabilistic learning process, but differs in that errors are blocked not by the use of the relevant verb in *any* construction (as for entrenchment), but only by constructions that constitute a relatively close paraphrase. Intuitively, preemption can be understood as the listener witnessing one particular formulation (e.g., *The hose spewed water onto the floor*) in a context where an alternative formulation would have suited the speaker's communicative intentions as well or better (e.g., **The hose spewed the floor with water*), and, as a consequence, learning that the former is preferred over the latter. The fact that, all else being equal, novel uses of more frequent verbs are less acceptable than novel uses of less frequent verbs, on this view, is not due to the fact that one verb is more frequent *overall*, but instead is due to the more frequent verb having been witnessed more often in a directly competing construction (Robenalt & Goldberg, 2015, 2016; Goldberg, 2011). Thus, on this account, apparent entrenchment effects have been found only because verbs that are highly frequent regardless of construction (entrenchment) are usually highly frequent in the constructions that compete with possible errors (preemption).

Readers unfamiliar with this literature are invited to consider the following analogy, which is designed to explain more intuitively the difference between entrenchment and preemption.[4] Suppose that a naïve observer is trying to figure out whether it is acceptable to use the name *Lizzy* when addressing the Queen of the United Kingdom (analogous to trying to figure out whether it is permissible to use *spew* in the ground locative construction; e.g., **The hose spewed the floor with water*).

- **Entrenchment** is summarized by the following internal monologue: "I've heard the name *Lizzy* used hundreds of times. Yet never, in all the royal greetings I've observed, have I heard someone address the Queen as *Lizzy*. Surely if this *were* allowed, I would have heard it by now. I will now therefore tentatively assume that it is not allowed".
- **Preemption** is summarized by the following internal monologue: "In all the royal greetings I've observed, people have addressed the Queen as *Your Majesty* and never as *Lizzy*, even though the latter would seem to convey the desired meaning (i.e., it *is* her name). I will now therefore tentatively assume that *Your Majesty*, rather than *Lizzy*, is the (more) permissible form of conveying this meaning (i.e., addressing the Queen).

To complete the analogy, consider a naïve observer who is trying to figure out whether it is acceptable to use *spew* in the ground locative construction (e.g., **The hose spewed the floor with water*). In fact (as the conventional asterisk indicates), it is not.

- **Entrenchment:** "I've heard *spew* used hundreds of times. Yet never, in all of the ground locative constructions (*[PERSON] [VERBed] [LOCATION] with [SUBSTANCE]*) I've observed, have I heard someone use *spew*. Surely if this *were* allowed, I would have heard it by now. I will now therefore tentatively assume that it is not allowed".
- **Preemption:** "For all of the spewing descriptions I have observed, people have said *[PERSON] spewed [SUBSTANCE] onto [LOCATION]* and never *[PERSON] spewed [LOCATION] with [SUBSTANCE]*. I will now

therefore tentatively assume that the former is the (more) permissible way of describing spewing events.

Or, for verbal *un-* prefixation (Studies 4–5).

- **Entrenchment:** "I've heard *come* used hundreds of times. Yet never, in all of the *un-[VERB]* constructions I've observed, have I heard someone use *come* (i.e., *uncome*). Surely if this *were* allowed, I would have heard it by now. I will now therefore tentatively assume that it is not allowed".
- **Preemption:** "For all of the reversals of coming actions I have observed, people have said *go* and never *uncome*. I will now therefore tentatively assume that the former is the (more) permissible way of describing reversals of coming events.

Five previous grammaticality judgment studies sought to compare and quantify effects of preemption and entrenchment by correlating corpus data with grammaticality judgment scores, with three age groups: 5–6 year olds, 9–10 year olds, and adults: Ambridge et al., 2012, (locatives); Ambridge et al., 2014 (datives); Ambridge, 2013; Blything et al., 2014; (verbal *un-*prefixation); Ambridge et al., 2015 (a cross section of eight different constructions). These studies yielded an inconsistent pattern, with preemption found to be the more important predictor for some age groups and sentence types, and entrenchment for others.[5] The main goal of the present article is to therefore provide a clearer formulation of preemption and entrenchment, and to investigate – by means of a series of reanalyses – whether the two mechanisms can be dissociated on the basis of these previous studies. We then present a new study which replicates the adult study of Ambridge (2013) with a larger dataset, in terms of both items and participants.

The structure of the paper is as follows. We begin by presenting an overview of the five previous studies (§1), and then discuss in more detail concerns which have been raised regarding the operationalization of preemption and entrenchment (§1.1), use of raw versus difference scores as the dependent measure (§1.2) and aspects of the statistical analyses (§1.3). We then present our reanalyses which address these problems for each study in turn (§2, §3,§4, §5) before presenting our new extended replication study (§6). In general, we find evidence for both preemption and entrenchment for almost every age group in every study, particularly when each is examined in isolation (i.e., in a single-predictor model). In many cases, the two predictors are too highly correlated to be dissociated. However, for the studies of various constructions (§4), *un-* prefixation (§5) and its extended replication (§6), both preemption and entrenchment explain unique variance. We conclude by arguing that effects of preemption and entrenchment, as well as semantics, are ultimately derived from a single unitary learning mechanism (e.g., Ambridge & Blything, 2016; Goldberg, in press; Barak, Goldberg & Stevenson, 2016; Ramscar, Dye & McCauley, 2013).

Overview of previous corpus-judgment studies

All five studies revisited in the current paper used a grammaticality judgment paradigm in which participants rate sentences (or, for the *un-*prefixation studies, word forms) for acceptability, in most cases using a 5-point smiley-face scale suitable for young children (see Ambridge, Pine, Rowland, & Young, 2008); in some cases, adults used a 7-point numerical scale. We note in passing that the grammaticality judgment paradigm does not provide a transparent window into the linguistic system, since there is evidence that participants' introspective judgments are unreliable, particularly for items from skewed distributions (e.g., Parducci, 1965), such as the distribution of verbs in the corpora used to derive predictors in the studies under discussion. These intuitions are presumably even more unreliable in the case of young children. Possible alternative paradigms, however, have other problems. Production paradigms (with or without priming) force participants to produce one or other form per trial, and so are ill-suited to capturing gradience in acceptability (Ambridge, 2017; Harmon & Kapatsinski, 2017). For example, several studies have shown that native speaking adults consider **The funny joke laughed him* to be less acceptable than **The funny joke giggled him* (Ambridge, Pine, Rowland, & Young, 2008; Ambridge, Pine, Rowland, Jones, & Clarke, 2009), though (presumably) none would utter either in a production task. In any case, there exist no production studies that have included anything like the range of verbs and constructions investigated in the studies reanalysed here.

For three of the five previous studies revisited here (the exceptions are the un- prefixation studies of Ambridge, 2013, and Blything et al., 2014), the dependent measure was the degree of (dis)preference for one construction over another ("difference score"), determined by subtracting, for each verb, the acceptability ratings of one construction (e.g., ground-locative or double-object[DO]-dative) from the other (e.g., figure-locative or prepositional-[PO]-dative). For the *un-*prefixation studies, raw ratings of *un-* forms were used (with ratings for the corresponding 'bare' form included as a control predictor). In all five studies, operationalization of the two key independent measures was calculated as follows: (a) *entrenchment*: the overall frequency of the verb in the relevant corpus and (b) *preemption*: the frequency of the verb in a specific competing construction. Crucially, for both entrenchment and preemption, these measures were calculated only for those verbs that were attested (in the relevant corpus) *exclusively* in one or other construction of each pair (e.g., figure- or ground-locative). For so-called "alternating" verbs (i.e., those attested in both constructions, even if only rarely) both measures were set to zero (note that even a single usage of a verb in a construction was enough to deem it *alternating*; a decision to which we return below).

The previously-reported findings regarding entrenchment and preemption (as well as a third relevant factor, verb semantics) are summarized in **Table 1**. Although there are more "YES"s for entrenchment than preemption, it is clear that no straightforward interpretation of the relative roles of these two factors

Table 1: Effects observed in five previous grammaticality judgment studies of entrenchment, preemption and verb semantics.

Study	Construction	Age	Semantics (at least one predictor)	Entrenchment	Preemption
Ambridge, Pine & Rowland (2012)	Figure-/Ground-locative;	5–6	YES	YES	NO
		9–10	YES	YES	NO
		18+	YES	YES	NO
Ambridge, Pine, Rowland, Freudenthal & Chang (2014)	PO-/DO-dative	5–6	YES	(YES)*	(YES)*
		9–10	YES	(YES)*	YES
		18+	YES	(YES)*	(YES)*
Ambridge (2013)	*un*-VERB;	5–6	YES	NO	NO
		9–10	YES	YES	YES
		18+	YES	YES	NO
Blything, Ambridge & Lieven (2014).	*un*-VERB	3–4	NO	NO	NO
		5–6	YES	NO	YES
Ben Ambridge, Bidgood, Twomey, Pine, Rowland & Freudenthal (2015)	Various constructions	5–6	NA	NO	NO
		9–10		YES	NO
		18+		YES	NO

*Significant if entered before, but not after, the other statistical predictor.

across constructions emerges. Furthermore, this table summarizes only the "core" analysis presented in each paper. If we were to include further supplementary analyses reported in the papers (e.g., verbs rated by adults only; analyses broken down by subsets of alternating and non-alternating verbs), the picture would become only more confused.

We argue that the analyses presented in these papers suffered from a number of shortcomings, specifically: (i) Inappropriate operationalization of entrenchment and preemption; (ii) Undesirable consequences of the use of difference scores; (iii) Problems relating to the statistical analyses, and (iv) Use of a small corpus (for the locatives study only). We discuss each of these problems in sections §1.1–1.4). The focus of the present paper is on entrenchment and preemption, because these are the predictors whose roles are – in previous work – most inconsistent and in need of clarification. However, an important secondary goal of the present paper is to verify that previously-observed effects of verb semantics, which seem to play a crucial role in the retreat from overgeneralization, are robust to the more rigorous analyses employed here. We cannot, for example, rule out *a priori* the possibility that previously-observed semantic effects may disappear after controlling for more precisely operationalized measures of entrenchment and preemption.

Operationalizing preemption and entrenchment

The most crucial shortcomings of the locative, dative and various-construction studies of Ambridge et al. (2012, 2014, 2015) relate to the entrenchment and preemption variables. A key problem is that all three studies (Ambridge et al., 2012, 2014, 2015) treated entrenchment and preemption as mechanisms that work only to block verbs in constructions that *never* occurred in the corpus. That is, *even* a single use of a verb in the disfavoured construction was taken as grounds to deem it an "alternating" verb, and hence to set both the entrenchment and preemption counts to zero. For example, in the corpus analysis for the datives study (Ambridge et al., 2014), heave – a prototypical example of a PO-only verb (e.g., Pinker, 1989) – appeared in the PO dative 9 times (e.g., He heaved the box to her). However, a single occurrence in the DO dative (e.g., *?He heaved her the box*) was sufficient for this verb to be deemed "alternating", and the entrenchment and preemption counts set to zero.

This is problematic for two related reasons, one theoretical, one statistical. The theoretical problem is that since both entrenchment and preemption are inherently *probabilistic* accounts, they do not predict that a *single* use in one construction will override hundreds or even thousands of competing uses. Indeed, it seems implausible that any learning mechanism could be so brittle, given that learners will presumably encounter the odd ungrammatical uses of many verbs (e.g., slips of the tongue, non-native speech etc.). In addition, there is clear evidence from studies of adult language processing that speakers are sensitive to verb-bias – the gradient degree to which even "alternating" *verbs – tend* to show a bias to one or other construction of a pair (e.g., PO- vs DO-dative; Garnsey et al., 1997; MacDonald, Pearlmutter, & Seidenberg 1994; Snedeker & Trueswell 2004; Trueswell, Trananaus & Kello 1993; Wonnacott et al., 2008). Thus, speakers must be

implicitly tracking the distributions of even "alternating" verbs; a fact was not captured by the entrenchment and preemption measures as operationalized in these previous studies.

The statistical problem with this operationalization is that the preemption and entrenchment predictors are likely to explain some variance in participants' judgment data simply by predicting higher acceptability scores across both constructions (or low difference scores) for verbs with zero than greater-than-zero scores on these measures. In other words, the preemption and entrenchment predictors were "told for free" whether a verb alternated between two constructions (indicated by a zero) or was restricted to one construction only (indicated by a positive number). For example, considering the dative study (Ambridge et al., 2014), *feed, give, pass, throw, sell* and *send* – as verbs attested in the corpus in both the PO- and DO-dative – were assigned entrenchment and preemption scores of zero. Hence the statistical model can do very well simply by predicting a difference score of zero for all of these verbs (i.e., predicting that participants will assign approximately equal ratings to the PO- and DO-sentence variants), even though they vary dramatically in their relative preference for the PO- vs DO-dative.

Partly in response to this problem, the *un*-prefixation study of Ambridge (2013: 516) and the various-construction study of Ambridge et al. (2015) looked for preemption and entrenchment effects across (*a priori*) ungrammatical forms only (a similar supplementary analysis was presented for locatives in Ambridge et al., 2012: 270). However, this solution is too harsh to the entrenchment and preemption predictors, as it robs them of their opportunity to predict which verbs may or may not grammatically appear in particular constructions – exactly what they were designed to do in the first place.

A solution to this problem is to reformulate both preemption and entrenchment as measures of contingency, rather than raw frequency (an approach already adopted, for dative overgeneralization errors, by Stefanowitsch, 2008). Indeed, in the wider domains of language learning (e.g., Ramscar, Dye & Klein, 2013; Ramscar, Sun, Hendrix and Baayen, 2017) and human and animal learning generally (e.g., Rescorla & Wagner, 1972; Allan, 1980; Gallistel, 2003), it has long been recognized that inferences are made on the basis of contingency rather than raw frequency. Consider the simple case of a rat learning the relationship between a buzzer and delivery of a food pellet. The rat's behaviour is predicted not by the raw frequency of any one stimulus, or even of any pair of stimuli, but by contingency, as can be summarized in a 2 × 2 matrix (**Table 2**). Assume that time is divided up into a number of discrete intervals – e.g., by a bell signalling the start of each trial – on which each stimulus (buzzer/food) can be either present or absent.

Table 2: Learning is based on contingency, not raw frequency.

	FOOD-YES	FOOD-NO
Buzzer – YES	8	2
Buzzer – NO	8	2

What predicts learning in this scenario? Not the raw frequency of buzzer+food pairings (8), not the proportion of buzzer trials on which food appears (8/10), but the *contingency* of the food on the buzzer. In fact, a rat in this scenario learns no relationship between buzzer and food. There is none to learn: Food appears on 80% of trials whether the buzzer sounds or not. A relationship is learned only when the proportion of food versus no food trials is greater when the buzzer sounds then when it does not (e.g., **Table 3**).

To take another example, if we want to know whether smoking causes cancer, it is not enough to know simply the number of smokers who get cancer. We also need to know the number of smokers who do not get cancer, and the number of *non*-smokers who do and do not get cancer.

Similarly, if we are interested in the co-occurrence relationship between a particular verb *fill* and a particular construction (e.g., figure locative as in **She filled the water into the tub*), we need to know the frequency with which the verb versus other verbs appear in that construction versus *the other construction(s) under consideration*. This gives us a new way to differentiate preemption and entrenchment: for preemption, the other relevant context is limited to a directly competing, near synonymous construction (e.g., the *ground locative* – **Table 4**); for entrenchment, *all* other constructions are relevant (**Table 5**).

Thus, the chi-square test quantifies the relative unlikelihood of the observed deviation between a given verb's distribution and all verbs' distribution, under the hypothesis that the given verb shares the same distributional properties of all other verbs. An important property of the chi-square test is that – unlike some other measures of contingency such as the odds-ratio – it is sensitive to the raw frequencies in each cell. For example, a verb with 100 figure-locative and 20 ground-locative

Table 3: An example of successful contingency learning.

	Food-YES	FOOD-NO
Buzzer – YES	8	2
Buzzer – NO	3	7

Table 4: Preemption as contingency.

	Figure locative	Ground locative
fill	A	B
all other verbs	C	D

Table 5: Entrenchment as contingency.

	Figure locative	All other constructions
fill	A	B
all other verbs	C	D

uses would yield a larger chi-square value (i.e., a larger *figure-locative* bias) than a verb with 10 *figure-locative* and 2-ground locative uses, even though the odds-ratio is identical for the two cases.

This is a desirable property of the chi-square measure, given the considerable evidence that, across a wide variety of domains, the language-learning mechanism is acutely sensitive to raw frequency (see Ambridge, Rowland, Theakston & Kidd, 2015 for a review). Indeed, four recent studies (two in each of Tatsumi, Ambridge & Pine, 2017, in press) have observed a relationship between a chi-square measure of input distribution and children's performance in a language-production task. That said, we do not claim that there is anything special about the chi-square statistic per se. Any measure of contingency that is sensitive to both raw frequency and base rate (i.e., the counts in the bottom two cells in **Tables 2–4**) would be suitable for our purposes, and would presumably yield very similar results (e.g., the *p* value of the Fisher-Yates exact test; Stefanowitsch & Gries, 2003; Stefanowitsch 2008; Gries 2012, 2015; Yule's, 1912, coefficient of colligation). Indeed, Hughes (in press) compared eight co-occurrence measures on their ability to predict the amplitude of Event Related Potential (ERP) responses to novel linguistic combinations, and found that the best combined both bias and some measure of raw frequency or effect size (from best to worst: *Z*-score, cubic association ratio (MI3), Dice coefficient, *T*-score, Frequency, Transition probability, Mutual Information, Log-Likelihood). For the analyses that follow, we selected the chi-square statistic as a simple and widely-understood test of contingency/independence that has this property, and that is sufficiently accurate when the expected values are large (as they are, in every case, for the present analyses).

Another desirable property of the chi-square measure (or any similar measure of contingency/independence) is that occasional errors that the learner may encounter (e.g., **She filled the water into the cup*; perhaps produced by a non-native speaker) do not need to be detected and discarded. Infrequent errors (e.g., the occasional use of *fill* in a figure-locative construction) will make only an extremely minimal contribution to the chi-square value, since they will be overwhelmed by grammatical forms: i.e., use of *fill* in ground locative constructions (preemption) or other, non-locative constructions (entrenchment).

Another problem with the previous analyses was that the preemption measure used in these studies – corpus frequency of the relevant verb in the single most nearly synonymous construction (e.g., sentences of the form *X poured Y into Z* preempt errors such as **Bart poured the cup with water*) – is a subset of the entrenchment measure – overall corpus frequency of the relevant verb, regardless of construction.[6] Thus, the two measures were not only highly correlated in practice (e.g., $r = 0.70$, $p < 0.001$ for the locatives study of Ambridge et al., 2012; $r = 0.70$, $p < 0.001$ for the datives study of Ambridge et al., 2014), but systematically related in principle. This is a problem as it makes the two predictors virtually impossible to disentangle statistically. In the current studies, we attempted to address this problem by excluding, when calculating the entrenchment predictor, all uses already counted towards preemption. For example, in **Table 5**, counts in cells B and D would *not* include utterances already counted in the equivalent cells in **Table 4**.

That is, as in Ambridge et al. (2012, 2014, 2015), the preemption measure is based on counts of near-synonymous uses only (e.g., *X poured Y [into Z]* for errors such as **Bart poured the cup [with water]*). However, the entrenchment measure is based not on *all* uses of the relevant verb (as in these previous studies), but only uses such as *It's pouring outside* which do not count towards preemption. It is important to bear in mind that this constitutes a very conservative test of entrenchment. The present entrenchment predictor does not instantiate the entrenchment hypothesis per se (which would require calculating the predictor on the basis of *all* uses). Rather, it tests a specific prediction of the entrenchment hypothesis: that attested occurrences of a particular verb will contribute to the perceived ungrammaticality of attested uses, *even when the two are not in competition for the same message*. This modification means that, in principle, the preemption and entrenchment measures used in the present study are independent. In practice, they nevertheless remain relatively highly correlated, presumably because verbs that are (in/)frequent in a given construction tend to be (in/)frequent across the board.[7] Due to this problem of collinearity (e.g., Westfall & Yarkoni, 2016) we therefore use model comparison to investigate whether one predictor adds predictive power above and beyond the other.

The use of raw versus difference scores

With the exception of the verbal *un-* prefixation studies (Ambridge, 2013; Blything et al., 2014) and the various-constructions study (Ambridge, 2015), the previous studies revisited here used as their dependent measure "difference scores", calculated, on a participant-by-participant and verb-by-verb basis, by subtracting the acceptability rating for one construction (B) from that for another construction (A), as in the following example (from Ambridge et al., 2012) where construction A is the figure-locative and construction B is the ground-locative (chosen as nearly-synonymous constructions):

(A) *Bart poured water into the cup: 5 (on the 5-point scale)*
(B) **Bart poured the cup with water: 1 (on the 5-point scale)*
Difference score = 5 − 1 = 4

Similarly, for the datives study (Ambridge et al., 2014), the PO-dative (e.g., *Homer gave a book to Bart*) and the DO-dative (e.g., *Homer gave Bart a book*) were used as constructions A and B respectively.

The advantage of difference scores is that they control for any general (dis)preferences that participants may exhibit for particular verbs (e.g., low-frequency verbs, those denoting socially undesirable actions), characters (e.g., Bart vs Lisa) etc. However, this advantage is outweighed by a number of more serious disadvantages.

First, any apparent preemption or entrenchment effect could, in principle, be a consequence of attested uses boosting the acceptability of the grammatical member of the pair. For example, if our preemption measure predicts a high difference scores for "pour", this could be due EITHER to the fact that pour is judged highly grammatical in the (attested) figure locative (e.g., *Bart poured water into the cup*) OR because it is judged highly ungrammatical in the unattested ground locative construction (e.g., **Bart poured the cup with water*). Although any statistical learning account predicts a boost in acceptability ratings for more frequent attested uses (as well as – in production – increased use of attested forms at the expense of competing formulations), the most stringent test of these hypotheses is whether attested uses independently reduce the rated acceptability of all other uses (entrenchment) or other competing uses (preemption) of that verb. Indeed, the locative study of Ambridge et al. (2012) included supplementary analyses of raw scores for exactly this reason.

Second, because our test of the preemption hypothesis assumes competition between two semantically-similar constructions, if we use as the dependent measure participants' relative preference for the two structures in question, there is a sense in which we are *assuming* the hypothesis that we are setting out to test. Thus, the use of difference scores may unfairly advantage preemption over entrenchment. Note that the opposite may be true (i.e., difference scores may unfairly advantage entrenchment over preemption) if we calculate difference scores using a non-preempting structure. For example, for verbal *un*-prefixation (the present Studies 4 and 5), the only way to calculate a difference score is to subtract ratings from (for example) **unclose* from *close*. While such a difference score is meaningful in one sense (i.e., it controls for the extent to which people like the semantic and phonological properties of the base verb, here *close*), it is not quite analogous to those calculated for the other studies, which reflect competition between verbs in two semantically-similar sentence constructions.

The use of difference scores also works against accounts based on verb semantics, including those with no explicit role for statistical-distributional properties (e.g., Pinker, 1989), which were also tested in the relevant previous studies. This is because certain semantic properties of verbs might be associated with increased acceptability of both constructions of a pair. For example, both the PO- and DO-dative (e.g., *Bart gave a present to Lisa; Bart gave Lisa a present*) are associated to some extent with transfer. Thus, the extent to which a verb denotes transfer might be expected to positively predict acceptability of this verb in both constructions – a possibility that would be impossible to detect using difference scores. Indeed, if the size of the effect were equal for *both* constructions, the effects would cancel each other out entirely, giving a mean estimate of zero for this semantic predictor.

Finally, using a single "difference score" to explore a pair of constructions does not allow us to detect differences in any baseline preferences for one construction over another. For example, participants might view one construction of the pair as more "open" than the other (e.g., due to type frequency, semantic generality); any such effect would be obscured in the previous analyses.

For these reasons, then, every study in the present article uses – for the main analysis – raw acceptability-judgment ratings as the dependent measure. However, because this is a contentious issue, and because – as we acknowledge above – they do have important advantages, we also report in each case a supplementary analysis using difference scores. For the raw-score analyses, noise due to semantic factors is – to some extent – controlled by the inclusion of semantic predictors in all statistical models. Any readers who (like a reviewer of this paper) remain concerned that this control may be insufficient are invited to disregard the raw-scores analyses and draw conclusions only on the basis of the difference-scores analyses.

Problems relating to the statistical analyses

In addition to the conceptual problems discussed above, each of these previous studies – while using modern statistical analyses (mixed-effects models using the lme4 package in R) – failed to conform to the current state of the art in at least one respect.

First, all except two (Blything et al., 2014; Ambridge et al., 2015) attempted to address the problem of collinearity by residualizing predictor variables against one another. Wurm & Fisicaro, 2014 point out that this is not appropriate, specifically discussing Ambridge et al. (2012) as a case study of an analysis in which residualization did not have the desired effect of allowing for assessment of the individual contributions of the predictors:

> Residualizing has no effect on the result for the residualized variable. The positive Betas [for preemption and entrenchment in a two-variable model] are what would have been observed in two-variable models even without residualization (p. 41)

The latter part of this quotation refers to the fact that Ambridge et al. (2012: 271) incorrectly interpreted a sign-change (preemption was a negative predictor in single-variable model, but a positive predictor in a two-variable model with entrenchment) as "a statistical quirk arising from the residualization process". In fact, as Wurm and Fisicaro (2014: 41) note, "what caused the changes in sign is not residualization, but moving from one-variable to two-variable statistical models".[8] This example illustrates a wider problem: Because these previous studies did not systematically report single-predictor models, we cannot know whether the failure of particular effects to reach significance (or, indeed, to run in the expected direction) was a consequence of the inclusion of other effects in the model, or whether the predictor simply did not correlate with judgments at all (or did so in the wrong direction).

Wurm and Fisicaro's (2014: 42) general recommendation of "simultaneous analysis with the original [i.e., non-residualized] predictors" is not feasible for the present analyses, due to collinearity between the predictor variables, which renders the estimates for individual predictors unreliable (as in the sign-change case discussed

above). Instead, we therefore (a) report a single-predictor nonpartial model for every predictor of interest (in each case a Bayesian mixed-effects model) and (b) investigate the unique contribution of each predictor using model comparison, specifically, the method recommended by Barr et al., 2013 (using frequentist mixed-effects models implemented in lme4). (Our perhaps-unusual decision to combine Bayesian and frequentist methods is discussed further below). All predictors are scaled into standard deviation units (Z scores) and centred; not because this reduces essential collinearity between them (as Wurm & Fisicaro, 2014, point out, it does not), but simply to allow for the use of the same prior for each for the Bayesian single-predictor models.

Another statistical shortcoming of the previous studies is that they did not take a consistent approach to either model building or significance testing. In some cases, simultaneous regression models were used; in others, predictors were entered in a theoretically-determined order, either individually or in batches. In some cases, *p* values for individual predictors were taken directly from the model summary table (i.e., calculated on the basis of the observed *t* value); in others, they were calculated using Markov Chain Monte Carlo sampling (yielding what are technically P_{MCMC} values, rather than *p* values per se; a distinction that we explain below).

In the present series of reanalyses, we use a consistent approach. First, for each single-predictor nonpartial analysis, we report a Bayesian mixed-effects model. The advantage of a Bayesian approach is that it generates P_{MCMC} values and credible intervals that – unlike frequentist *p* values and confidence intervals – each yield an intuitive interpretation, and eschew arbitrary cut-offs (e.g., $p < 0.05$). Bayesian mixed-effects models work by generating, for each fixed effect, a sample of plausible mean (Beta) values on the basis of (a) the observed values and (b) the specified distribution: here, in all cases a normal (Guassian) distribution with a mean of 0 and a standard deviation of 1. This prior was chosen to be conservative, on the basis that an increase of one SD for any single predictor (all were scaled into SD units) is likely to result in a change of considerably less than 1 point on the 5-point grammaticality judgment scale. For example, in the study of Ambridge, Pine, Rowland and Young (2008) – the first to use this scale – the mean difference between ratings for errors with high and low frequency verbs (e.g., **The man fell/tumbled the boy into the hole*) was between 0.6 and 0.8 points on the 5-point scale for all three age groups (5–6 years, 9–10 years and adults; the same age groups as in the studies reanalysed here). The P_{MCMC} value for a particular fixed effect is simply the proportion of samples that have values of zero or lower (or, for negative effects, zero or higher). Thus, the P_{MCMC} value yields an intuitive interpretation (one that is often incorrectly ascribed to frequentist *p* values; e.g., Cohen, 1994): the probability that the true mean value for the effect in question is zero or lower (for a positive effect). Similarly, Bayesian credible intervals have a more straightforward interpretation than their frequentist equivalent (confidence intervals): The probability that the true value of the mean lies within the credible interval is 0.95 (or whatever interval was calculated). That said, it is important to note that, with a relatively uninformative prior – such as that used here – Bayesian and frequentist analyses generally arrive at similar conclusions.

The single-predictor models described above cannot, of course, tell us whether or not one predictor (e.g., preemption) explains variance above and beyond the other(s) (e.g., entrenchment, verb semantics, control predictors). In principle, this question could be investigated by means of simultaneous regression models. For the present analyses, however, the high degree of collinearity between predictors (particularly preemption and entrenchment) would render such models essentially uninterpretable. We therefore followed the model-comparison approach recommended by Barr, Levy, Scheepers and Tily (2013), using frequentist models implement in lme4. Although, for the reasons outlined above, we would have preferred to use Bayesian models, this proved to be computationally infeasible, since the model-comparison procedure (leave-one-out validation) requires the calculation of several thousand models for each dataset.[9] Barr et al's (2013) model-comparison approach sidesteps the problem of collinearity, because it works by comparing a full model against a model with the predictor of interest removed. Thus, the predictor of interest is never evaluated 'in situ' in a model containing other predictors with which it may share collinearity.

A final statistical shortcoming is that all previous studies except Ambridge et al. (2015) reported random-intercept-only models without random slopes (though all except Ambridge, 2013, verified that the addition of random slopes for significant predictors did not significantly improve model fit). In a series of simulation studies, Barr et al. (2013) present evidence that exclusion of random slopes is anti-conservative. However, there is some debate in the literature as to how to determine which slope structure to use: Barr et al. (2013) argue for a "maximal approach" approach – i.e. including all possible random slopes that are relevant for the design, while a recent paper by Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017) suggests that that this approach may be overly conservative at the expense of power, arguing for the use of model comparison in determining slope structure, particularly for relatively small data sets. Another problem with the approach advocated by Barr et al. (2013) is that fully maximal models often fail to converge (or to converge in a reasonable time; e.g., Eager & Roy, submitted). This necessitates a by-hand simplification process that is not only overly laborious (the present analyses required 262 separate lme4 models), but results in very different model structures for similar datasets (e.g., for adults and children completing the same task), which makes comparison across these datasets problematic. As a compromise that allows for a uniform approach across all frequentist models, we therefore decided on a structure that models random effects on the intercept and all slopes, but not correlations between them (convergence-failures is not a problem for the single-predictor Bayesian models).

Problems of corpus size

A final problem is specific to the locative study of Ambridge et al. (2012): the corpus – the 1-million-word ICE-GB – used to generate distributional statistics was not sufficiently large (this corpus was chosen over larger alternatives because it is fully parsed, making it relatively easy to extract locative constructions). Consequently, many verbs were not attested in either of the locative construction(s), (although they are attested when a larger corpus is used), which puts the preemption predictor at a distinct disadvantage as compared with entrenchment. We address this problem by using counts from the 100-million word BNC. The data for the locative construction were automatically parsed using the Stanford parser (Klein & Manning, 2003). Verb-in-construction counts, for deriving the preemption measure, were obtained by hand-coding a random sample of 100 transitive verb uses and pro-rating on the basis of the overall number of verb uses in the corpus (or, for verbs with overall frequency < 100), by hand coding all uses.

The present study

In summary, the main goal of the present study is to reanalyze the acceptability judgment data from previous studies of overgeneralization errors involving locatives (Ambridge et al., 2012), datives (Ambridge et al, 2014), various constructions (Ambridge et al, 2015) and verbal *un*-prefixation (Ambridge, 2013; Blything et al, 2014), (1) using de-confounded (as far as possible) entrenchment and preemption predictor variables, (2) operationalized using the chi-square statistic as a measure of verb-construction contingency, (3) calculated on counts obtained from a large corpus (BNC or SUBTLEX). The statistical analyses use (4) Bayesian single-predictor models and frequentist model comparison (with maximal and near-maximal random structure respectively), (5) non-residualized predictor variables and (6) both raw and difference-score outcome variables.

Study 1: Locatives (Ambridge et al, 2012)

For **locatives**, an overgeneralization error occurs when a verb that is grammatical in only the *figure-locative* construction (e.g., *Bart poured water into the cup*) appears in a *ground-locative* construction (e.g., **Bart poured the cup with water*). An error also occurs when – vice versa – a verb that is grammatical in only the *ground-locative* (e.g., *Lisa filled the cup with water*) appears in a *figure-locative* (e.g., **Lisa filled water into the cup*). Note also the existence of some verbs that "alternate" between the two constructions (e.g., *Lisa sprayed water onto the flowers/Lisa sprayed the flowers with water*). Thus, for errors involving the ground-locative construction (e.g., **Bart poured the cup with water*), the most natural preempting construction is the *figure locative/caused-motion* construction, and vice-versa. Consequently, the prediction of the preemption hypothesis tested by Ambridge et al (2012) was of a negative correlation between the acceptability of errors (relative to grammatical uses, since difference scores were used) and the frequency of the relevant verb in the opposite locative construction. The prediction of the entrenchment hypothesis tested in this previous study was of a negative correlation between the acceptability of errors (relative to grammatical uses, since difference scores were used) and overall verb frequency (including uses of, for example, *pour* in neither construction; e.g., *It's raining, it's pouring*).

Finally, the prediction of the verb-semantics hypothesis tested was of a positive correlation between the relative acceptability of (a) *figure-locative versus* (b) *ground-locative* forms and the extent to which the relevant verb was judged (by independent raters) to exhibit semantic properties associated with (a) X *causing Y to GO (IN/ON)TO Z in a particular MANNER* versus (b) *X causing Z to undergo a STATE CHANGE*; the meanings of these constructions. For example, one can *pour water into a cup* (GO IN in a particular MANNER) whether or not the cup ends up full (i.e., even if there is no STATE CHANGE). Conversely, one can *fill a cup with water* (causing the cup to undergo a STATE CHANGE) regardless of the particular MANNER used (pouring, turning on a tap, dipping it in a bath etc.).

Note that, throughout this paper, when we refer to semantics or semantic predictors, we are talking about semantic properties rated at the verb level. That said, it is important to bear in mind that the relevant verb-level semantic properties (and hence those rated) are delineated by the semantics of the constructions under investigation. For example, for Study 1, the verb-level semantic properties were to do with MANNER and STATE change not – say – animacy of the second argument (a property that is relevant when the DO-dative is under investigation, as in Study 2). Like the authors of the original studies, we are agnostic as to where these verb-level semantic ratings originally come from. Participants' meanings for *pour* (which allow them to rate the extent to which this verb means "GO IN a particular MANNER") could derive from real-world cross-situational learning of events described with *pour*. Alternatively (or additionally) they could be learned by means of surface-level distributional analysis (e.g., one *pours juice* or *wine*; not *cups* or *glasses*); see Baayen, Milin and Ramscar (2016) for an analysis of the latent semantic structure in large spoken corpora, and Twomey, Chang and Ambridge (2014) for a computational model that learns the semantics of locative verbs in this way. Thus we are agnostic as to whether any semantic effects observed participants' acceptability judgments are a consequence of real-world learning of individual verb forms, or of this latent semantic structure.

Method

Participants. The judgment data reanalyzed here were provided by 48 children aged 5;10–6;8 (M = 6;3), 48 aged 9;10–10;9 (M = 9;4) and 30 adults (18–21). Fewer adults than children were required because every adult completed every test trial, with each child completing only a subset of 40.

Verb frequency counts. One important difference between the present reanalysis and the original study of Ambridge et al. (2012) is the use of a much larger corpus for deriving the entrenchment and preemption counts: the 100-million-word BNC, as opposed to the 1-million-word

ICE-GB). The present study used the core set of 59 verbs rated by both adults and children in this previous study. These were originally chosen on the basis of three criteria: (a) they were included in Levin (1993)'s list of "locative" verbs because all occurred in one or the other (or both) locative constructions, (b) they were reasonably frequent and therefore likely to be familiar to children as young as five, and (c) they were split roughly evenly between alternating (N = 20), strongly figure-biased (N = 20) and strongly ground-biased verbs (N = 19; intended to be N = 20, but *stain* was accidentally included twice in different sentences).

For the present reanalysis, the first step necessary was to obtain counts of verb frequency (a) overall (used for calculating the entrenchment measure), and (b) in the figure-locative and ground-locative constructions (used for calculating both the entrenchment and preemption measures). All uses of the relevant verbs tagged as VERB in the British National Corpus were extracted. These uses were then tagged using the Stanford parser (Klein & Manning, 2003), and 100 transitive uses[10] (or, if there were less than 100 transitive uses in the corpus, all transitive uses) extracted and hand-coded for construction type: (a) figure-locative, (b) ground-locative or (c) other. As in the original study, not all arguments needed to be overtly realized, as long as the meaning was clear in context. For example, sentences such as *He poured the water* and *He filled the cup* were coded as instances of figure-locative and ground-locative respectively. Hand coding was performed by two independent coders, with an agreement rate of 0.76 using Cohen's kappa measure. Disagreements were resolved using a third coder. These counts were pro-rated to give an estimate of the total number of occurrences in each construction for each verb.

Preemption predictor. The verb-in-construction counts outlined above were used to calculate the preemption predictor: a measure of the relative association of each verb with the two mutually-preempting constructions (see **Table 6**). For example, the verb *pour* appears 3310 times in the corpus. After pro-rating, as described above, 1031 uses were classified as *figure-locative*, 6 as *ground-locative* and the remaining 2273 as *other*. The chi-square statistic is then calculated to determine whether the observed distribution between figure and ground locatives (3310 vs 6) is different from the expected distribution calculated on the basis of all verbs in the set. This essentially gives us a measure of verb-bias towards/away from one of two constructions when compared with other verbs in the data set. On the basis of all 69 verbs, this set shows roughly a 25%–75% split in favour of the ground-locative construction. We can be reasonably confident in this split, given that an independent analysis conducted in a different way (Twomey, Chang and Ambridge, 2014) reported a very similar proportion.

The Pearson chi-squared statistic (without Yates' correction) is calculated according to the standard formula below:

$$\frac{(A*D-B*C)^2 * (A+B+C+D)}{(A+C)*(B+D)*(A+B)*(C+D)}$$

$$\frac{(1031*14202-6*4742)^2 * (1031+6+4742+14202)}{(1031+4742)*(6+14202)*(1031+6)*(4742+14202)}$$

The very large chi-square value for *pour* (2648.25) reflects that fact that the proportion of *pour* locatives that are figure locatives (>99%) is significantly greater than the proportion of other-verb locatives that are figure locatives (roughly 25%). Because the resulting chi-square values – like the frequency counts from which they are derived – have a long-tailed distribution, they were natural log (LN(0.001 + N)) transformed (a small constant was added due to the presence of very occasional zeros; in the present dataset, just one [for *splotch*]). For example, for *pour* the chi-square value of 2648.25 becomes 7.88. Because the chi-square test is non directional, we set the sign to positive if the ratio of the target to the preempting construction (for the sentence being rated) was greater for that verb than for all other verbs in the corpus, and negative if it was smaller. For example, the preemption predictor for *pour* is 7.88 for trials in which it is rated in a figure-locative construction (reflecting a strong bias towards this construction and away from the ground-locative), and –7.88 for trials in which it is rated in a ground-locative construction (reflecting a strong bias away from this construction and towards the figure locative). The use of polarity (+/–) to indicate whether a verb is attracted or repelled by a particular construction is standard in this type of analysis (see, e.g., Gries, 2015: 525). Since the supplementary difference-score analysis (arbitrarily) calculated difference scores as figure-minus-ground locative, we used the chi-square values calculated for figure locatives, such that positive values indicate bias towards the figure-locative, away from the ground-locative, and vice-versa for negative values.

Entrenchment predictor. The entrenchment predictor was calculated in a similar way, except that, for each verb, two different calculations were necessary: (a) entrenchment towards (+)/away from (–) the figure-locative construction (for trials in which the figure-locative construction was being rated) and (b) entrenchment towards (+)/away from (–) the ground-locative construction (for trials in which the ground-locative construction was being rated). Again,

Table 6: Example of the calculation of the preemption measure for the verb *pour*. *pour*: preemption.

	figure-locative **construction (*into/onto*)**	***ground-locative*** **construction (*with*)**
Pour	(A) 1031	(B) 6
all other verbs (summed)	(C) 4742	(D) 14202

the direction of the sign (+/–) was determined on the basis of whether the target verb was biased towards or against the target construction, relative to all other verbs in the corpus.

Table 7 illustrates the calculation of the entrenchment predictor for *pour*. As discussed in Section 1.1, for the present analysis, the entrenchment predictor includes only sentence types that are *not* counted under the preemption predictor (recall that this constitutes a conservative test of the entrenchment hypothesis). For example, the sentences counted as (ii) entrenching *pour* away from the ground-locative construction are the 2273 "Other" (i.e., non-locative) uses of *pour* (e.g., *it's pouring outside*), but – unlike in previous studies – NOT the 1031 figure-locative uses of *pour* that have already been allocated to preemption.

(i) = 1585.14 ln(i+1) = 7.36 (positive value used because 1031/2273 > 4742/46587)
(ii) = 671.51. ln(ii+1) = 6.51 (negative value used because 6/2273 < 14202/46587)

By way of comparison, the entrenchment predictor for *fill* (a ground-biased verb) was calculated as –6.75 for the figure-locative construction and 6.99 for the ground locative construction.

Recall that, for reasons discussed in Section 1.2, in addition to a main analysis conducted on participants' raw ratings, we additionally present an analysis with difference scores (rating for figure- minus ground-locative) as the dependent measure. The semantic and preemption predictors used in the raw and difference-score analyses are identical. However, because the raw entrenchment predictor reflects entrenchment towards/away from only *one* construction of the pair, it was necessary to calculate a new entrenchment predictor specifically for the difference-score analyses. This predictor was calculated by subtracting the chi-square ground predictor (e.g., **Table 7[ii]**) from the chi-square figure predictor (e.g., **Table 7[i]**). In some respects, this composite predictor is too generous to the entrenchment hypothesis, as it is giving it – albeit in a roundabout way – counts of the verb in the two rival constructions; information usually considered the sole preserve of the preemption hypothesis. However, the alternative – having entrenchment against (say) the figure locative predict participants' relative preference for the figure *over ground* locative – seemed to instantiate entrenchment even less satisfactorily (and more harshly). We therefore used (for the supplementary difference-score analysis only) a composite difference-score entrenchment predictor.

Note that it is not necessary to calculate, for the difference-scores analyses, a new preemption predictor, which, by definition, already represents a verb's degree of bias both towards one of the two defined target constructions *and* away from the other. Neither is it necessary to calculate new semantic predictors. Indeed, it is unclear how this could be done meaningfully given that each semantic factor is (by hypothesis) positively associated with one construction of the pair and negatively associated with the other (see the following paragraph). For example, two predicted correlations here are between (a) higher Manner scores and a preference for figure- over ground- locatives and (b) higher End State scores and a preference for ground- over figure- locatives.

To sum up, all predictors other than entrenchment were the same in the main analysis (which uses raw sentence ratings as the DV) and the difference-scores analysis (which uses difference score sentence ratings as the DV). A raw entrenchment predictor was used in the main analysis; a difference-score entrenchment predictor was used in the difference-score analysis. Unfortunately, despite the steps taken to de-confound the preemption and entrenchment predictors, they remained very highly correlated for analyses of figure-locative sentences (r = 0.81), ground locative sentences (r = 0.76) and difference scores (r = 0.79). Indeed, due to commonalities

Table 7: Example of the calculation of the entrenchment measure for the verb *pour*.
[i] entrenchment of *pour* towards (+) away from (–) the *figure-locative* construction.

	***figure-locative* construction**	**All uses of *pour*/*all other verbs* *except* ground* or figure locatives**
Pour	(A) 1031	(B) 2273
all other verbs (summed)	(C) 4742	(D) 46587

*ground locatives are not counted here because they were allocated to the preemption predictor; figure locatives are not counted because they are allocated to (A) and (C).

[ii] entrenchment of *pour* towards (+) away from (–) the *ground-locative* construction.

	***ground-locative* construction**	**All uses of *pour*/*all other verbs* *except* ground or figure* locatives**
Pour	(A) 6	(B) 2273
all other verbs (summed)	(C) 14202	(D) 46587

*figure locatives are not counted here because they were allocated to the preemption predictor; ground locatives are not counted because they are allocated to (A) and (C).

in the way these predictors are calculated (i.e., both take into account the frequency of the verb in the construction in which it is being rated), the correlation between these predictors is even higher than in the original study (r = 0.7). Although this of course reduces the likelihood of observing dissociable effects of preemption and entrenchment, the model-comparison procedure used to investigate this possibility is not affected by collinearity (since the predictor under investigation is removed from the model altogether). Thus any observed effect of preemption above and beyond entrenchment (or vice versa) is not called into question by the existence of collinearity between these predictors.

Semantic (and morphophonological) predictors. Ambridge et al (2012) used seven composite semantic predictors, each denoting the extent to which each verb was judged – by a separate group of adult raters– to exhibit a particular cluster of semantic properties (determined by Principle Components Analysis). Two of these composite predictors (Manner, End State) related to Pinker's (1989) broad-range semantic rules on the locative constructions; the remainder (Splattering, Joining, Stacking, Gluing and Smearing) to Pinker's narrow-range semantic classes. These predictors were used unchanged (other than z-transforming).

Dependent variables. Participants rated figure locative and ground locative uses of each of 59 verbs, using a 5-point scale (children) or a 7-point scale adults. We first present separate analyses for ratings of figure- and ground-locative sentences – in full simultaneous (**Figure 1**) and nonpartial models (**Figure 2**) – followed by a supplementary analysis on difference scores (figure-minus-ground locative) – again in simultaneous (**Figure 3**) and nonpartial models (**Figure 4**).

Results and Discussion

The data were analysed using R 3.3.2 (R Core Team, 2016). Maximal single-predictor Bayesian mixed effects models were fitted by using the glimmer and map2stan functions of the rethinking package (McElreath, 2016), to pass reformatted data and lme4 syntax (Bates, Maechler, Bolker & Walker, 2015) to the rstan package (Stan Development Team, 2015a, 2015b; Hoffman & Gelman, 2014; Carpenter et al., 2016). PMCMC values and 95% credible intervals were calculated for the single predictor in each model. For the model-comparison procedure we began by building – for each age group and each sentence type (figure-locatives/ground-locatives) – a (near) maximal model (Barr et al., 2013) with random effects of Verb and Participant on the intercept, and by-participant random effects on the slopes for all predictor variables (but no correlations between random effects included), as shown below in lme4 syntax (note the || used to exclude correlations between random effects).

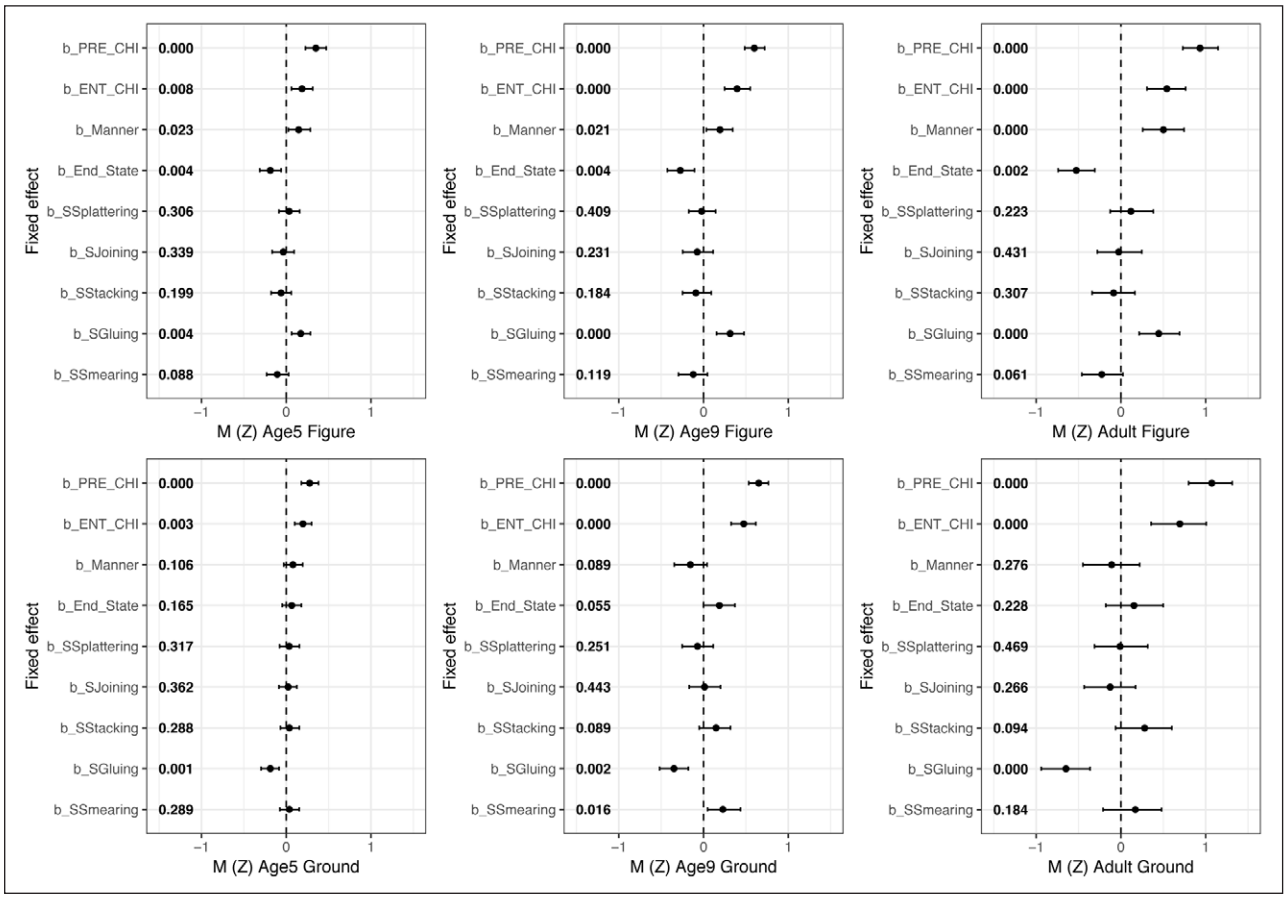

Figure 1: Study 1: Locatives, nonpartial analysis. Fixed effects (each from a separate regression model) for participants' judgments of (top) figure-locatives and (bottom) ground-locatives, and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

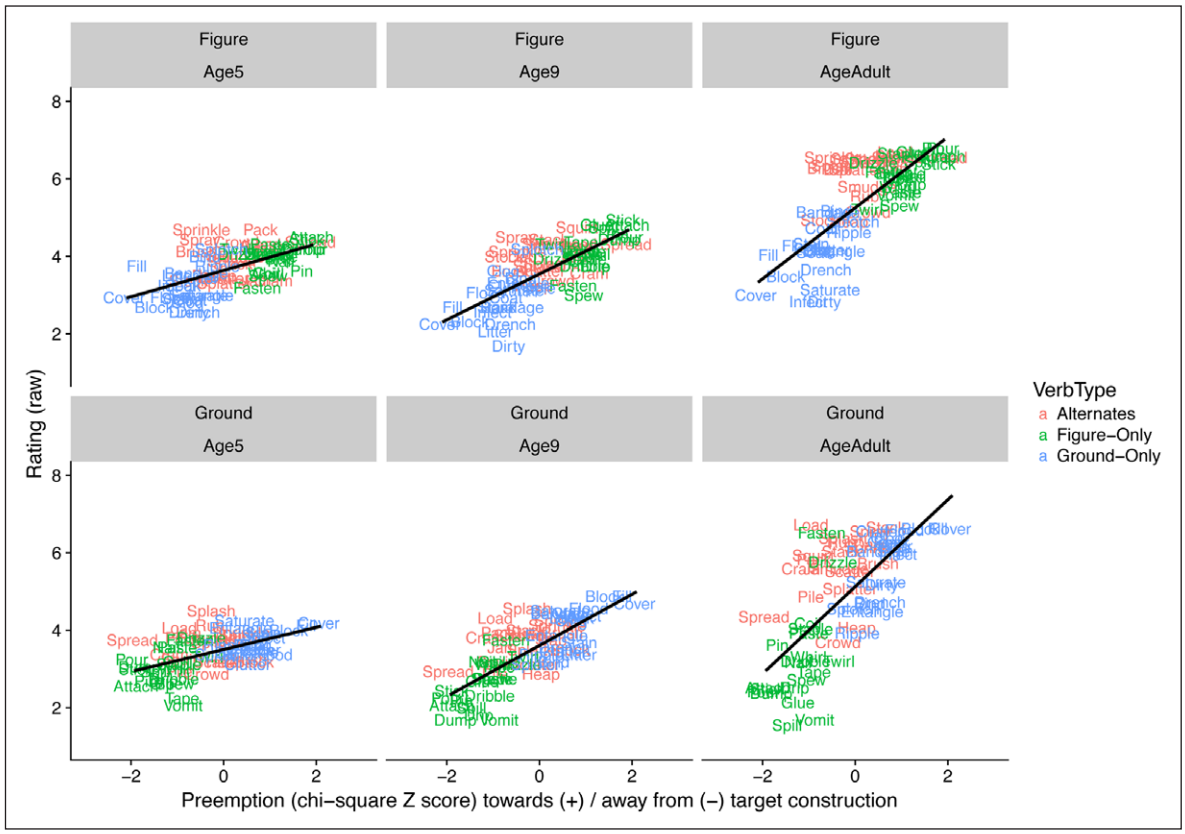

Figure 2: Study 1: Locatives. Relationship between (X axis) the preemption predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) figure-locatives and (bottom) ground-locatives on the on the 5-point (children) or 7-point scale (adults).

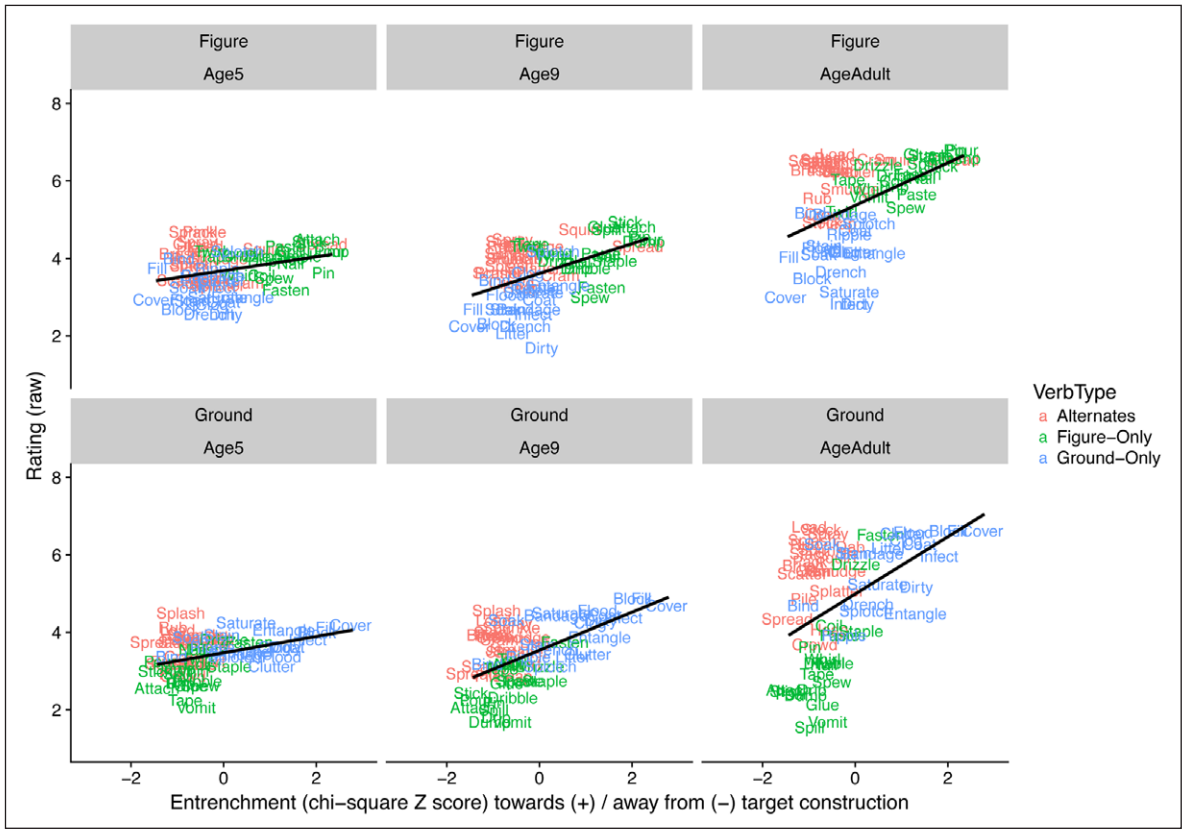

Figure 3: Study 1: Locatives. Relationship between (X axis) the entrenchment predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) figure-locatives and (bottom) ground-locatives on the on the 5-point (children) or 7-point scale (adults).

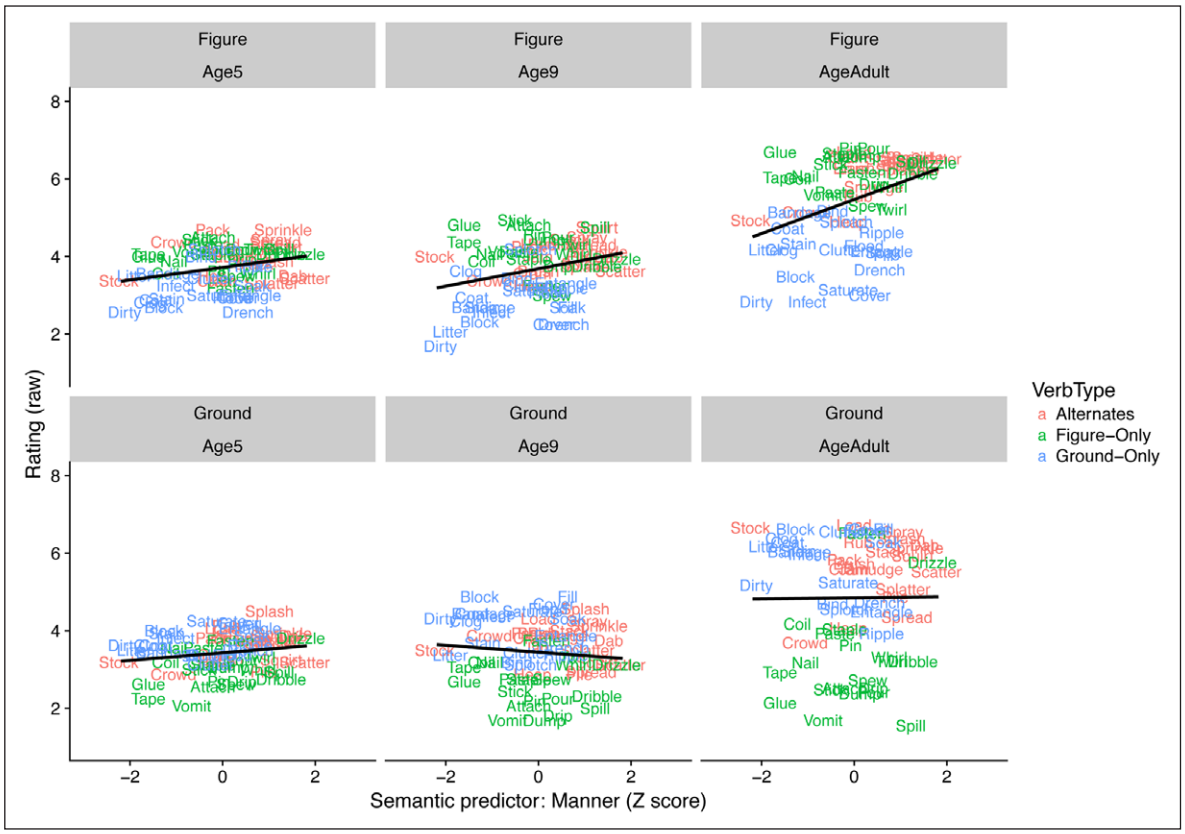

Figure 4: Study 1: Locatives. Relationship between (X axis) the semantics "Manner" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) figure-locatives and (bottom) ground-locatives on the on the 5-point (children) or 7-point scale (adults).

Rating ~ (1 + PRE_CHI + ENT_CHI + Manner + End_State + SSplattering + SJoining + SStacking + SGluing + SSmearing || Participant) + (1|Verb) + PRE_CHI + ENT_CHI + Manner + End_State + SSplattering + SJoining + SStacking + SGluing + SSmearing

We then used the *drop1* function of lme4 to remove each predictor individually (i.e., with replacement), in each case retaining all other fixed and random effects (including the by-participant random slope and intercept for the fixed effect in question), as recommended by Barr et al. (2013). This function then compares the reduced model to the full model using a likelihood-ratio-test and returns a *p* value calculated from the chi-square distribution. As noted in the Introduction, this method avoids the problem of attempting to interpret a predictor 'in situ' in a model that contains other predictors with which it shares collinearity. It is important to note that although the likelihood-ratio test is directional in the sense that it tests which model provides a better fit to the data (i.e., the model with or without the predictor of interest), it is non-directional with regard to the predictor of interest; i.e., it does not indicate whether the predictor of interest is positively or negatively related to the dependent measure. Thus we interpret the direction of each significant predictor as the direction of the nonpartial correlation between this predictor and the dependent measure (in the relevant single-predictor Bayesian model). This decision is motivated by the fact that, in cases of collinearity, the direction of the relationship between a predictor and a dependent measure can flip, in a way that does not reflect the true relationship between the two (see Footnote 8).

The correlations between predictor variables are shown in Appendix Table A1. A handful are potentially indicative of some collinearity ($r = 0.47 - 0.56$), but only the correlation between preemption and entrenchment ($r = 0.81$ and $r = 0.76$ for figure and ground-locatives respectively) gives real cause for concern.

Figure 1 shows the mean, 95% credible interval and (in bold) direction-corrected p_{MCMC} value for each single-predictor regression model. Since, for ease of interpretation, we present these data graphically, we do not report diagnostic indices such as the number of effective samples or the Gelman-Rubin convergence diagnostic (R-hat). However, we verified that for all models reported throughout this paper the latter was well below the conventional cut-off of 1.1. **Figures 2–7** plot against participants' judgments (Y axis), each predictor whose 95% CI did not include zero, for at least one construction (figure/ground locative) and one age group (Preemption, Entrenchment, Manner, End State, Gluing, Smearing). Appendix Table A2 presents the results of the model-comparison analysis. Although the four semantic predictors survive this more stringent analysis only for the adults, and only for figure locatives, preemption is shown to explain variance above and beyond entrenchment for both figure and ground locatives, for all three age groups.

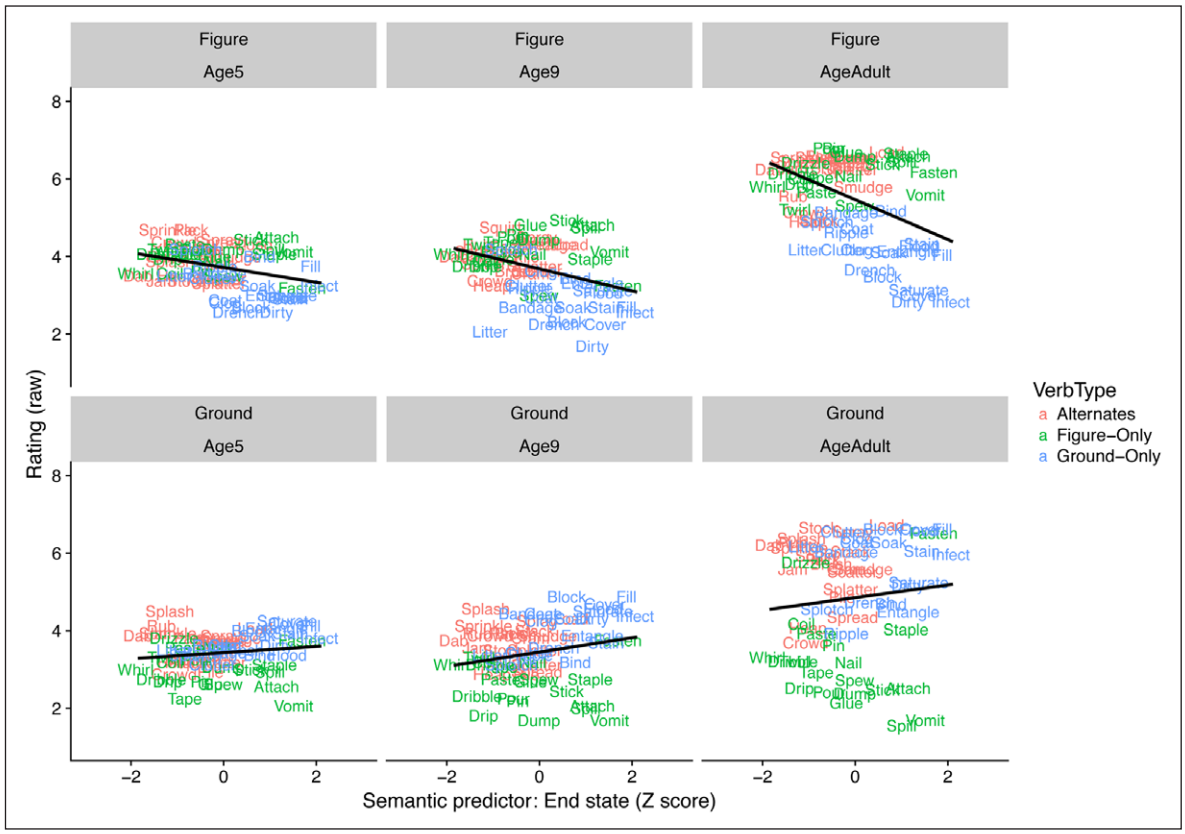

Figure 5: Study 1: Locatives. Relationship between (X axis) the semantics "End State" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) figure-locatives and (bottom) ground-locatives on the on the 5-point (children) or 7-point scale (adults).

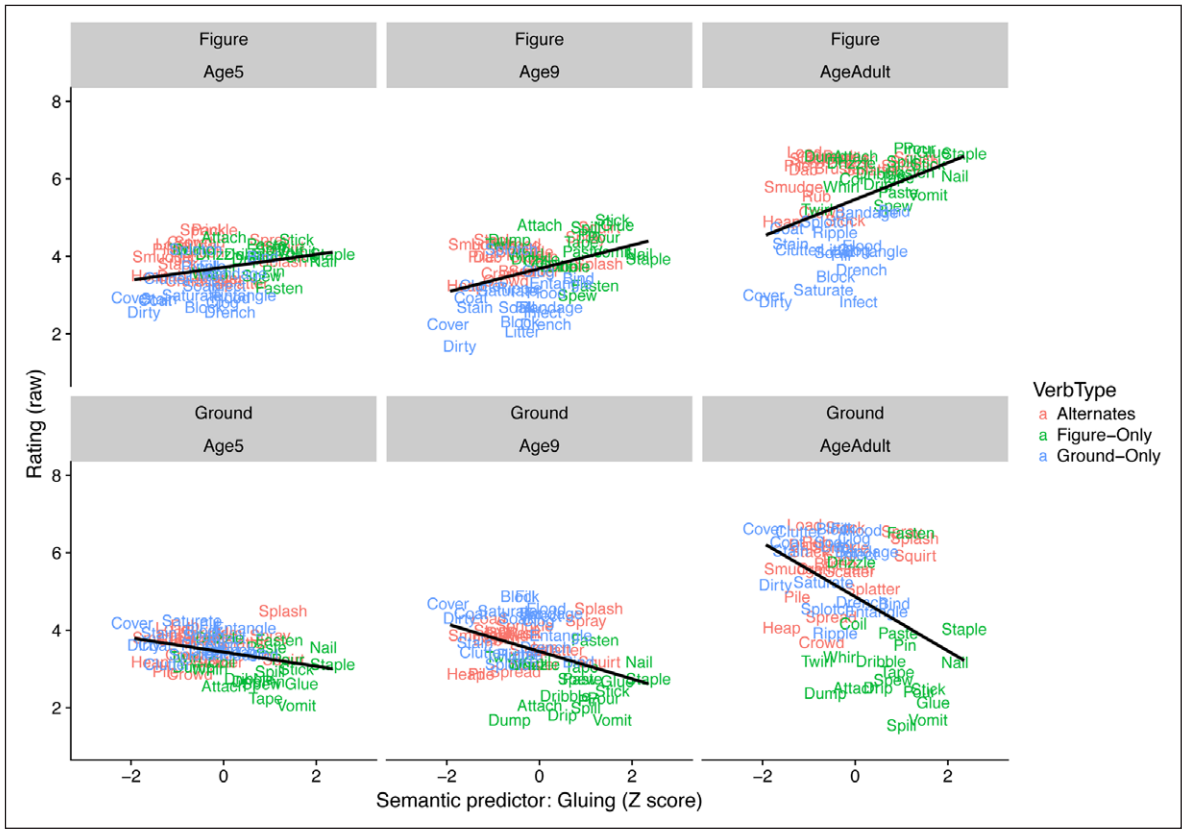

Figure 6: Study 1: Locatives. Relationship between (X axis) the semantics "Gluing" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) figure-locatives and (bottom) ground-locatives on the on the 5-point (children) or 7-point scale (adults).

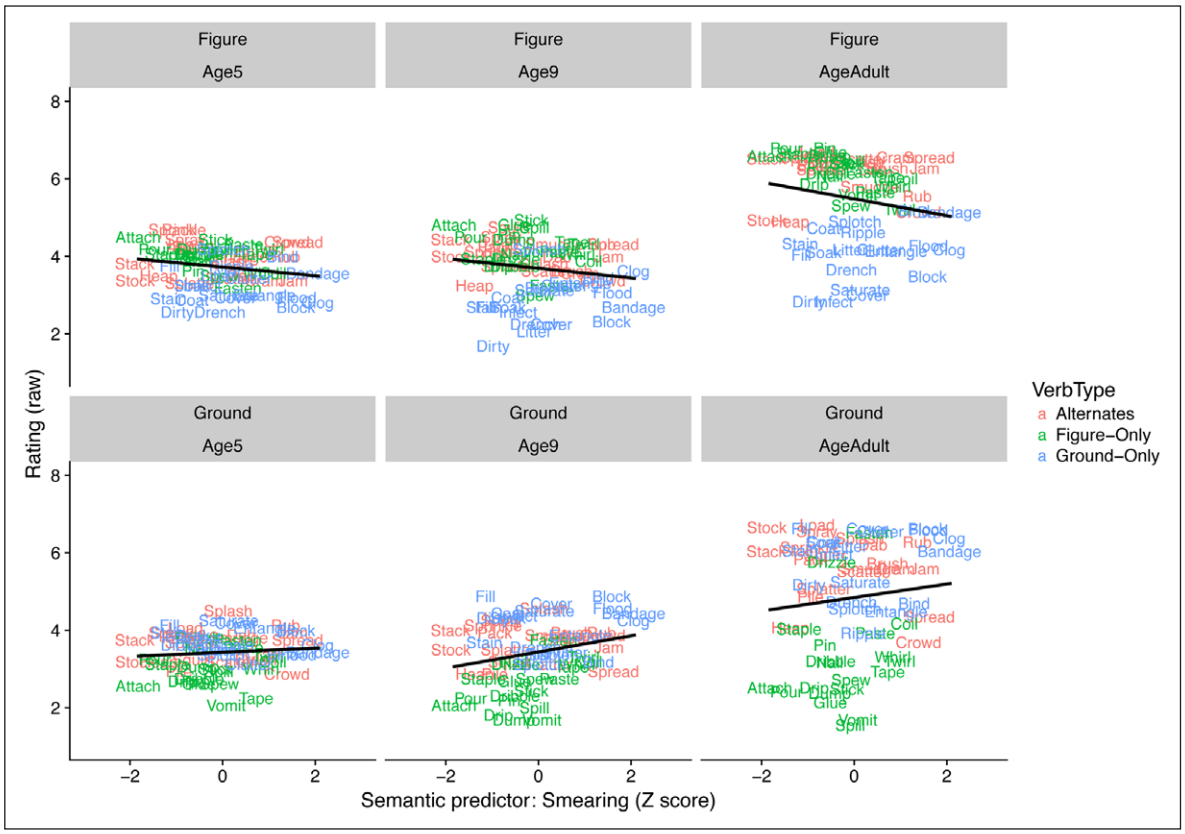

Figure 7: Study 1: Locatives. Relationship between (X axis) the semantics "Smearing" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) figure-locatives and (bottom) ground-locatives on the on the 5-point (children) or 7-point scale (adults).

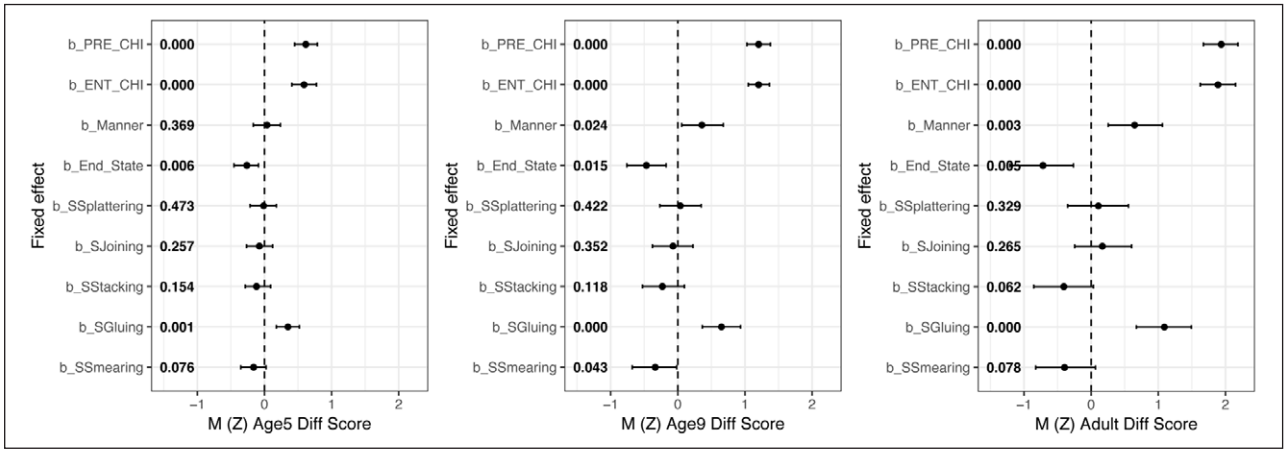

Figure 8: Study 1: Locatives, nonpartial analysis of difference scores. Fixed effects (each from a separate regression model) for participants' difference scores (figure- minus ground-locatives) and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

Furthermore, for figure locatives only, entrenchment explains variance above and beyond preemption, at least for adults and 5–6 year olds (for 9–10 year olds, $p = 0.05006$). The difference-score analyses largely replicate this pattern, with clear effects of preemption, entrenchment and four semantic predictors in the nonpartial analysis (**Figure 8**), but – apart from Gluing (for adults) – only preemption explaining variance above and beyond the other predictors for adults ($p < 0.001$), 9–10 year olds ($p = 0.03$) and – marginally – 5–6 year olds ($p = 0.055$).

To sum up, preemption displayed a clear effect above and beyond entrenchment (and all other predictors), reversing the null finding observed in the original study. An effect of entrenchment above and beyond preemption (and all other predictors) was observed for figure locatives, but not ground locatives (or difference scores). Any post-hoc explanation for this unpredicted difference could well constitute an over-interpretation of noise but – pending future replication – one possibility is that this difference is due to the fact that

overgeneralizations involving the ground-locative construction (e.g., **She poured the cup with water*) are rated as more unacceptable across the board than overgeneralizations involving the figure-locative (e.g., **She filled water into the cup*; Ambridge et al., 2012), and so may be more subject to floor effects. In conclusion, then, this reanalysis suggests independent effects of both preemption and entrenchment, though the evidence is considerably stronger in the former case. Future studies could attempt to dissociate these effects further by including verbs that have high overall frequency, but low frequency in either of the two constructions. The preemption account predicts that such uses should be relatively acceptable, provided they are semantically appropriate, while the entrenchment account predicts that they should not be (see Robenalt & Goldberg, 2015, 2016, for studies along these lines, though using different designs and methods).

Study 2: Datives (Ambridge et al, 2014)

For **datives**, an overgeneralization error occurs when a verb that is grammatical in only the *prepositional-object* (*PO*) or *ditranstive* construction (e.g., *Bart said something to Lisa*) appears in the *double-object* (*DO*) construction (e.g., **Bart said Lisa something*). For these constructions, errors in the opposite direction, though theoretically possible (e.g., **Bart cost $5 to Homer; c.f., Bart cost Homer $5*), are, at most, a marginal phenomenon. However, note again the existence of some verbs that "alternate" between the two constructions, including many of the post prototypical dative verbs, such as *give* (e.g., *Bart gave a present to Lisa/Bart gave Lisa a present*). Thus, for errors involving the *double-object* (*DO*) construction (e.g., **Bart said Lisa something*), the most natural preempting construction is the *prepositional-object* (PO) construction (e.g., *Bart said something to Lisa*, which clearly competes semantically with the error).

Consequently, the prediction of the **preemption** hypothesis tested by Ambridge et al (2014) was of a negative correlation between the acceptability of such errors (relative to grammatical uses, since difference scores were used) and the frequency of the relevant verb in the PO-dative construction. The prediction of the **entrenchment** hypothesis tested in this previous study was of a negative correlation between the acceptability of such errors (relative to grammatical uses, since difference scores were used) and overall verb frequency (including uses of, for example, *say* in neither construction; e.g., *Bart said "Hi"*). Finally, the prediction of the **verb-semantics** hypothesis tested in this previous study was of a positive correlation between the relative acceptability of (a) PO- versus (b) DO-dative forms (i.e., the difference-score measure) and the extent to which the relevant verb was judged to exhibit semantic properties associated with (a) *X causing Y TO GO TO Z* versus (b) *X causing Z to HAVE Y*; the meanings of these constructions. For example, one can send a child to bed but not **send bed a child* (DO), because the event is one of causing to GO, not causing to HAVE. Conversely, one can *give someone a headache* (DO) but not **give a headache to someone* (PO), because the event is one of causing to HAVE, not causing a headache to GO from one person to another.

Method

Participants. The judgment data reanalyzed here were provided by 36 children aged 5;2–6;1 (M = 5;7), 36 children aged 9;2–10;1 (M = 9;8), and 30 adults aged 18–21.

Preemption and Entrenchment predictors. Since the original corpus counts were obtained from the BNC, it was not necessary (unlike for Study 1) to obtain new counts. The original corpus counts were simply used to calculate new preemption and entrenchment predictors analogous to those used in Study 1. That is, the preemption predictor reflects the extent to which a verb's preference for PO- *vs.* DO-datives differs from that of the other verbs in the corpus (the counts for all other verbs in the corpus were obtained for the entire set of 301 verbs used in the larger study; not just the 44 verbs rated by adults and children). The entrenchment predictor reflects the extent to which the ratio of (a) PO-dative to non-dative uses or (b) DO-dative to non-dative uses (depending on the form being rated) differs between the verb being rated and all other verbs in the corpus. As for Study 1, we also calculated a difference-score entrenchment predictor (entrenchment-vs-PO minus entrenchment-vs-DO) for use in the supplementary difference-score analysis.

Unfortunately, despite the steps taken to de-confound the preemption and entrenchment predictors, they remained highly correlated for DO-dative sentences (r = 0.81) and difference scores (r = 0.41). For PO-dative sentences, these predictors were moderately correlated, but in the opposite direction to that predicted (r = –0.24). Given that, for PO-dative sentences, preemption is based on DO-dative frequency while entrenchment is based on nondative frequency, the negative correlation between these two predictors suggests a negative correlation between DO-dative and nondative frequency. This makes sense on the assumption that verbs that are highly frequent in the DO-dative construction, such as *give* (e.g., *He gave her a book*) tend not to appear in nondative sentences. Whether or not this explanation is correct, the high degree of collinearity between the preemption and entrenchment predictors for the DO-dative and difference-score analyses is again problematic for an analysis designed to differentiate the two predictors.

Semantic predictors. Ambridge et al. (2014) used seven composite semantic predictors, each denoting the extent to which each verb was judged – by adult raters who did not complete the main grammaticality judgment task – to exhibit a particular cluster of semantic properties (determined by Principle Components Analysis). Three of these composite predictors related to Pinker's (1989) broad-range semantic rules on the dative constructions (one to the PO-dative, two to the DO-dative); the remainder (Speech, Mailing, Bequeathing, Motion) to Pinker's narrow-range semantic classes. A final predictor, derived from the same PCA, related to a Morphophonological constraint: at least some Latinate verbs seem to be dispreferred in the DO-dative (e.g., *That gives/*suggests me an idea*).

Dependent variable. Participants rated *PO-dative* and *DO-dative* uses of each of 44 verbs, using a 5-point scale (children) or a 7-point scale (adults). The verbs were chosen to be split evenly between alternating verbs ($N = 22$) and PO-only verbs that are ungrammatical in the DO-dative (e.g., *Bart said 'hi' to Lisa; *Bart said Lisa 'hi'*). While all analyses presented in Ambridge et al (2014) used difference scores, we again present a main analysis based on raw ratings, and a supplementary analysis based on difference scores.

Results

In the same way as for the previous reanalysis, we built (a) a series of maximal single-predictor Bayesian models and – for the purposes of model comparison – (b) a series of near-maximal frequentist models (correlation between random effects was not included in the models), by removing each predictor in turn from the model specified below (in lme4 syntax).

Rating ~ (1 + PRE_CHI_LOG + ENT_vs_DO_LOG + BROAD_PO + BROAD_DO_1 + BROAD_DO_2 + SPEECH + MAILING + BEQUEATHING + MOTION + LATINATE || Participant) + (1|Verb) + PRE_CHI_LOG + ENT_vs_DO_LOG + BROAD_PO + BROAD_DO_1 + BROAD_DO_2 + SPEECH + MAILING + BEQUEATHING + MOTION + LATINATE)

The correlations between predictor variables are shown in Appendix Table A1. Again, a handful are potentially indicative of some collinearity ($r = 0.42 – 0.62$), but – given the main goal of the present reanalysis – it is only the correlation between preemption and entrenchment for DO-datives ($r = 0.81$), and, to a lesser extent, for the difference scores ($r = 0.42$), that gives real cause for concern.

Figure 9 shows the mean, 95% credible interval and (in bold) direction-corrected p_{MCMC} value for each single-predictor regression model. **Figures 10–17** plot against participants' judgments (Y axis), each predictor whose 95% CI did not include zero, for at least one construction (DO-/PO-dative) and one age group (Preemption, Entrenchment, Broad PO, Broad DO1, Speech, Mailing, Motion, Latinate). Appendix Table A2 presents the results of the model-comparison analysis. Although the semantic predictor of Broad DO 1 (5/6 datasets) and the morphosyntactic Latinate predictor (3/6 datasets) survive this more stringent analysis, entrenchment and preemption essentially cancel one another out: Although both effects are observed for DO Datives in the single-predictor models, only in one case – entrenchment for the 5–6 year olds – does either explain variance above and beyond the other predictors. For PO Datives, neither predictor shows evidence of an effect, even in the nonpartial analysis.

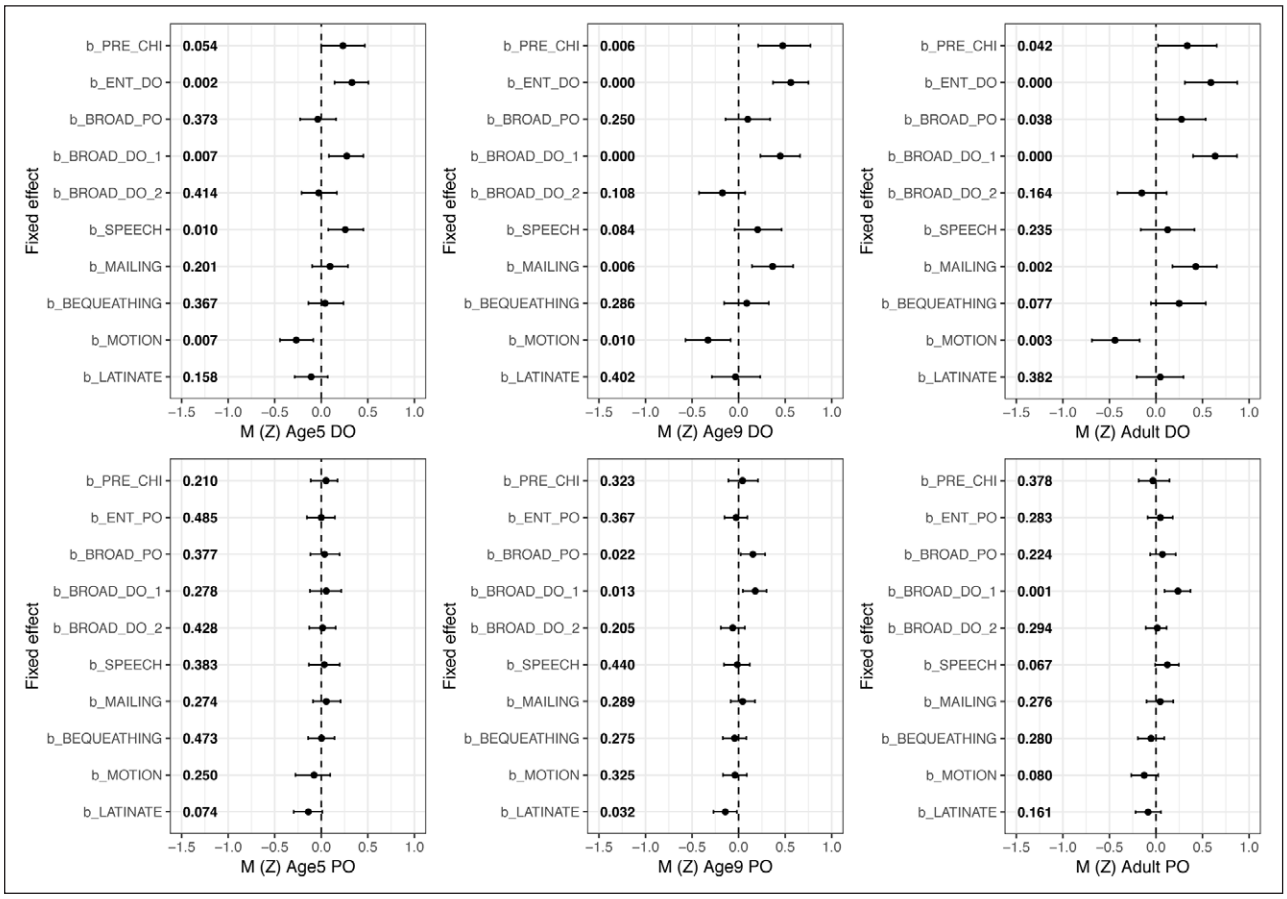

Figure 9: Study 2: Datives, nonpartial analysis. Fixed effects (each from a separate regression model) for participants' judgments of (top) DO-datives and (bottom) PO-datives, and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

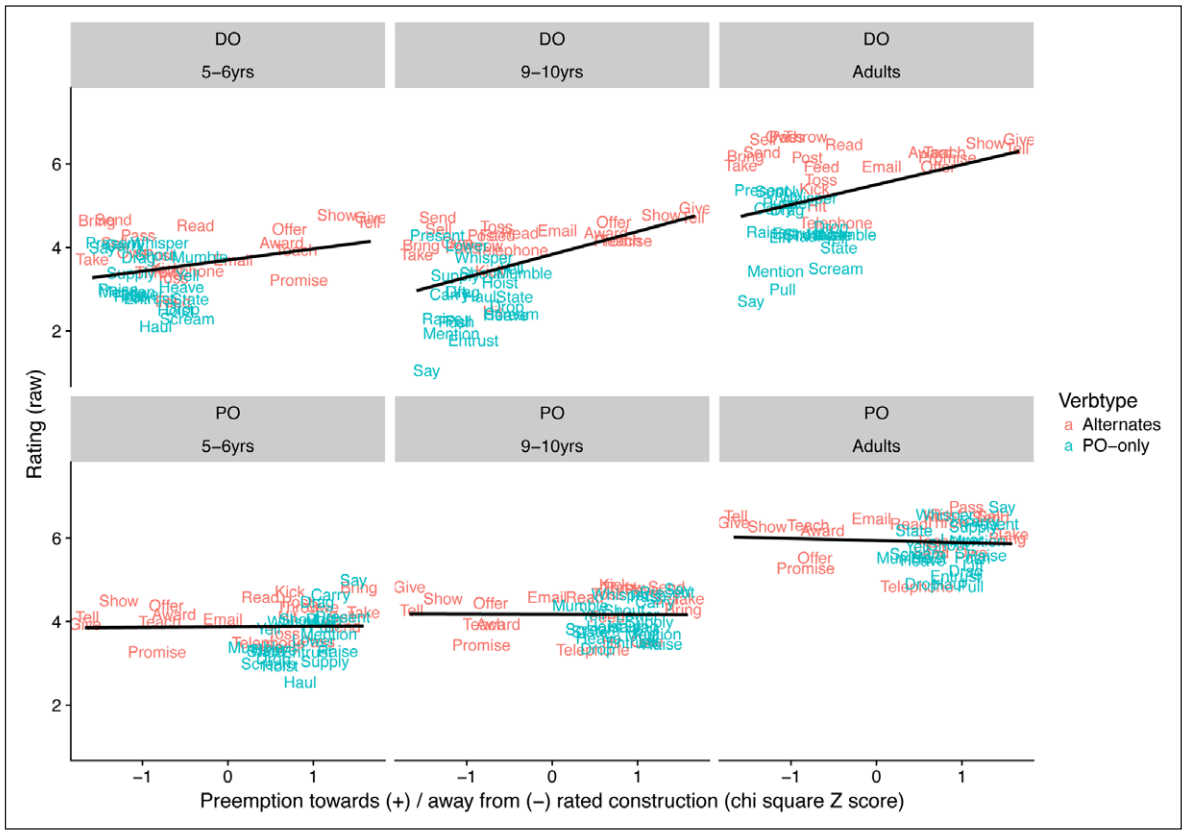

Figure 10: Study 2: Datives. Relationship between (X axis) the preemption predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

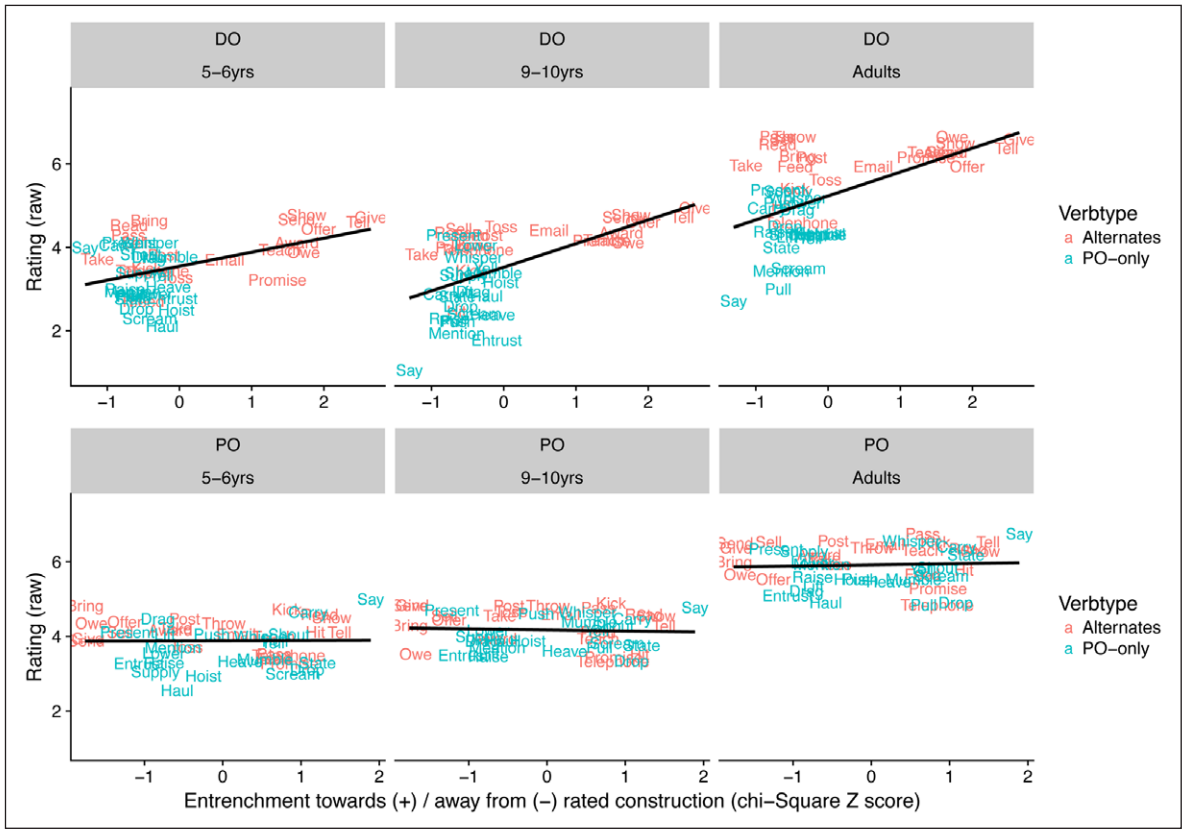

Figure 11: Study 2: Datives. Relationship between (X axis) the entrenchment predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

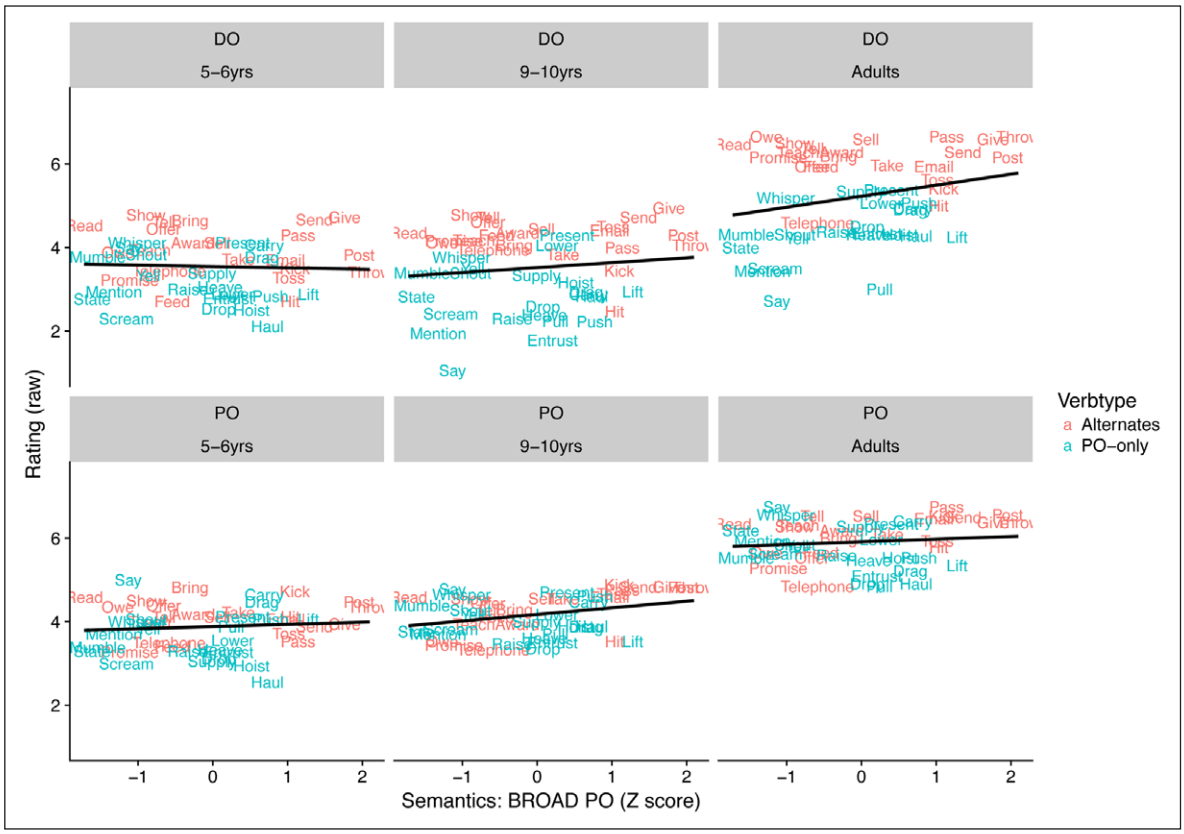

Figure 12: Study 2: Datives. Relationship between (X axis) the semantics "Broad PO" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

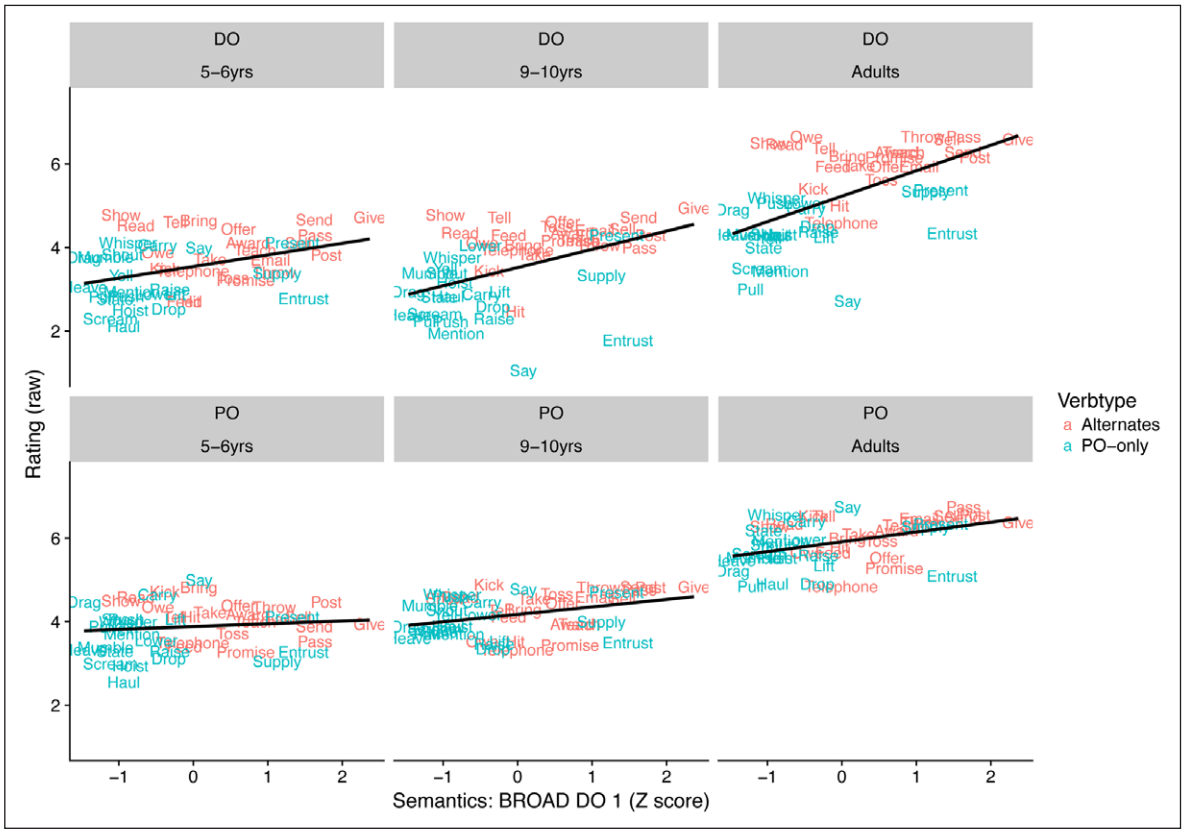

Figure 13: Study 2: Datives. Relationship between (X axis) the semantics "Broad DO1" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

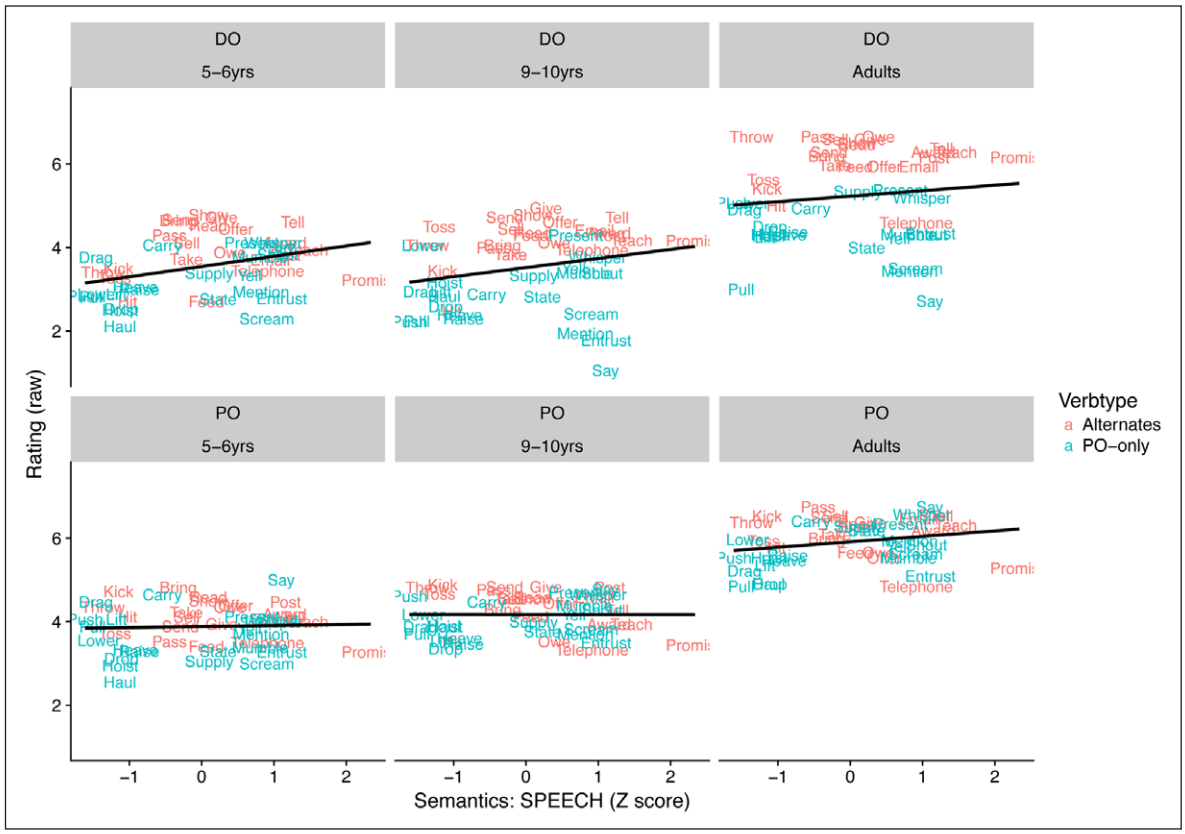

Figure 14: Study 2: Datives. Relationship between (X axis) the semantics "Speech" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

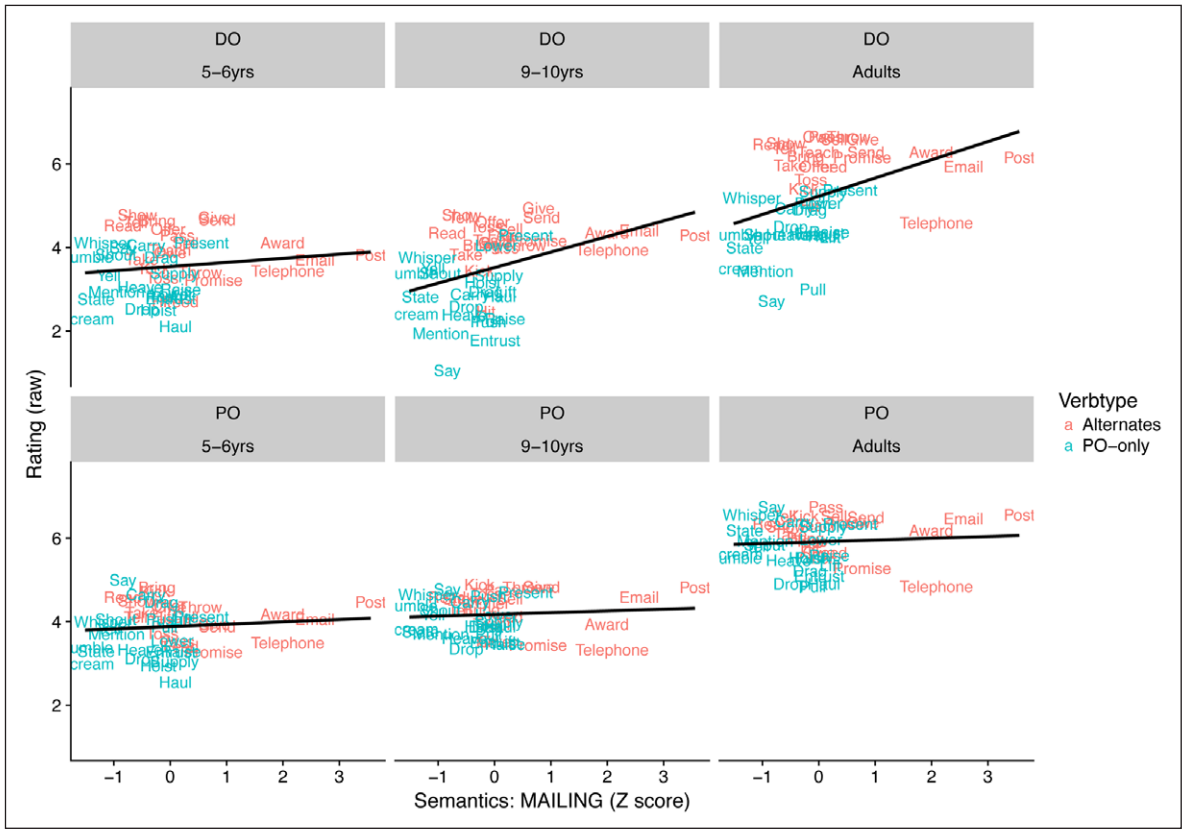

Figure 15: Study 2: Datives. Relationship between (X axis) the semantics "Mailing" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

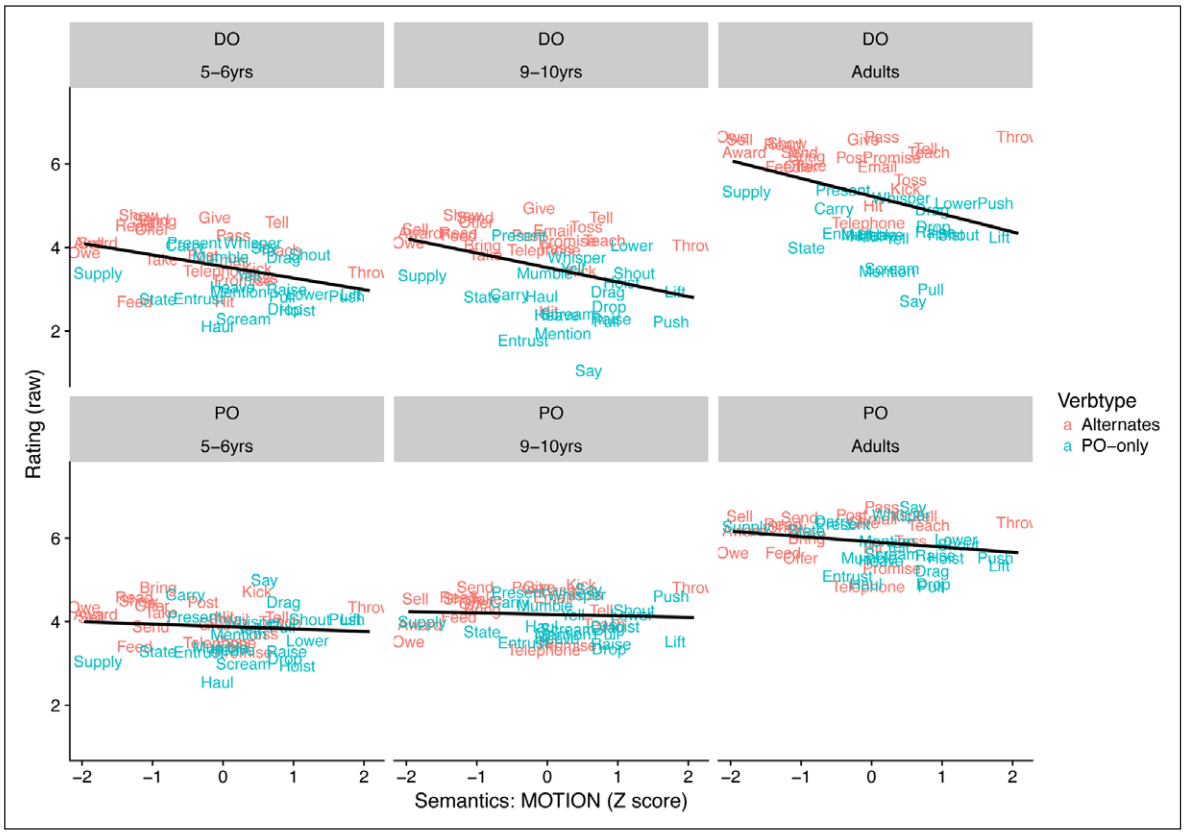

Figure 16: Study 2: Datives. Relationship between (X axis) the semantics "Motion" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

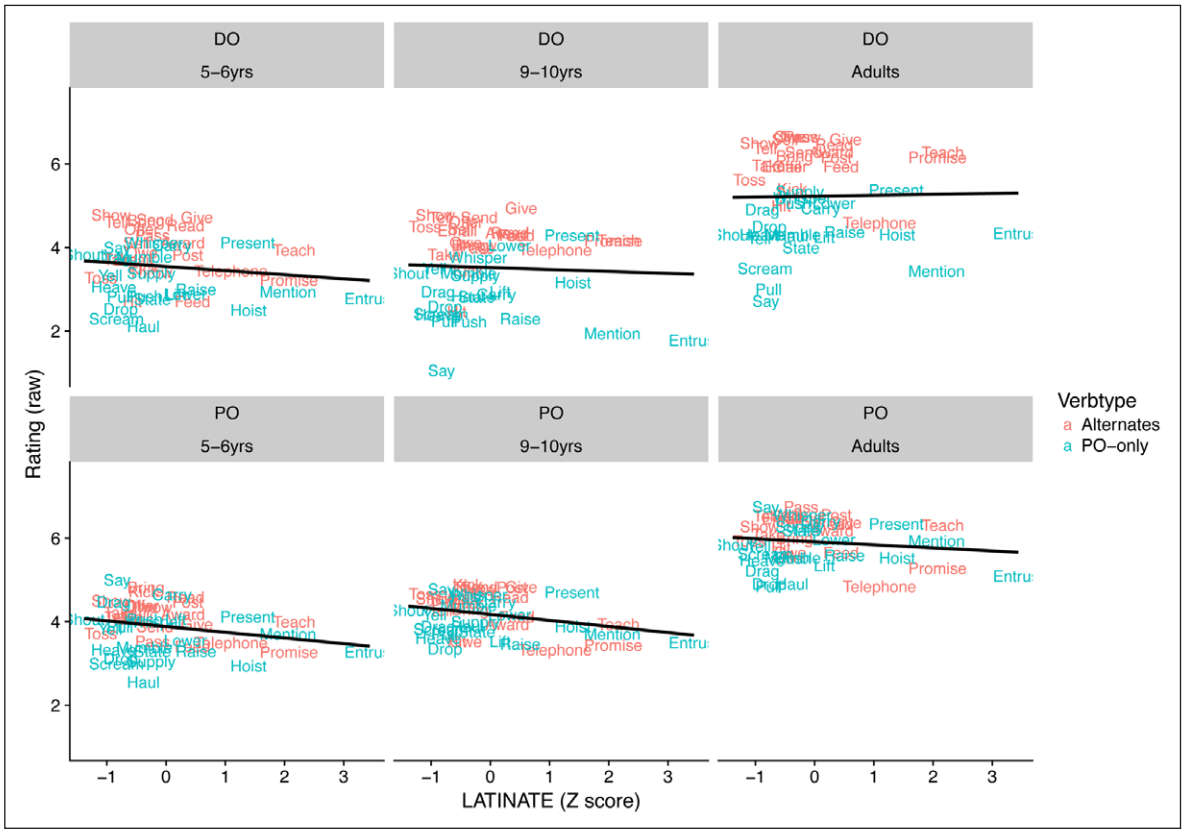

Figure 17: Study 2: Datives. Relationship between (X axis) the morphophonological "Latinate" predictor, in standard deviation units (Z scores), and participants' raw sentence ratings for (top) DO-datives and (bottom) PO-datives on the on the 5-point (children) or 7-point scale (adults).

Because the preemption and entrenchment predictors are much less highly correlated in the difference-score dataset ($r = 0.41$, as opposed to $r = 0.81$ for DO datives), it is here that we have the greatest opportunity to observe unique effects of each; not least because both predictors show large effects in the single-predictor models for this difference score data (**Figure 18**). And, indeed, dissociable effects of both preemption and entrenchment are observed. Although the former reaches significance for only the two older groups, and the latter only for the two younger groups, it would be premature to interpret this as evidence for a meaningful developmental pattern, since the nonsignificant effects – at $p = 0.13$ and $p = 0.09$ – can hardly be taken as convincing evidence of no effect (e.g., Altman & Bland, 1995; Gelman & Stern, 2006; Dienes, 2014).

In summary, then, while collinearity between these two predictors renders DO-dative judgment data inconclusive, the difference-score data suggests – like the locative data for Study 1 – dissociable effects of both preemption and entrenchment. That said, it is somewhat surprising that both do such a poor job of predicting PO-dative judgments. Recall that neither preemption nor entrenchment shows any effect, even in a single-predictor model (**Figure 9**). A possible explanation is that all verbs in this study were chosen to be grammatically acceptable in the PO-dative (since DO-dative-only verbs like *bet* and *wager* are both infrequent and unfamiliar to children), leaving little variance to explain. On the other hand, the very fact that all of these verbs *are* relatively acceptable in the PO-dative potentially constitutes a problem for preemption and entrenchment, since verbs do vary considerably in their predicted bias towards/against this construction, on the basis of these two predictors (see the bottom panels of **Figures 10**–**11**). A possible solution is that semantic and morphophonological factors are overriding distribution here. Consistent with this possibility, the only significant predictors in the model-comparison analysis are Broad DO 1 (associated with possession transfer) and Latinate (which differentiates PO-preferring verbs such as *donate* and *transfer* from PO/DO alternating verbs such as *give* and *send*).

Nevertheless, if we set aside the analyses for which preemption and entrenchment were highly correlated (DO-datives) and did not seem to be operational at all (PO-datives), and focus on difference scores, for which both effects were observed, the conclusion again is that – as for locatives (Study 1) – dissociable effects of both predictors can be seen in the same dataset.

Study 3: Various constructions (Ambridge et al, 2015).

Ambridge et al. (2015) compared the effects of entrenchment and preemption across eight different constructions (including the locative and dative constructions also investigated in Studies 1–2), but did not investigate verb semantics. Participants (5–6 year-olds, 9–10 year olds and adults) again rated sentences using a 5-point smiley-face scale. Participants rated verbs in eight constructions – intransitive, transitive, periphrastic causative, PO-dative, DO-dative, figure locative, ground locative and passive – see **Table 8**, which also shows the constructions taken to be the preempting construction for each target.

Although this study used counts from a sufficiently large corpus (the 200-million-word British television subtitle corpus, SUBTLEX-UK), raw scores (as opposed to difference scores) and maximal models, it shares two important problems with the studies revisited above. First, the entrenchment and preemption counts were based on raw frequency of occurrence rather than the chi-square measure. Consequently, these counts were obtained for – and all analyses therefore restricted to – *a priori* ungrammatical uses only. Second, because the entrenchment and preemption predictors were highly correlated, they were not compared in the same model.

Method

Participants. The judgment data reanalyzed here were provided by 72 children aged 5;2–6;8 (M = 5;10), 72 children aged 9;2–10;6 (M = 9;11) and 72 adults aged 18;1–22;2 (M = 19;1).

Entrenchment and Preemption predictors. The present reanalysis uses (in principle) de-confounded preemption and entrenchment counts (i.e., each corpus sentence counts towards only one or the other predictor),

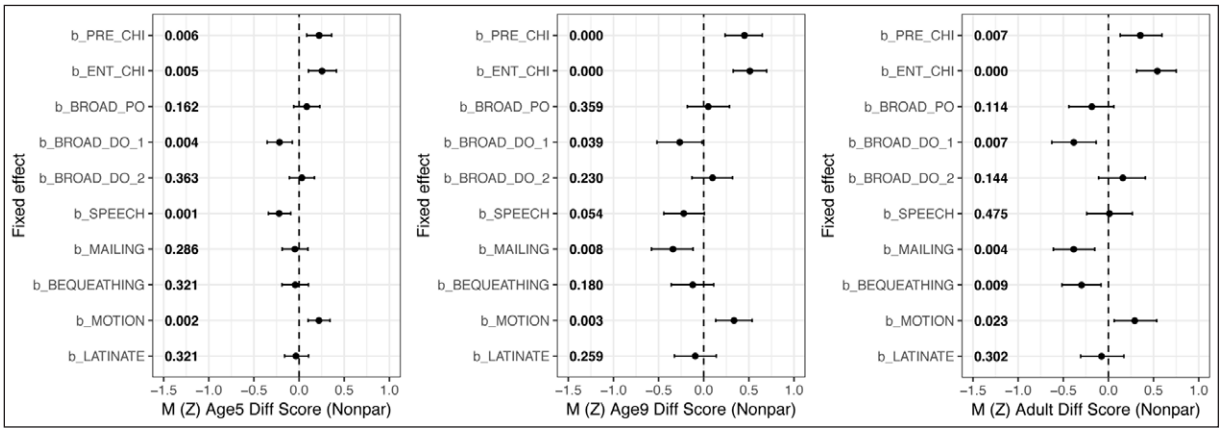

Figure 18: Study 2: Datives, nonpartial analysis of difference scores. Fixed effects (each from a separate regression model) for participants' difference scores (PO- minus DO-) and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

Table 8: Target and preemption constructions for Study 3: Various constructions.

Construction	Example	Preempting construction
Intransitive	The girl laughed/giggled (intransitive-only verb)	Passive (**Y was laughed/giggled [by X]*)
Transitive	*Bart laughed/giggled the girl	Periphrastic (*X made Y laugh/giggle*)
Intransitive	*The money took/removed (transitive-only verb)	Passive (*Y was taken/removed [by X]*)
Transitive	Marge took/removed the money	Periphrastic (*X made Y take/remove*)
Intransitive	The toy moved/rolled (alternating verb)	Passive (Y was moved/rolled)
Transitive	Marge moved/rolled the toy	Periphrastic (*X made Y move/roll*)
PO	Marge screamed/shrieked the warning to Homer (PO-only verb)	DO (*X screamed/shrieked Z Y*)
DO	*Marge screamed/shrieked Homer the warning	PO (*X screamed/shrieked Y to Z*)
PO	*Marge refused/denied the beer to Homer (DO-only verb)	DO (X *cost/fined Z Y*)
DO	Marge refused/denied Homer the beer	PO (*X refused/denied Y to Z*)
PO	Lisa showed/taught the answer to Homer (alternating verb)	DO (*X showed/taught Z Y*)
DO	Lisa showed/taught Homer the answer	PO (*X showed/taught Y to Z*)
Figure	Marge spilt/dribbled juice onto the rug (figure-only verb)	Ground (*X split/dribbled Z with Y*)
Ground	*Marge spilt/dribbled the rug with juice	Figure (*X spilt/dribbled Y onto Z*)
Figure	*Bart covered/coated mud onto Lisa (ground-only verb)	Ground (*X covered/coated Z with Y*)
Ground	Bart covered/coated Lisa with mud	Figure (*X covered/coated Y onto Z*)
Figure	Homer splashed/spattered water onto Marge (alternating verb)	Ground (*X splashed/spattered Z with Y*)
Ground	Homer splashed/spattered Marge with water	Figure (*X splashed/spattered Y onto Z*)
Active	Lisa looked like/resembled Marge (active-only verb)	Passive (**Y was looked like/resembled [by X]*)
Passive	*Marge was looked like/resembled by Lisa	Active (X *looked like/resembled Y*)
Active	Marge pushed/chased Homer (alternating verb)	Passive (Y was pushed/chased [by X])
Passive	Homer was pushed/chased by Marge	Active (X pushed/chased Y)

with the aim of comparing these predictors in the same model. In practice, however, they were again highly correlated for both the raw- and difference-score analyses ($r = 0.58$ and $r = 0.48$ respectively); a degree of collinearity that is again cause for concern. Unlike Studies 1–2, each of which looked at a single construction pair, the present study included eight different constructions (see **Table 8**). Descriptively, and in terms of how the study materials were put together, these constructions can be understood as forming four "alternation pairs" (intransitive/transitive, PO-/DO-dative, figure-/ground-locative, active/passive), with (except for active/passive), an equal number of verbs grammatical in (a) the first construction of the pair only (b) the second only or (c) both. However, neither "pair" nor "verb-type" (a-only, b-only, alternating) is entered as a factor in the main analysis, which looks at raw ratings for individual sentences.

In order to calculate the preemption predictor (in both the original study and the present reanalysis), it was necessary to stipulate – for each construction – the most closely semantically related, and hence potentially preempting, construction. The preempting constructions chosen are shown in the right-hand column of **Table 8**. The choice of preempting construction is straightforward (and, we hope, uncontroversial) with one exception: Errors in which transitive-only verbs are used in intransitive sentences (e.g., **The money took* meaning 'somebody took the money') were held to be preempted by the passive construction (e.g., *The money was taken [by somebody]*), with or without a by- phrase. This decision was taken (following Brooks & Tomasello, 1999) because the passive – like the intransitive inchoative (e.g., *The ball rolled*) – promotes the underlying semantic PATIENT to SUBJECT position, either dropping the underlying AGENT altogether, or demoting it to the by-phrase (e.g., *The ball rolled* is to *The ball was rolled* as **The money took* is to *The money was taken*).

Since, for completeness, the present study adds a difference-score analysis not present in the original, a decision had to be taken regarding the pairs of constructions used to calculate these scores. Given that a difference score represents the preference for one construction over a closely semantically related construction, the decision taken was to use each construction's preempting construction (defined as a close semantic competitor) when calculating difference scores. This necessitated the exclusion of intransitive sentences from the difference-score analysis, as no judgments were

collected for ratings of the relevant verbs in the designated semantically-related construction, the passive. **Table 9** shows how these difference scores were calculated, along with – in each case – a verb that is grammatical in the first construction, but not the second (the rating for which is subtracted from the rating for the first). Note, however, that this is purely for illustrative purposes; difference scores were calculated for all verbs, whether held to be grammatical in one or both constructions in the pai.

Results and Discussion

In terms of the statistical analyses, the only substantive difference between the present analyses and those presented in Studies 1–2 is that, rather than running separate analyses for each construction, we include target construction as a random effect (slope and intercept), as in the original analysis. This decision was taken because the aim of original study was to investigate whether preemption and entrenchment generalize across constructions and we wish to run an equivalent analysis. As in the original study, a separate analysis for each construction would not be possible, given the paucity of data, with just four verbs (for the active and passive constructions) or six verbs (all other constructions) rated in each (see **Table 8**). As for Studies 1–2, we built (a) a series of maximal single-predictor Bayesian models and – for the purposes of model comparison – (b) a series of near-maximal frequentist models (but with noncorrelated random effects), by removing each predictor in turn from the model specified below (in lme4 syntax).

```
Rating ~ (1 + PRE_CHI_LOG + ENT_CHI_LOG ||
Stype) + (1+PRE_CHI_LOG + ENT_CHI_LOG ||
Participant), PRE_CHI_LOG + ENT_CHI_LOG)
```

The correlations between predictor variables are shown in Appendix Table A1. Again, a potentially-problematic degree of collinearity between the predictor variables of preemption and entrenchment was observed for both the raw (r = 0.81) and difference scores analyses (r = 0.78).

Figure 19 shows the mean, 95% credible interval and (in bold) direction-corrected p_{MCMC} value for each single-predictor regression model. **Figures 20–21** plot against participants' judgments (Y axis), the preemption and entrenchment predictor respectively; both of which had 95% CIs that did not overlap with zero for all age-groups, in their respective single-predictor models. However, the model-comparison analysis (Appendix Table A2) revealed that while entrenchment explained variance above and beyond preemption, at least for the two older groups, the reverse was not the case.

For the difference score analysis, the 95% CIs for the single-predictor models always included zero, for both preemption and entrenchment (**Figure 22**). It is therefore somewhat surprising that the model-comparison procedure (see Appendix Table A2) suggested (just) significant effects of both preemption and entrenchment for the adults (p = 0.044 and p = 0.047 respectively), and of entrenchment only for the older children (p = 0.044).

Table 9: Calculation of difference scores for Study 3: Various constructions (A minus B).

Construction A (minuend)	Construction B (subtrahend)
Periphrastic causative (e.g., Bart made the girl laugh)	Transitive (e.g., *Bart laughed the girl)
PO-dative (e.g., Marge screamed the warning to Homer)	DO-dative (e.g., *Marge screamed Homer the warning
Figure-locative (e.g., Marge spilt juice onto the rug)	Ground-locative (e.g., *Marge spilt the rug with juice)
Active (e.g., Lisa looked like Marge)	Passive (e.g., *Marge was looked like by Lisa)

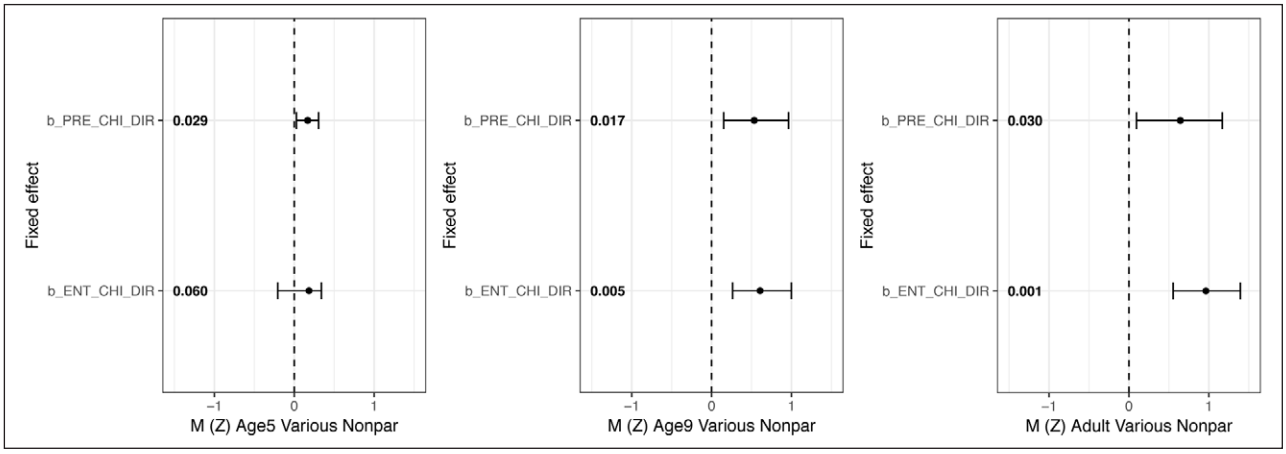

Figure 19: Study 3: Various constructions, nonpartial analysis. Fixed effects (each from a separate regression model) for participants' sentence judgments, and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

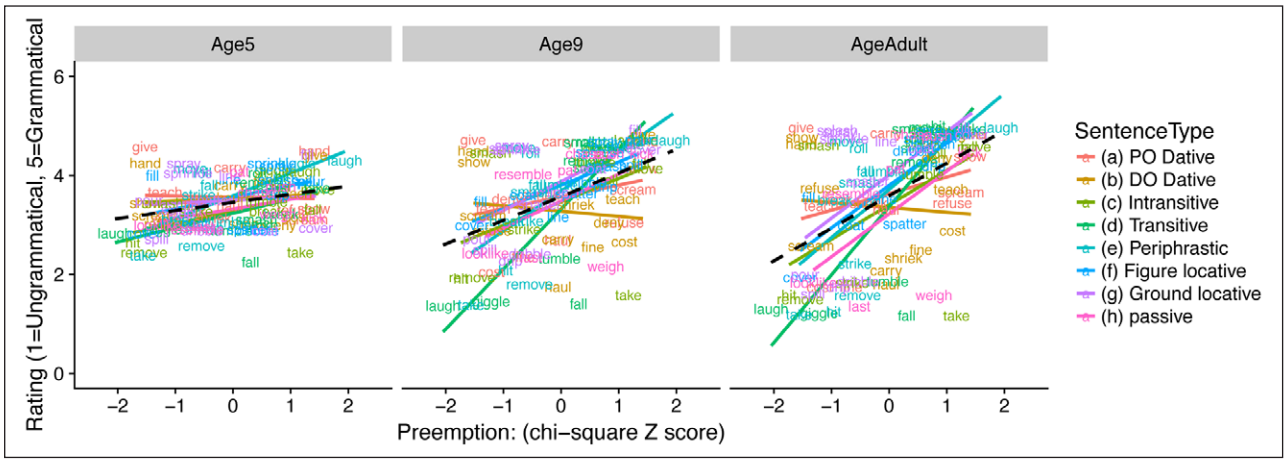

Figure 20: Study 3: Various constructions. Relationship between (X axis) the preemption predictor, in standard deviation units (Z scores), and participants' raw sentence ratings on the 5-point scale (used for both children and adults).

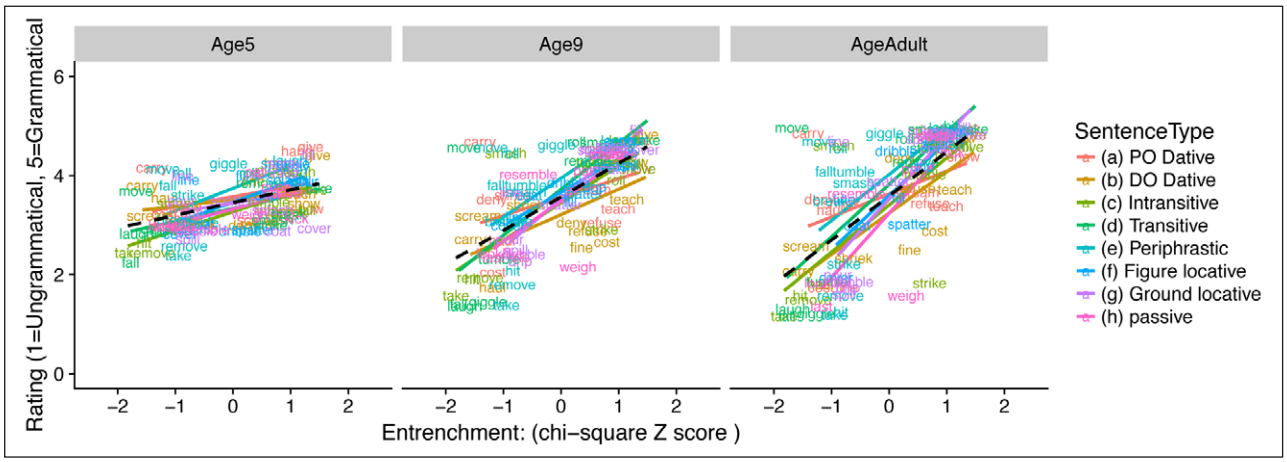

Figure 21: Study 3: Various constructions. Relationship between (X axis) the entrenchment predictor, in standard deviation units (Z scores), and participants' raw sentence ratings on the 5-point scale (used for both children and adults).

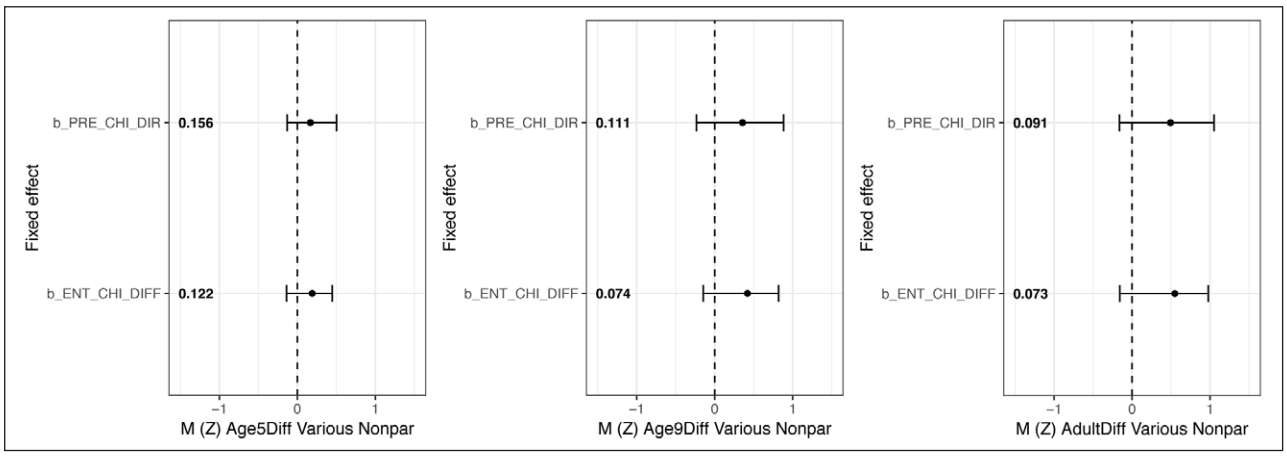

Figure 22: Study 3: Various constructions, nonpartial analysis of difference scores. Fixed effects (each from a separate regression model) for participants' difference scores (see Table 9) and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

This pattern is not easy to interpret, but given that both predictors show large effects in the raw-scores single-predictor analyses, and the two are highly correlated for both the raw and difference-scores analyses, probably the fairest conclusion is that preemption and/or entrenchment seems to be operational for this dataset, but we cannot tell which. Nevertheless, given that each explains unique variance above and beyond the other in the difference-scores

analysis (which may be particularly appropriate here, since it corrects for differences between construction pairs), this pattern is at least consistent with our conclusion from Studies 1–2 that, generally, both effects are observed.

Study 4: un- Prefixation (Ambridge, 2013; Blything et al., 2014)

Studies 1–3 reanalysed the data from three previous studies of overgeneralizations of verb argument structure involving locatives (Study 1), datives (Study 2) and various constructions (Study 3), with the aim of mediating between the preemption and entrenchment hypotheses. So far, our tentative conclusion is that, generally, both effects are observed, though it is extremely difficult to differentiate them, given that they show high-to-moderate correlations in almost all of the datasets analysed so far. Indeed, as we discussed in the Introduction, when studying verb argument structure constructions, a large correlation between measures of preemption and entrenchment is virtually inevitable, since verbs that occur with high frequency in a particular construction (preemption) tend to be frequent overall (entrenchment). As we will see in more detail shortly, the **verbal *un-* prefixation** construction does not, in principle, suffer from this shortcoming, because any preemption or entrenchment effect operates at the level of individual words (i.e., *different* lexical items), rather than sentence-level constructions in which the *same* verbs are used. This makes *un-* prefixation particularly valuable from the point of view of picking apart effects of preemption and entrenchment.

In this domain, an overgeneralization error occurs when a verb that may not appear grammatically with the prefix *un-* appears in this form (e.g., **unsqueeze; *uncome*),[11] to denote reversal of an action. It is important to note (as for the argument structure cases in Studies 1–3) the existence of some verbs that do undergo this generalization (e.g., *button/unbutton; fasten/unfasten*). For these errors (e.g., **unsqueeze; *uncome*), the most natural preempting form is the verb that expresses the intended meaning (e.g., *release; go*). Consequently, the prediction of the preemption hypothesis tested by Ambridge (2013) and Blything et al. (2014) was of a negative correlation between the acceptability of such errors (i.e., using raw rather than difference scores) and the summed frequency of the two nearest semantic competitor verbs (e.g., for **unsqueeze, release* and *loosen*). The decision to use – for each *un-* form – the *two* nearest competitor verbs, rather than one or three, was taken simply because this gave better coverage of the data (Ambridge, 2013, speculated that this is because most *un-* forms do not have a single perfect synonym, but casting the net wider than two catches less relevant, more distant synonyms). The prediction of the entrenchment hypothesis tested in these previous studies was of a negative correlation between the acceptability of such errors and overall verb frequency (e.g., all uses of *squeeze*, without the prefix *un-*). Finally, the prediction of the verb-semantics hypothesis tested in these previous studies was of a positive correlation between the acceptability of forms prefixed with *un-* and the extent to which the relevant verb was judged to exhibit a constellation of semantic properties thought to characterize the verbs that can appear with this prefix (e.g., covering, enclosing, surface-attachment, circular motion, hand-movements, change-of-state).

The reason that this construction is potentially particularly useful for distinguishing preemption and entrenchment is as follows: Given a particular overgeneralization error (e.g., **unsqueeze*), uses that preempt (e.g., *release, loosen*) and entrench away from this error (i.e., all non-*un-*prefixed forms of *squeeze*) are *different verbs* (rather than, as for Studies 1–3, the same verb in different sentence-level argument-structure constructions). Thus, unlike the three other studies reported in this paper, this study need not suffer from the problem that verbs that are frequent in the relevant preempting construction tend also to be frequent across the board (entrenchment). Unfortunately, however, it turns out that, in practice, the preemption and entrenchment predictors are highly correlated for this particular dataset (r = 0.66 for both the raw and difference-score analyses; for reasons set out below, the predictors are the same in the two analyses).

There are two apparent reasons for this large correlation. First, verbs with high overall frequency (entrenchment) such as *come, give, go* and *stand* denote common human actions, and – as such – actions that are also commonly reversed or undone. Consequently, the synonyms suggested by participants as preempting **uncome, *ungive, *ungo* and **unstand* are also of high frequency: *go, take, come* and sit. Second, when calculating the chi-square statistics for entrenchment and preemption (see **Table 10**), the values in the two leftmost cells – frequency of (a) the target verb and (b) all other verbs in *un-* form – are identical. This is entirely appropriate, and indeed unavoidable, since both measures reflect a trade of between witnessed *un-*forms and competitors (non-un- forms of the same verb for entrenchment, competing synonyms for the *un-* form, for preemption). An unavoidable consequence, however, is that verbs that occur very frequently in *un-* form will yield a very high chi-square statistic on both measures – entrenchment and preemption – thus driving up the correlation between them (note that this problem also applies to Studies 1–3).

Method

Similar to Studies 1–3, the main changes from the original analyses are the use of (a) chi-square predictors, (b) both single-predictor Bayesian models and frequentist model-comparison, (c) raw and difference scores (the original analyses used only the former) and (d) analyses conducted across all forms as opposed to (in the original study) separate

Table 10: Calculation of the entrenchment predictor for ***unbuckle***. Numbers refer to counts in the British National Corpus. Chi^2 = 299, reflecting a strong bias in favour of the ***un-*** vs bare form (relative to other verbs in the corpus).

	Freq of verb with *un-*	Freq of verb without *un-*
Buckle	15	246
All other verbs	3,714	1,413,504

analyses for *a priori* grammatical and ungrammatical un- forms. However, note that, although participants rated both *un-* forms (e.g., *unbutton, *unsqueeze*) and bare forms (e.g., *button, squeeze*) we do not conduct a separate analysis for ratings of bare forms: The question that such an analysis would answer ("How do learners determine the acceptability of verbs in their bare, citation form?") is not relevant to our purposes here. Indeed, except perhaps for a handful of very unusual verbs that are more frequent in *un-* than bare form (in the present set, just *uncork, unleash* and *unveil*), it is not clear that this question is meaningful. As for Studies 1–3, we reanalyze data from 5–6 year olds (collapsing across Ambridge, 2013 and Blything et al., 2014), 9–10 year olds (Ambridge, 2013) and adults (Ambridge, 2013). We do not reanalyze the 3–4 year-olds' data from Blything et al. (2014), given the original authors' conclusion that these data are "too noisy for detection of any mechanisms of restriction" (p. 3).

Participants. The judgment data reanalyzed here were provided by 38 children aged 5–6 (18 from Ambridge, 2013, *M* = 5;6, and 20 from Blything et al., 2014, *M* = 6;0), 18 children aged 9;10–10;10 (*M* = 10; 5), and 18 adults aged 18–21 (both older groups from Ambridge, 2013).

Preemption and Entrenchment predictors. Because the entrenchment and preemption counts were originally obtained from the BNC, there was no need to obtain new counts. As for Studies 1–3, we calculated new versions of these predictors based on the chi-square statistic (see **Tables 10–11**). Again, polarity (+/–) is used to indicate whether the (log transformed) chi-square value represents a bias towards or away from the *un-* form relative to the other verbs in the set.

While the chi-square predictor makes intuitive sense for entrenchment (proportion of uses with/without *un-* for the target verb versus all other verbs), this is less straightforwardly the case for preemption. On the face of it, it makes little sense to compare the ratio of *buckle: release+loosen* (the top two synonyms for *unbuckle*) to the combined ratio of *chain: release* + *free, pack: empty* + *remove, zip:open* + *reveal* etc., given that (unlike for the verb argument structure constructions in Studies 1–3) an entirely different set of lexical items is involved. On reflection, however, moving away from the level of surface forms and focussing on the underlying mechanism assumed by preemption, the use of a chi-square predictor is sensible: We want to know how likely is it that the reversal of *buckle* is expressed by (a) a completely different verb versus (b) the same verb prefixed by *un-, as compared to other verbs in the language*. Again, the comparison with other verbs in the language is crucial. Generally, un- forms are extremely rare. There is no *un-* form (at least in the present set) that is not, in absolute terms, vastly outnumbered by tokens of its two most-suggested synonyms. But there are a quite a few *un-* forms (*unbend, unbuckle, unbutton, undo, unfasten, unfreeze, unhook, unleash, unlock, unpack*) that are less outnumbered by their synonyms than are un- forms in general; and it is these forms that are predicted to be rated as particularly acceptable under the preemption hypothesis.

Table 11: Calculation of the preemption predictor for ***unbuckle***. Numbers refer to counts in the British National Corpus. $Chi^2 = 0.41$, reflecting a very small bias in favour of the ***un-*** form vs its synonyms (relative to other verbs in the corpus).

	Freq of verb with *un-*	Freq of top two synonyms for *un-* form
Buckle	15	8,327
All other verbs	3,714	1,747,148

Finally, note that, although we include a difference-score analysis, it is not appropriate to calculate a difference-score version of the entrenchment predictor for this analysis, as we did for Studies 1–3 (which is why the correlation between the entrenchment and preemption predictors is the same – $r = 0.66$ – for the raw-and difference-score analyses). Such a predictor is needed when we have counts from three categories (e.g., PO-datives, DO-datives, non-datives). In this case, we calculated entrenchment of each verb (a) away from the PO-dative (i.e., PO-datives vs non-datives), and (b) away from the DO-dative (i.e., DO-datives vs non-datives), and subtracted (b) from (a). For *un-*prefixation, we have only two counts: *un-*forms (e.g., *unbuckle, unbuckles, unbuckled*, etc.) and bare, non-*un-* forms (e.g., *buckle, buckles, buckled*, etc.). Thus the entrenchment predictor outlined in **Table 10** already represents each verb's relative bias towards the *un-* form and away from the bare form. The perfect complementary distribution of *un-* and bare (non-*un*) forms means that if we *did* decide to calculate the converse predictor – representing bias towards the bare form and away from the *un-* form – we would discover that it turned out to be simply the same chi-square value with opposite polarity (so that subtracting it from the original predictor would simply double the size of the latter). As in Studies 1–3, a difference-score version of the preemption predictor (for use in the analysis with difference scores as the dependent measure) would be superfluous (and mathematically equivalent to a raw-score version) since this predictor already measures a verb's relative bias towards *un-* versus bare form.

Semantic predictor. Like the original studies, the present analysis used a single semantic predictor derived, using principle components analysis, from semantic ratings obtained by Li and MacWhinney (1996). These authors asked 15 native English speakers to rate each verb as to whether it instantiates each of 20 semantic features thought to relate to the semantic cluster (or "cryptotype") of verbs prefixable with *un-* (e.g., *circular movement, change of state, manipulative action*).

Control predictors. The original analysis included control predictors of (a) verb type, a binary variable reflecting whether or not the *un-* prefixed form of the verb appears in the BNC, and (b) *un-*form frequency, the frequency of the *un-* form in this corpus. These control predictors were not included in the present reanalysis because the existence and frequency of the *un-* form is incorporated into the chi-square entrenchment and preemption predictors (see **Tables 10–11**). Two control predictors were retained unchanged from the original analysis. First, in lieu of a separate analysis for bare forms (which, as discussed above, would be unmotivated) we

included each participant's rating of the corresponding bare form as a control predictor in the main analysis (but, not in the difference-score analysis, where the difference score is calculated as bare-minus-*un*-form). Second, we included a separate group of participants' ratings of reversibility, to control for the possibility that the ability to appear with *un-* is simply a proxy for the extent to which the action denoted by the verb is semantically reversible.

Dependent variable. Participants rated bare and *un*-forms (e.g., *squeeze; *unsqueeze*) of 48 verbs (24 each that were listed as taking/not taking *un-* in the study of Li & MacWhinney, 1996), using a 5-point scale. Like the original analyses of Ambridge (2013) and Blything et al. (2014), the main analysis was conducted directly on participants' ratings of individual *un-* prefixed forms (i.e., raw scores), rather than difference scores. However, as for Studies 1–3 above, we also added difference score analyses, calculated as bare-minus-*un-* form (as usual, on verb-by-verb and participant-by-participant basis). Although it seemed important to include this difference-score analysis for completeness, it should be interpreted with extreme caution, given that – unlike the difference scores in Studies 1–3 – these difference scores (e.g., ratings for *squeeze* minus **unsqueeze*) do not reflect two alternative formulations of the same (or very similar) message.

Results and Discussion

As for Studies 1–3, we built (a) a series of maximal single-predictor Bayesian models and – for the purposes of model comparison – (b) a series of near-maximal frequentist models (but without correlation between random effects being included in the model), by removing each predictor in turn from the model specified below (in lme4 syntax).

UnRating ~ (1 + Preemption + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics || Participant) + (1|Verb) +Preemption + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics)

The correlations between predictor variables are shown in Appendix Table A1. Again, a potentially-problematic degree of collinearity was observed between the predictor variables of preemption and entrenchment (r = 0.66), as well as between semantics and entrenchment (r = 0.64), placing additional importance – for this dataset – on the model-comparison analysis.

Figure 23 shows the mean, 95% credible interval and (in bold) direction-corrected p_{MCMC} value for each single-predictor regression model. **Figures 24–28** plot against participants' judgments (Y axis), all five predictors – Preemption, Entrenchment, Bare-form rating, Reversibility

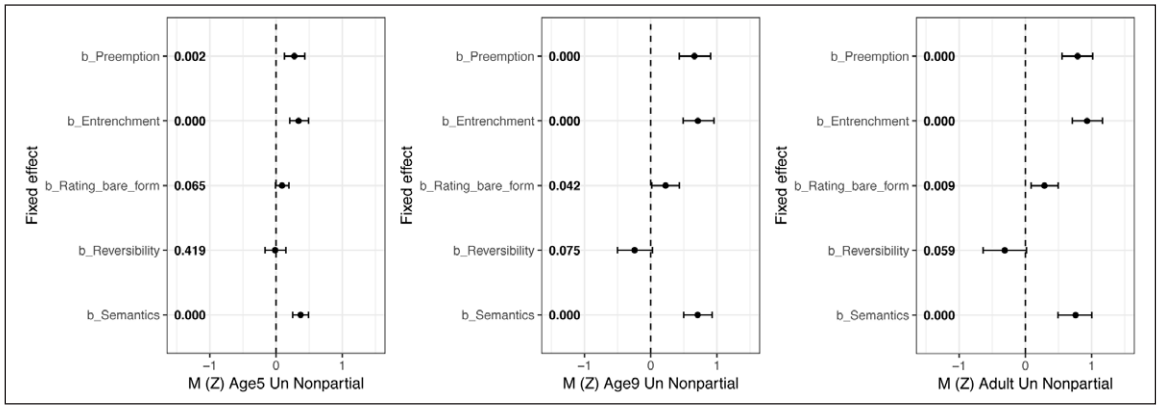

Figure 23: Study 4: *Un-* prefixation, nonpartial analysis. Fixed effects (each from a separate regression model) for participants' judgments of *un-* forms, and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores). However, see the main text for concerns regarding the interpretability of a nonpartial analysis for this particular dataset.

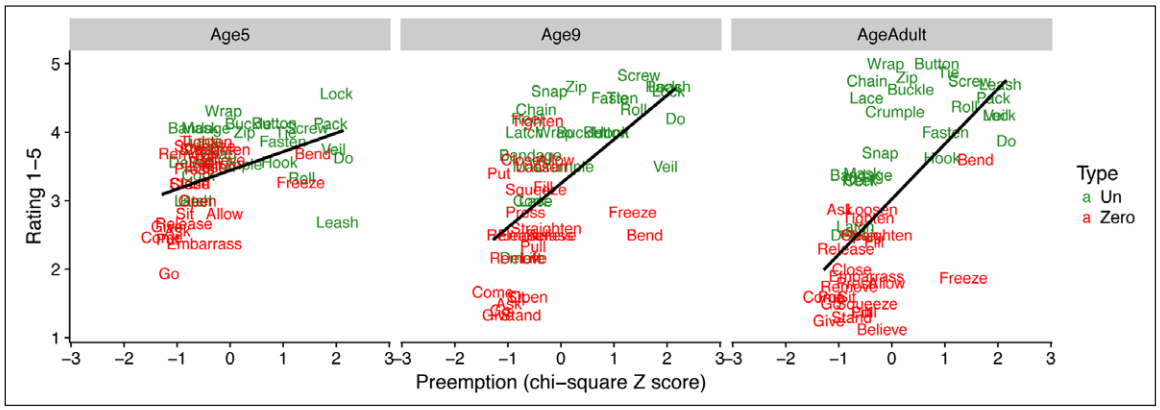

Figure 24: Study 4: *Un-* prefixation. Relationship between (X axis) the preemption predictor, in standard deviation units (Z scores), and participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

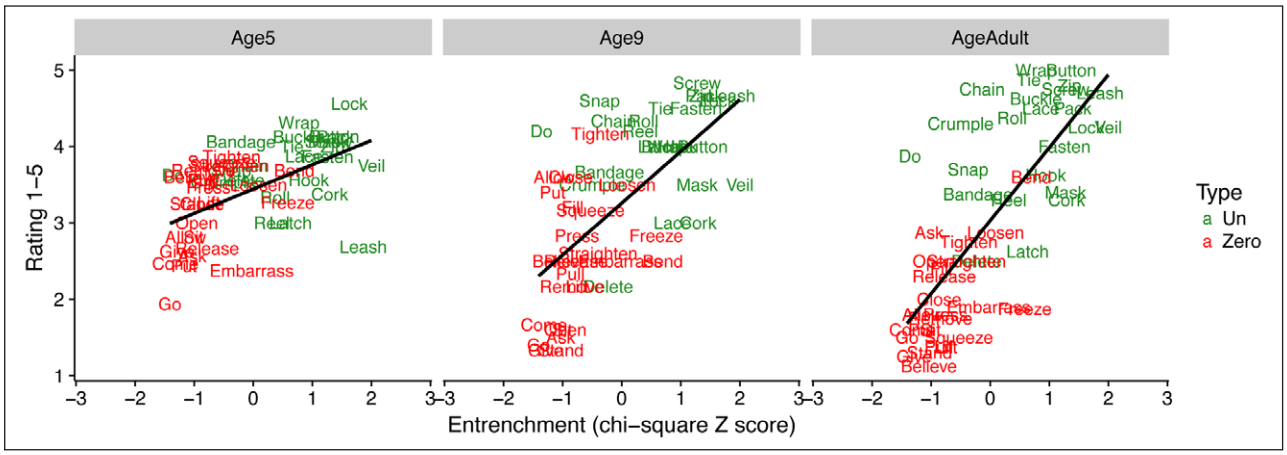

Figure 25: Study 4: *Un-* prefixation. Relationship between (X axis) the entrenchment predictor, in standard deviation units (Z scores), and participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

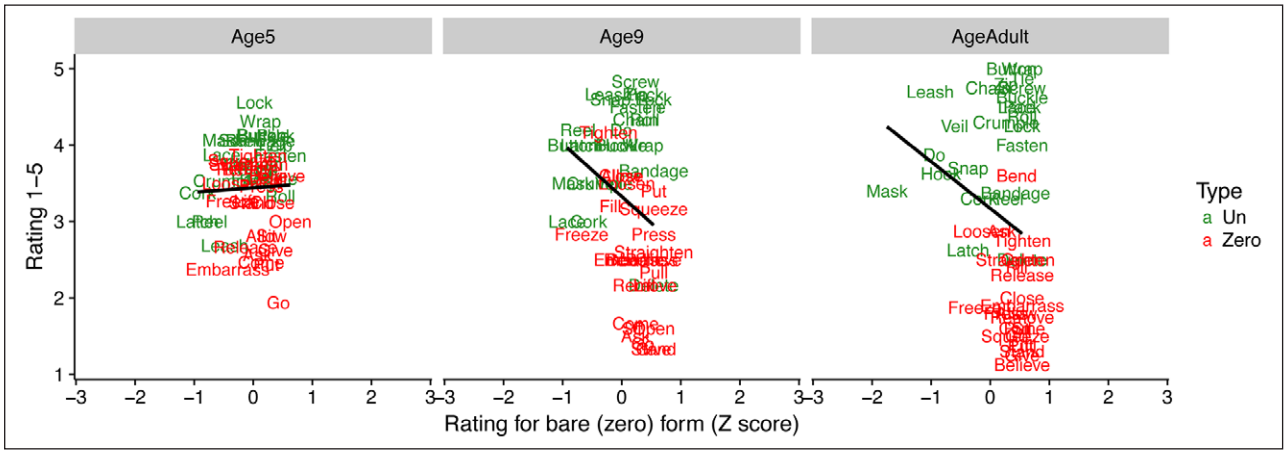

Figure 26: Study 4: *Un-* prefixation. Relationship between (X axis) participants' bare-form ratings, in standard deviation units (Z scores), and participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

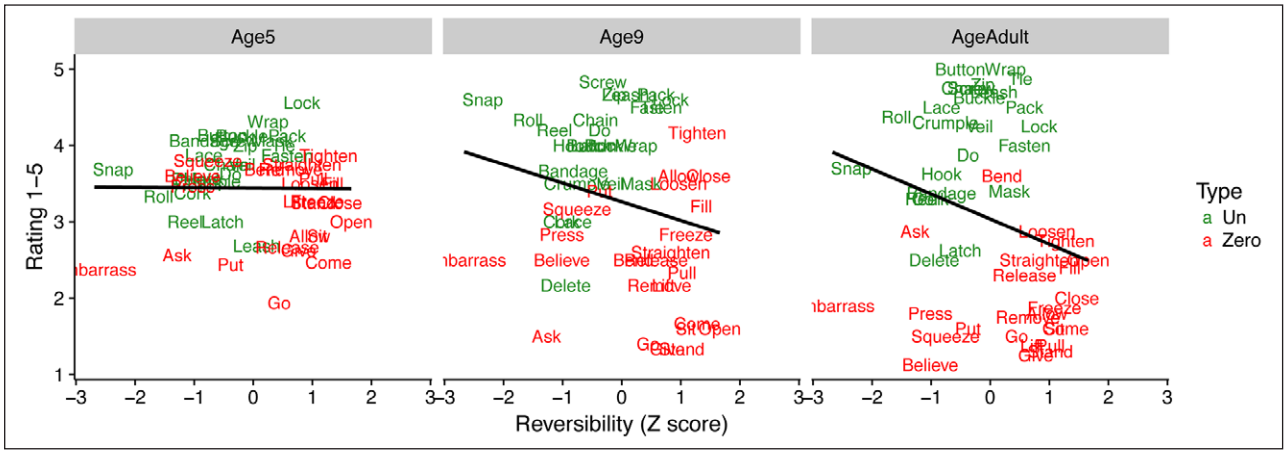

Figure 27: Study 4: *Un-* prefixation. Relationship between (X axis) participants' reversibility ratings, in standard deviation units (Z scores), and (different) participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

and Semantics – all of which had CIs that did not overlap zero for at least one age group. Indeed, the three non-control predictors, Preemption, Entrenchment and Semantics, all yielded p_{MCMC} values of exactly zero (i.e., all samples from the posterior distribution were greater than zero), for all three age groups. Furthermore, in the model-comparison analysis (see Appendix Table A2), all five predictors explained unique variance for adults. For the older children, *all* except

Figure 28: Study 4: *Un-* prefixation. Relationship between (X axis) participants' semantic ratings (from Li & MacWhinney, 1996), in standard deviation units (Z scores), and (different) participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

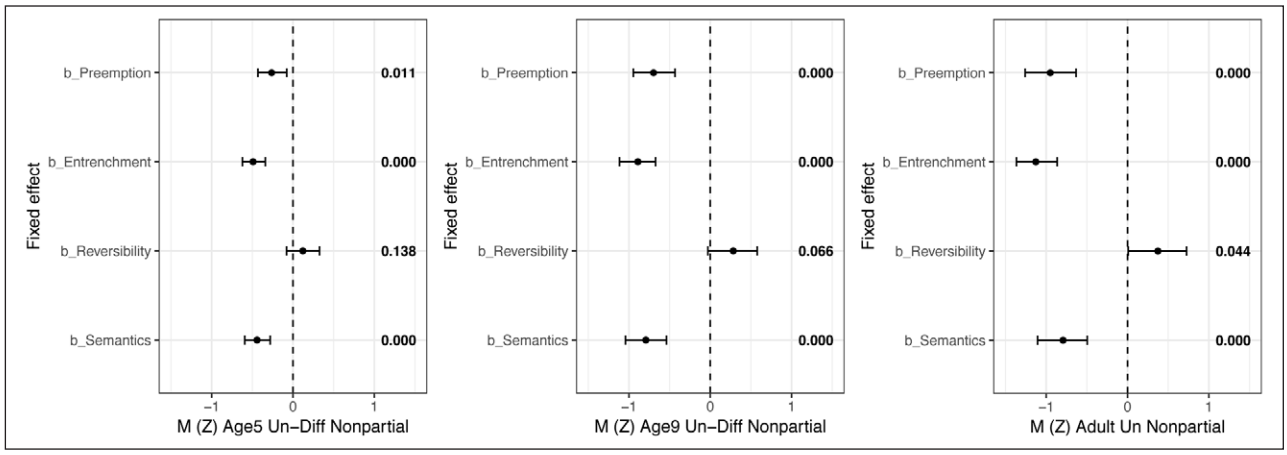

Figure 29: Study 4: *Un-* prefixation nonpartial analysis of difference scores. Fixed effects (each from a separate regression model) for participants' difference scores (bare minus un- forms) and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores). However, see the main text for concerns regarding the interpretability of a nonpartial analysis for this particular dataset, and regarding the use of bare-minus-*un*-difference scores.

entrenchment explained unique variance, while – for the younger children – only semantics did so.

This pattern is largely confirmed by the difference-scores analysis. Preemption, Entrenchment and Semantics again yield p_{MCMC} values of exactly zero in the single-predictor models, for all three age groups (**Figure 29**). In the model-comparison analysis, all predictors except Semantics ($p = 0.06$) explain unique variance for adults and – again – all except entrenchment do so for the older children. One difference is that entrenchment explains unique variance for the younger children (though neither entrenchment nor preemption did so in the raw-scores analysis).

It would be possible, if tricky, to attempt a developmental explanation of why – at least on the basis of difference scores – 5–6 year olds show a unique effect of entrenchment only, 9–10 year olds of preemption only, and adults of both. But, as for Study 1, such an explanation would seriously risk constituting an over-interpretation of noisy data, particularly given that (a) younger children do not show a unique effect of entrenchment for the raw (as opposed to difference-score) and (b) both entrenchment and preemption and observed for every age group in single-predictor models. A more robust conclusion, then – particularly if we treat the adult data as a gold standard – is that unique effects of both preemption and entrenchment are observed for verbal *un-* prefixation.

In retrospect, the finding of an effect of preemption above and beyond entrenchment for verbal *un-* prefixation should not be surprising. After all, preemption was originally devised to explain overgeneralizations involving derivational morphology, albeit at the noun level (e.g., **cooker* for *cook*; Clark & Clark, 1979). Indeed, perhaps more so than at the sentence level, preemption enjoys a great deal of intuitive plausibility at the morphological single-word level. Intuitively, the reason we don't say *cooker* (for the person) is because we say *cook* (the preempting alternative). Intuitively, the reason we don't say **uncome* or **unsit* is because we say go or stand. So, it would be odd if the availability of these competing alternatives (as measured by preemption) did not explain variance in the

(un)acceptability of such forms. But, from this standpoint, the finding of an effect of entrenchment above and beyond preemption is quite surprising. Ungrammatical *un-* forms such as **uncome* or **unsit* do not compete semantically with their bare forms (e.g., *come, sit*), unless they do so extremely indirectly (e.g., *he sat down, and then I told him not to sit there anymore*). Thus, it is difficult to explain why the availability of such bare forms (as measured by entrenchment) explains variance in the (un) acceptability of these un- forms (e.g., **uncome* or **unsit*).

We therefore decided to conduct, with adults only, an extended replication of this final study, in order to explore the robustness of our finding that preemption and entrenchment seem to explain unique variance in participants' judgments of un- forms.

Study 5: A new study of un- prefixation

This final study differed from the adult part of Ambridge (2013) in two important respects. First, in order to ensure the robustness of the findings, we ran a larger sample in terms of both participants (N = 50, as opposed to N = 18) and verbs (all 160 verbs originally studied by Li and MacWhinney, 1996, as opposed to a subset of just 48). Most of the additional verbs were of considerably lower frequency, since the original 48 had been selected for their suitability for use with young children, and so were relatively common. This also increased the number of verbs that did not have a straightforward pre-empting alternative (though, of course, we did not know this until we had completed the part of the study in which participants suggest pre-empting alternatives). Second, in order to allow preemption and entrenchment to be further differentiated, in a way that was not possible in Ambridge et al (2013), we allowed participants who took part in the task of suggesting preempting alternatives to *un-* forms to answer "none", where this seemed appropriate to them. This allows us to conduct an additional test of preemption (following the logic of Robenalt & Goldberg, 2015, 2016) by comparing ratings for *un-* forms for which participants, as a group, did and did not suggest a competing, pre-empting alternative. This also allows us to conduct an additional test of entrenchment, by looking for an effect of entrenchment solely across un- forms which are deemed not to have a pre-empting alternative.

Ethics

This study was approved by the University of Liverpool Research Ethics Committee. Participants provided consent via an online form.

Participants

Participants were recruited via Prolific (http://Prolific.ac). Fifty participants were recruited for the main part of the study (acceptability judgments of *un-* prefixed and bare verb forms). Each participant provided 160 ratings of *un-* forms (as well as 160 ratings of bare forms), for a total of 8,000 datapoints (as compared to just 864 adult datapoints in Ambridge, 2013). An additional 15 (different) participants were recruited to provide reversibility ratings and to suggest possible preempting synonym forms (the same number as in Ambridge, 2013). In this case, power is not affected by the number of participants, because the reversibility ratings and pre-empting synonym forms are combined across participants to yield predictor variables. Prolific's screening criteria were used to recruit only first language speakers of English aged 18–60, with A Levels/High School.

Acceptability judgment task

Each participant rated all 160 verbs from Li and MacWhinney (1996), in both *un-* prefixed and bare form, on a 5-point visual scale. Stimuli were presented in random order using Qualtrics (https://www.qualtrics.com). Participants were given the following instructions.

> Your task in this study is to rate 320 VERBs for grammatical acceptability on a five-point scale. Acceptability is a sliding scale, not a yes/no judgment, so please try to use the whole of the scale. First, here are some warm-up/practice ratings, with suggested answers shown afterwards. For context, the VERB is shown in a full sentence, though with some words replaced with [X] and [Y]. You should mentally fill in the [X] and [Y] with whatever makes the best sentence for you. For example, you might read "The [X] BROKE the cup" as "The girl BROKE the cup". However, this is just to give you some context for the VERB. Your task is to rate the acceptability of the VERB itself (always shown in CAPITALS), rather than the sentence as a whole. Please complete the warm-up sentences below, then click NEXT to see the suggested answers.

Participants completed seven warm-up practice trials (see Appendix of Ambridge, 2013, for details), before completing the main part of the study. For the main part of the study, all verb forms (both *un-* and bare) were presented in the sentence *The [X] VERBED the [Y].* In Ambridge et al (2013), verbs were presented in sentences (e.g., *Homer unbroke the plate*). However, this is somewhat problematic in that the acceptability of the sentence also varies according to the nouns chosen (e.g., *Homer unsnapped the buckle/strap/ruler*), with noun preferences presumably varying from participant to participant. In the present study, we switched to presenting verbs in an abstract frame, with the aim of achieving ratings of the *maximum possible* acceptability of each form in a sentence context that is ideal for that participant.

Reversibility and synonyms task

Fifteen participants rated the reversibility of all 160 verbs, as per the following instructions (identical to those used in Ambridge (2013):

> Some **actions** are reversible. For example, if a shopkeeper raises his prices, he can reverse this action by lowering them. Some actions are not reversible. For example, if a chef **bakes** a cake he cannot reverse this action to end up with the raw ingredients. Some actions are somewhere in between. For example, if a chef **boils** his soup he can reverse this

> action by cooling it down again, but the reversal will not be quite complete as the flavour and texture of the soup will have changed. The first part of this study comprises a list of 160 actions. For each **action**, your task is to rate the extent to which the **action** is or is not reversible on a 7-point scale.

The same participants then completed a synonym-generation task in which they suggested synonyms for (i.e., preempting, competing-alternatives for) the *un-* form of each verb, as per the following instructions:

> For each action below, your task is to think up one or (maximum) two words that mean the **reversal** of this action (if you put two words, please separate them with a comma; e.g., word1, word2). We're not looking for words that are just *opposites* of the action, but that actually mean the **reversal** of that action: putting things back to how they were before (e.g., for the word *connect*, you might choose to write *disconnect*, since this means the reversal of the connecting action). If there is no suitable word that means the reversal of that particular action, please put **none**. **IMPORTANT:** You should NOT write words that you would consider "ungrammatical" (i.e., not real English words). **VERY IMPORTANT: You MAY NEVER write an *un-* word**, even if this word has the right meaning. For example, if the action is bolt, then unbolt would have the right meaning (as it reverses the bolting action) **BUT YOU MAY NOT WRITE UNBOLT.** Instead, you must try to come up with alternatives that **do NOT start with *un-*** (or put **none**).

Note that (as in Ambridge, 2013) the last part of these instructions (beginning "VERY IMPORTANT") prevents participants from suggesting *un-* forms of verbs *other* than the target verb. For example, given the action *connect*, this instruction prevents the suggestion not only of *unconnect*, but also of *undo, untie* etc. However, this was deliberate, as, otherwise, we would have risked participants using a handful of light *un-* verbs (such as undo) as suggestions for denoting the reversal of almost any action.

This task was very similar to that used in Ambridge (2013), but with two important changes. First, participants were asked to generate a maximum of two synonyms, as opposed to five in Ambridge (2013). This is because, in this previous study, any synonyms beyond the first two were always such distant synonyms to barely qualify as such (and, in practice, participants almost never suggested more than two). Second, and more importantly, participants were given the option of writing "none" if no synonym was available (as opposed to, in Ambridge, 2013, being encouraged to generate synonyms for every *un-* form, no matter how indirect they might be). This change is crucial, as it allows us to test for an effect of preemption by comparing ratings for un- forms for which participants, as a group, did and did not suggest a competing alterative (following Robenalt & Goldberg, 2015, 2016). An *un-*form was deemed to have a competing alternative (has-CA) if the number of participants suggesting either of the two most-suggested forms was greater than the number of participants suggesting "none". Otherwise it was deemed not to have a competing alternative (no-CA).

- **Has-CA verbs (*N* = 60):** *unagree, unallow, unappear, unapprove, unarrange, unassemble, unbegin, unbelieve, unbend, unbreak, unbring, uncapture, uncharge, unclasp, unclose, uncome, unconnect, uncontinue, undelete, undetach, unengage, unfasten, unfill, unfind, unfree, unfreeze, unget, ungive, ungrip, unhate, unhold, uninfect, unintegrate, unkeep, unlearn, unlift, unlike, unlive, unlock, unloosen, unmelt, unmount, unobey, unopen, unpress, unpull, unput, unremove, unseparate, unshow, unsit, unsqueeze, unstand, unstart, unstop, unstraighten, untake, untighten, untrust, unwrite.*
- **No-CA verbs (*N* = 100):** *unaffect, unaffiliate, unarm, unask, unbandage, unbecome, unbind, unbolt, unbraid, unbuckle, unbury, unbutton, uncall, unchain, unclear, unclench, unclog, uncoil, unconfirm, uncork, uncover, uncrumple, uncurl, undeprive, undo, undress, unembark, unembarrass, unentangle, unexpel, unfold, ungo, ungrow, unhang, unhear, unhelp, unhinge, unhitch, unhook, uninvite, unlace, unlatch, unleash, unlink, unload, unlocate, unlook, unmake, unmantle, unmask, unmove, unpack, unpat, unpay, unpeel, unplace, unplant, unplay, unplug, unpose, unpossess, unprove, unravel, unreach, unreel, unrelease, unreverse, unroll, unrun, unsay, unscramble, unscrew, unsee, unsettle, unsheathe, unslip, unsnap, unsolve, unspeak, unspill, unsplit, unsprinkle, unstrap, unstrip, untack, untalk, untangle, untell, untie, unturn, untwist, ununite, unuse, unveil, unwait, unwalk, unwind, unwork, unwrap, unzip.*

Frequency counts and predictors

New frequency counts were taken from the British National Corpus all texts (http://corpora.lancs.ac.uk/BNCweb/), and were used to calculate the chi-square entrenchment and preemption predictors in the same way as for Study 4, with two exceptions: First, the preemption predictor was calculated using the frequency of the *single* most commonly-suggested synonym (competing alternative) as opposed to – in Ambridge (2013) and Study 4 above – the sum of the *two* most commonly-suggested synonyms. The reason for this change was that the present, much larger verb set contains many more verbs for which few direct synonyms are available (presumably due to the addition of a large number of lower-frequency verbs that denote very specific actions). As a consequence, the number of participants suggesting the second most popular synonym was always low: 0 participants (14 verbs), 1 participant, meaning that the second-most-popular synonym was chosen arbitrarily (17 verbs), 2 participants (21 verbs), 3 participants (5 verbs) or 4 participants (3 verbs). Never was the second-most-popular synonym mentioned by 5 or more participants (out of 15). This we took as an indication that, as a rule, these second-tier synonyms were too distant to qualify as such. Second, the preemption

predictor was, of course, calculated only for the 60 verbs for which participants suggested a potentially-preempting competing alternative form.

Semantic predictor

Semantic feature ratings from Li and MacWhinney (1996) were used to create a single semantic predictor following the Principal Components Analysis procedure outlined in Ambridge (2013). The procedure used was identical, except for the fact that it was run over 160 as opposed to 48 verbs.

Results

The final dataset included ratings of *un-* prefixed and bare forms of 160 verbs, for 60 of which ("has-CA verbs") participants suggested a potentially-preempting competing alternative to the *un-* form.

Main analysis

Before proceeding to more detailed analyses designed to further dissociate preemption and entrenchment, we first conducted an analysis that took the same form as that conducted for Study 4 (which required the exclusion of the 100 verbs for which no preemption statistic was calculated; i.e., the no-CA verbs). That is, we again built (a) a series of maximal single-predictor Bayesian models and – for the purposes of model comparison – (b) a series of (in this case) maximal frequentist models, by removing each predictor in turn from the model specified below (in lme4 syntax).

UnRating ~ (1 + Preemption + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics | Participant) + (1|Verb) +Preemption + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics)

The correlations between predictor variables are shown in Appendix Table A1. Again, a potentially-problematic degree of collinearity was observed between the predictor variables of preemption and entrenchment (r = 0.82), as well as – to a lesser extent – between semantics and preemption (r = 0.44), and between semantics and entrenchment (r = 0.54), again emphasizing the importance of the model-comparison analysis.

Figure 30 shows the mean, 95% credible interval and (in bold) direction-corrected p_{MCMC} value for each single-predictor regression model. **Figures 31–35** plot against participants' judgments (Y axis), all five predictors – Preemption, Entrenchment, Bare-form rating, Reversibility

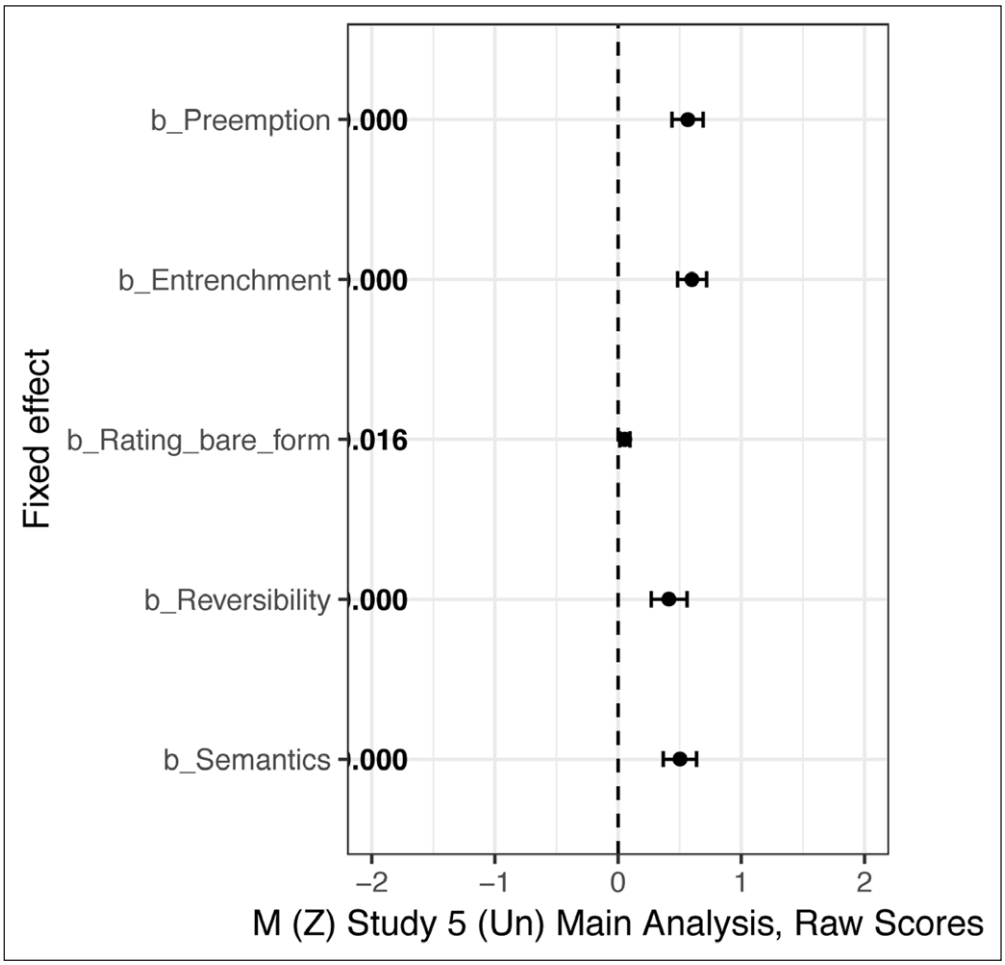

Figure 30: Study 5: New adult study of *un-* prefixation, nonpartial analysis. Fixed effects (each from a separate regression model) for participants' judgments of *un-* forms, and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores). However, see the main text for concerns regarding the interpretability of a nonpartial analysis for this particular dataset.

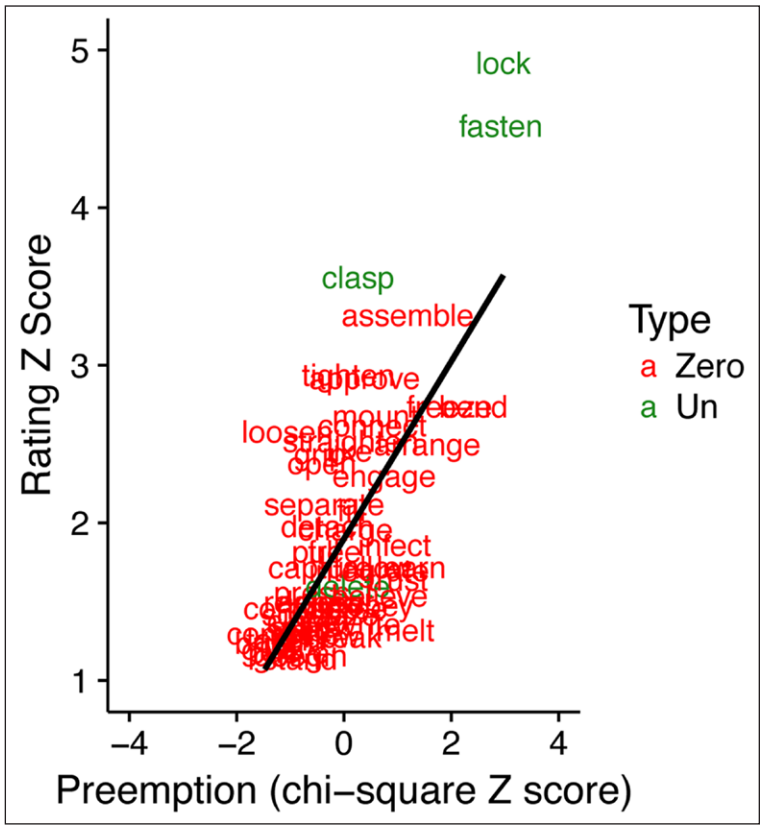

Figure 31: Study 5: New adult study of *un-* prefixation. Relationship between (X axis) the preemption predictor, in standard deviation units (Z scores), and participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

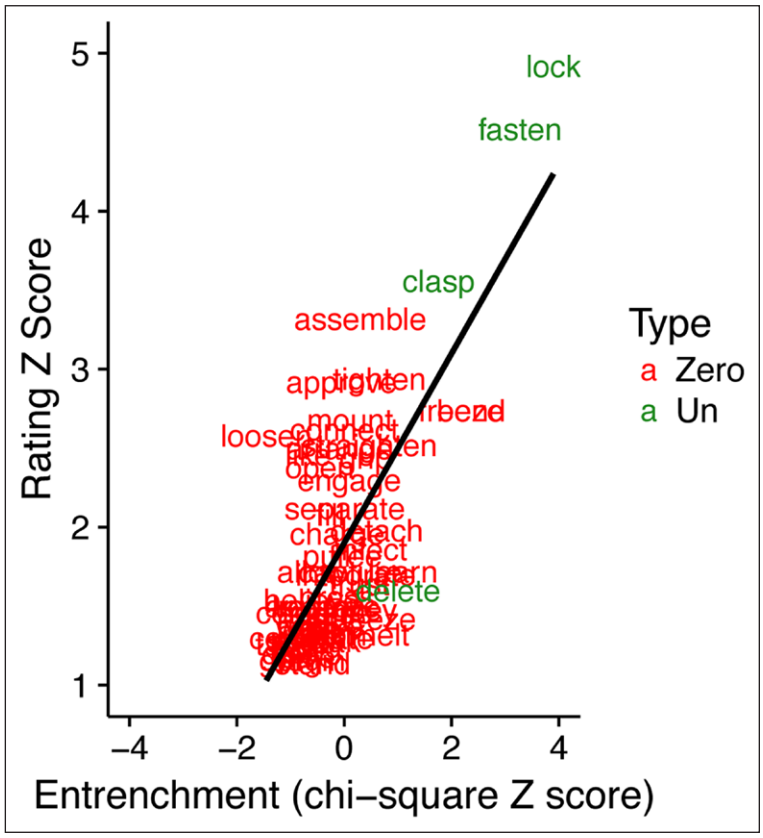

Figure 32: Study 5: New adult study of *un-* prefixation. Relationship between (X axis) the entrenchment predictor, in standard deviation units (Z scores), and participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

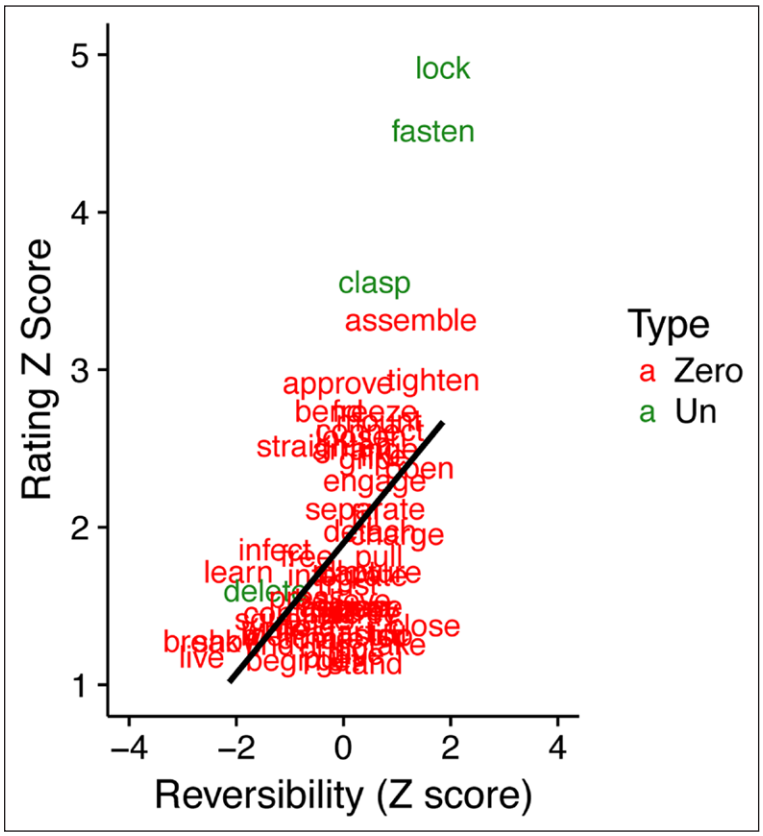

Figure 33: Study 5: New adult study of *un-* prefixation. Relationship between (X axis) participants' bare-form ratings, in standard deviation units (Z scores), and participants' raw un- form ratings on the 5-point scale (used for both children and adults).

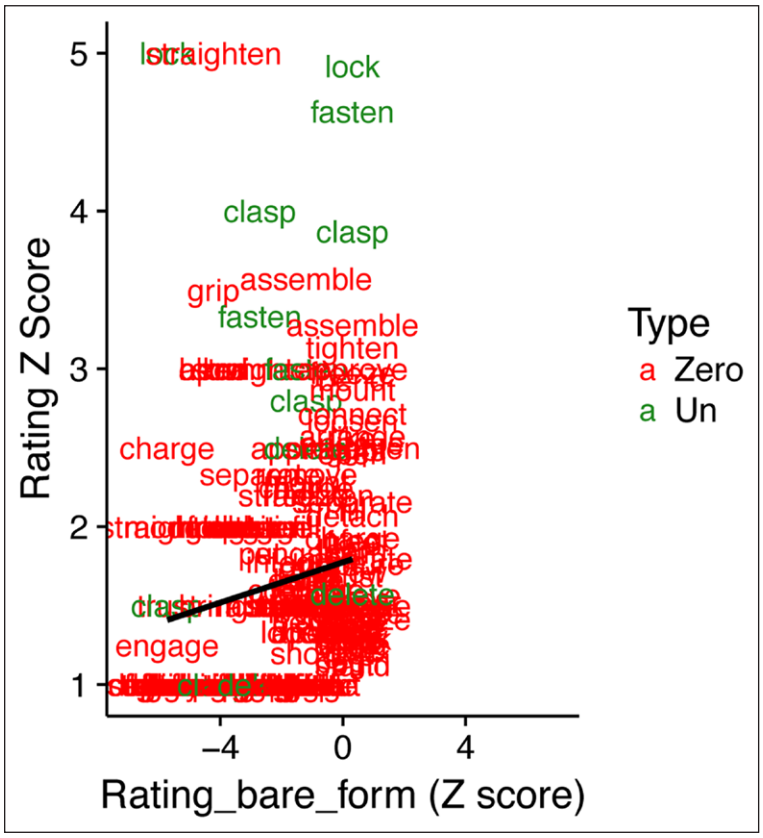

Figure 34: Study 5: New adult study of *un-* prefixation. Relationship between (X axis) participants' reversibility ratings, in standard deviation units (Z scores), and (different) participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

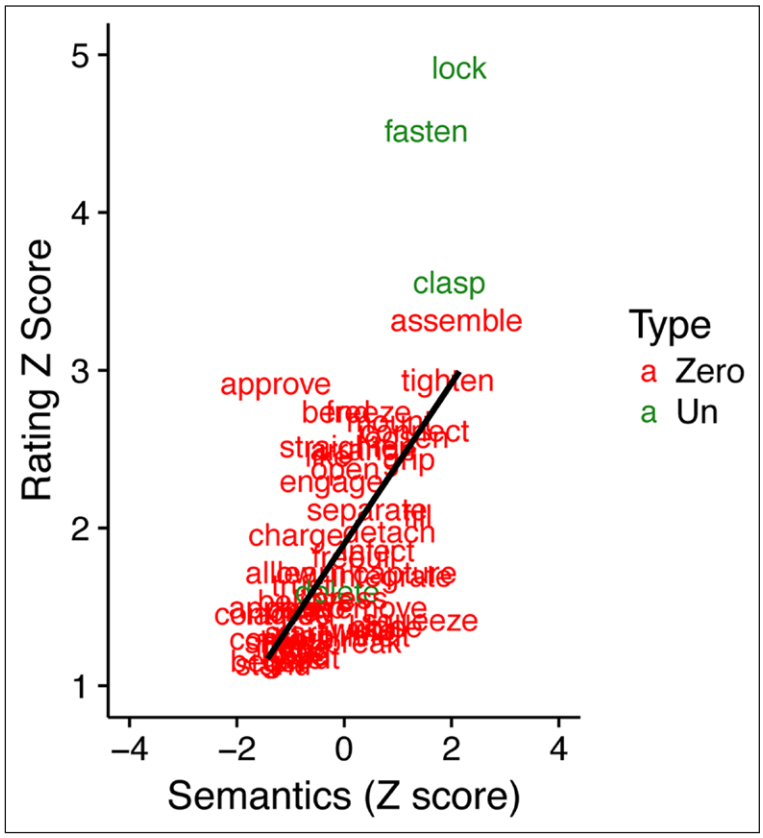

Figure 35: Study 5: New adult study of *un-* prefixation. Relationship between (X axis) participants' semantic ratings (from Li & MacWhinney, 1996), in standard deviation units (Z scores), and (different) participants' raw *un-* form ratings on the 5-point scale (used for both children and adults).

and Semantics – all of which had CIs that did not overlap zero. Indeed, all but the control predictor of rating for the bare form yielded p_{MCMC} values of exactly zero, meaning that all samples were in the predicted direction. The model comparison analysis (see Appendix Table A2) revealed that all five predictors explained unique variance, replicating from Study 4 the adult finding of dissociable effects of preemption, entrenchment and semantics with a new, larger group of participants, and a new, larger verb set. The difference-score analysis confirmed this pattern, with all four predictors (bare-form rating is not a predictor, since it is included in the difference-sore calculation) displaying non-zero-overlapping CIs in single-predictor models (**Figure 36**), and explaining unique variance in the model-comparison analysis (Appendix Table A2).

An additional test of the preemption hypothesis

Recall that, unlike in Ambridge et al (2013), participants in the present Study 5 who were invited to suggest potentially-preempting competing alternatives for *un-* forms were allowed to answer "none", where they felt this was appropriate. This allows for an additional test of preemption; albeit a rather narrower test that treats preemption as an all-or-nothing affair, when – in reality – a central assumption of the account is its gradient nature. This caveat notwithstanding, the preemption hypothesis clearly predicts that the very existence of a plausible competing alternative form (e.g., *disappear*) will reduce the acceptability of the relevant *un-*form (in this case, *unappear*). In order to test this prediction, we investigated whether a binary variable reflecting the existence or not of a competing alternative form (has-CA/no-CA, coded as 1/-1) predicts participants' acceptability judgments of *un-*forms in a model-comparison analysis; i.e., after controlling for entrenchment, semantics, and the control predictors of reversibility and the rating for the corresponding bare form, as per the following lme4 syntax:

> UnRating ~ (1 + Has_CA + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics | Participant) + (1|Verb) + Has_CA + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics).

This has-CA predictor was only moderately correlated with the Entrenchment predictor, and in the opposite direction to that predicted ($r = –0.33$, point bi-serial correlation; see Appendix Table A1). Again, the existence of collinearity between – in this case – semantics and entrenchment ($r = 0.58$), and between semantics and reversibility ($r = 0.46$) highlights the importance of model comparison.

Figure 37 shows, for this additional preemption analysis, the mean, 95% credible interval and (in bold) direction-corrected p_{MCMC} value for each single-predictor regression model. **Figures 38–42** plot against participants' judgments (Y axis), all five predictors – Has_CA (i.e., the new preemption predictor), Entrenchment, Bare-form rating, Reversibility and Semantics – all of

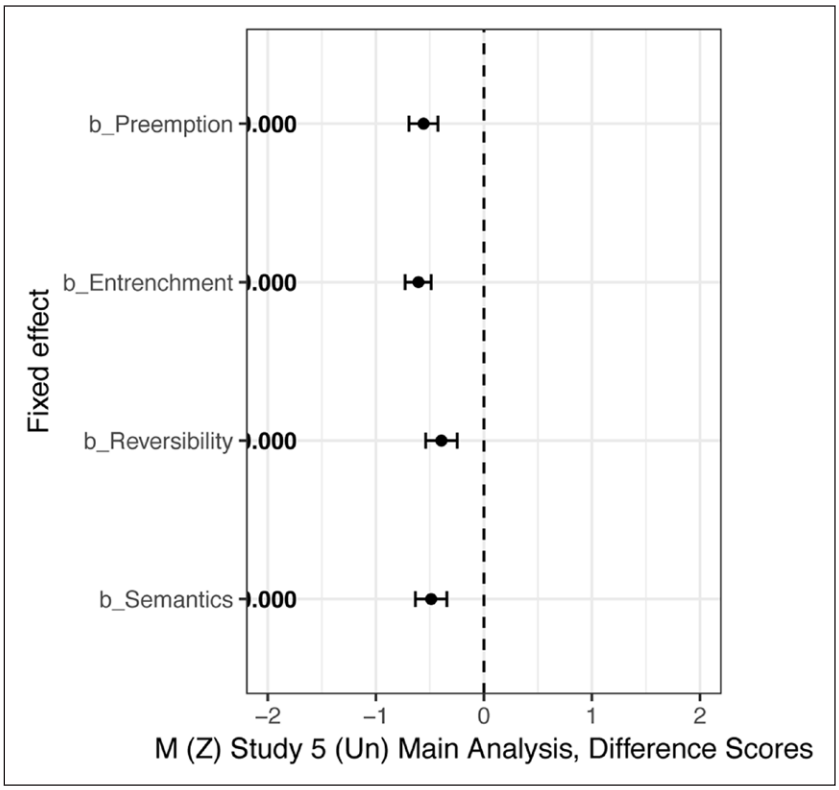

Figure 36: Study 5: New adult study of *un-* prefixation, nonpartial analysis of difference scores. Fixed effects (each from a separate regression model) for participants' difference scores (bare minus un- forms) and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores). However, see the main text for concerns regarding the interpretability of a nonpartial analysis for this particular dataset, and regarding the use of bare-minus-*un-* difference scores.

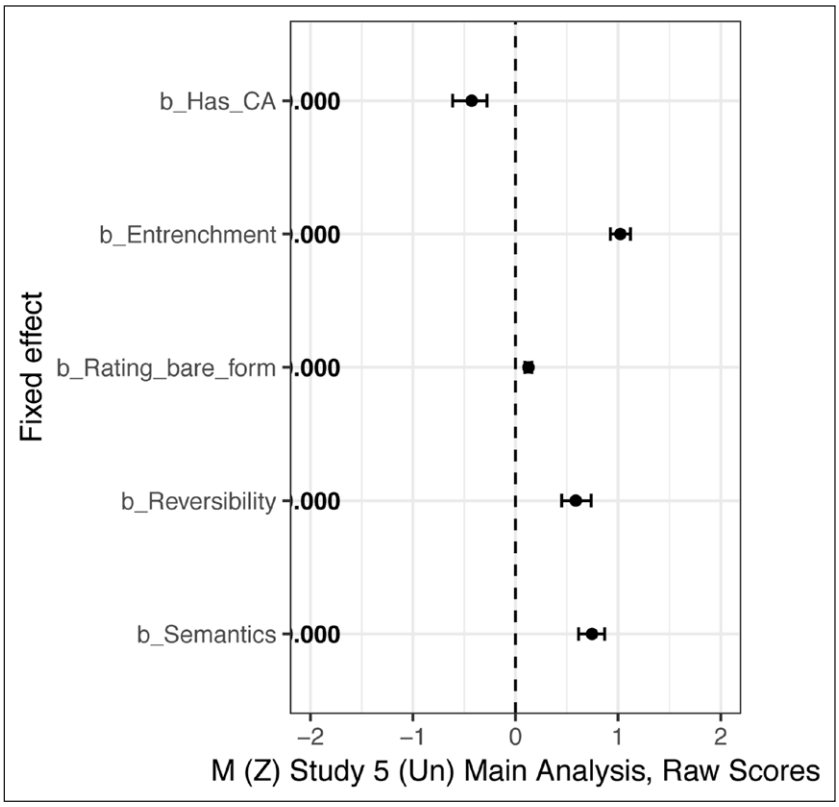

Figure 37: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Fixed effects (each from a separate regression model) for participants' judgments of *un-* forms, and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

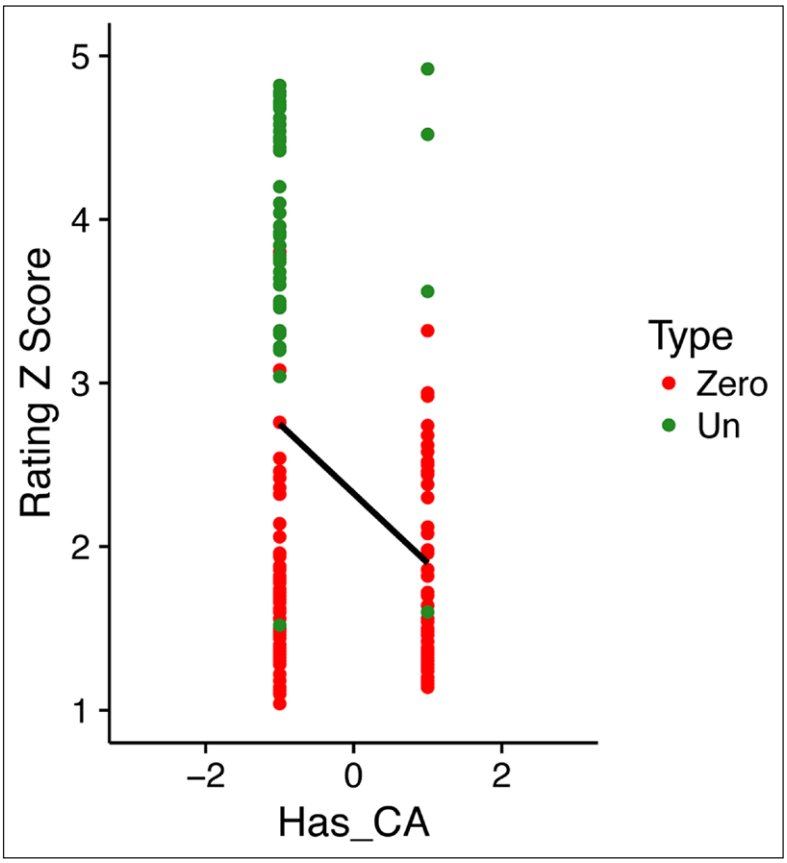

Figure 38: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Mean acceptability ratings for *un-* forms with (1) and without (−1) Competing Alternative forms (Has_CA).

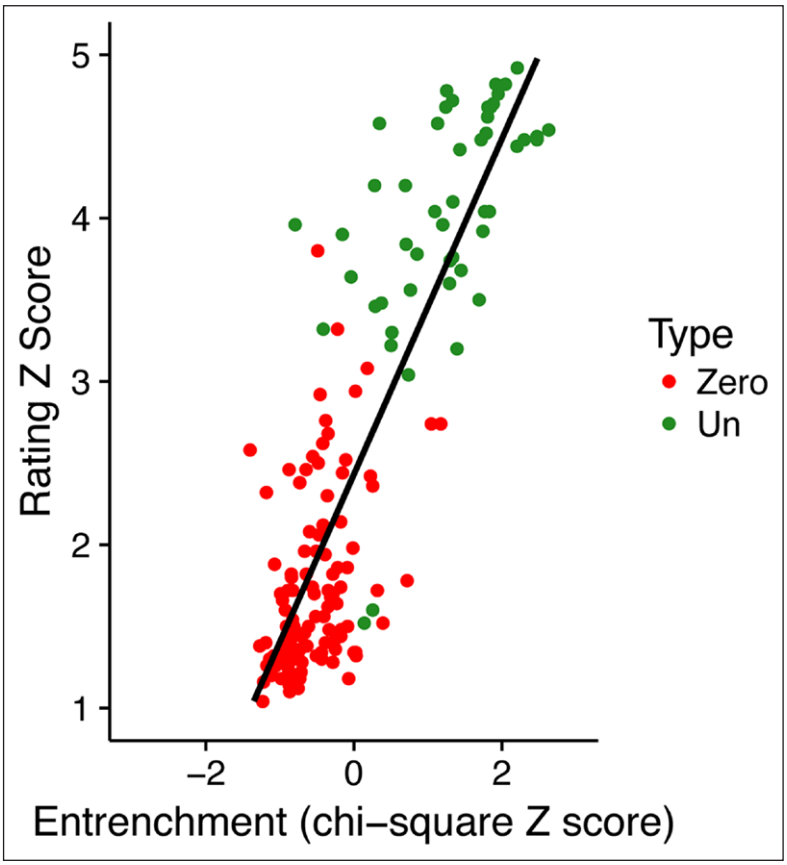

Figure 39: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Entrenchment predictor.

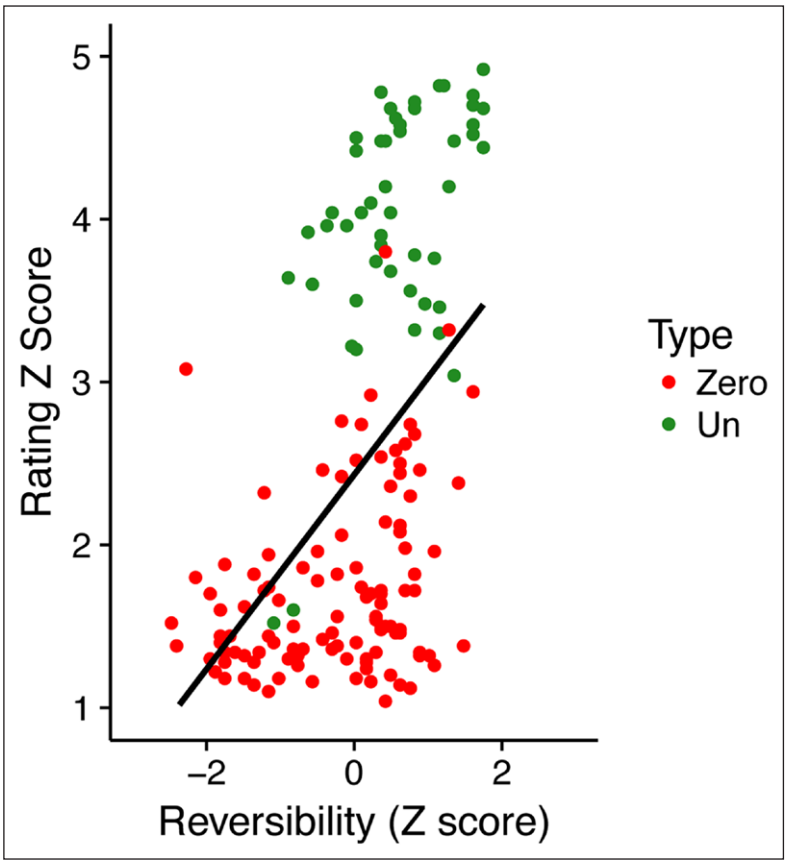

Figure 40: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Reversibility predictor.

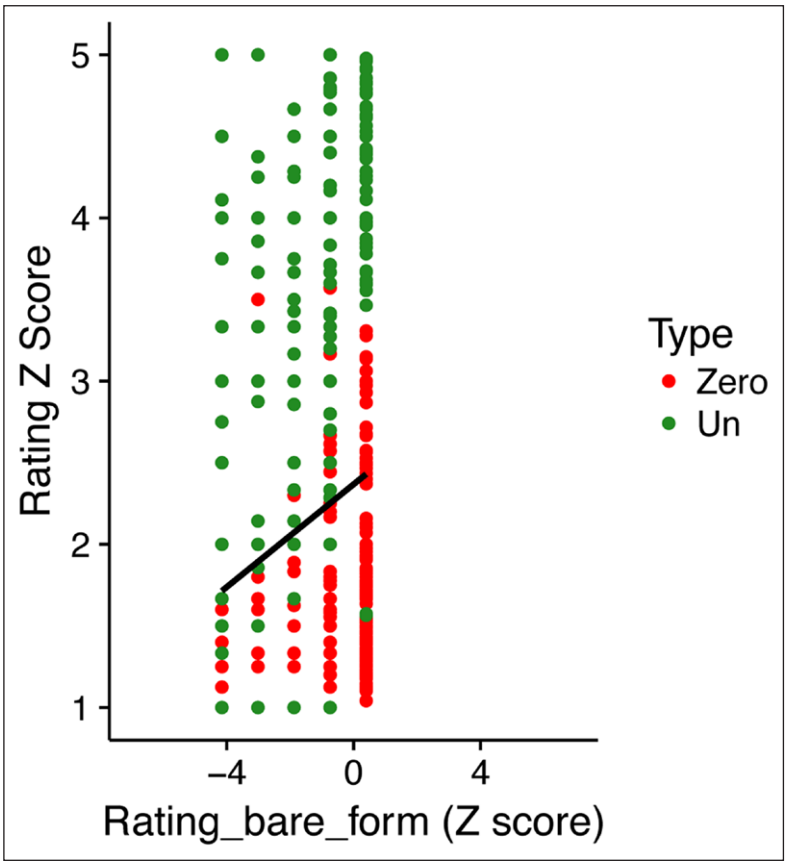

Figure 41: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Bare-form-rating predictor.

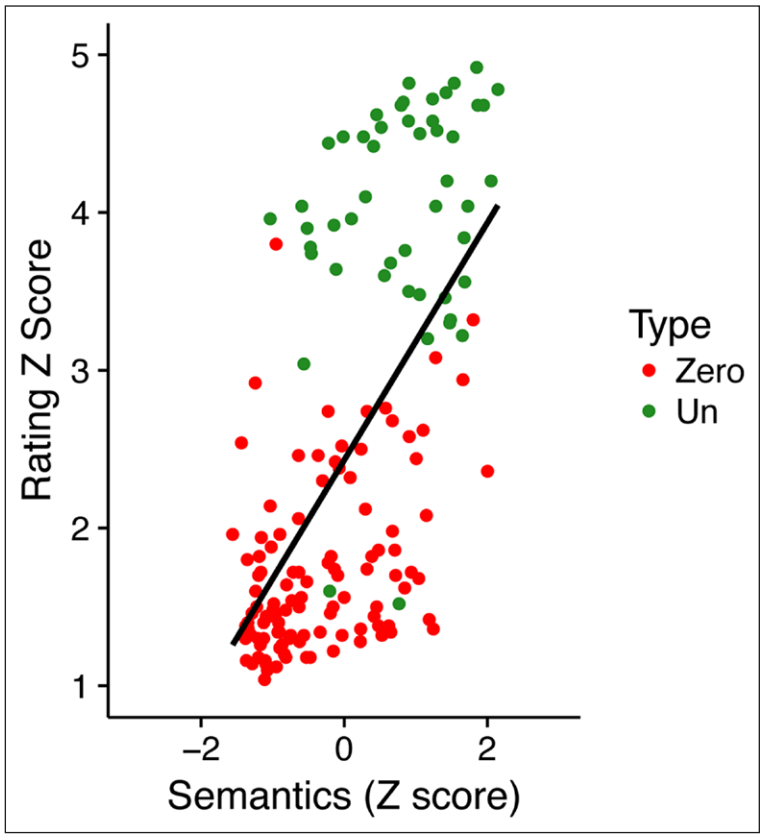

Figure 42: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Semantics predictor.

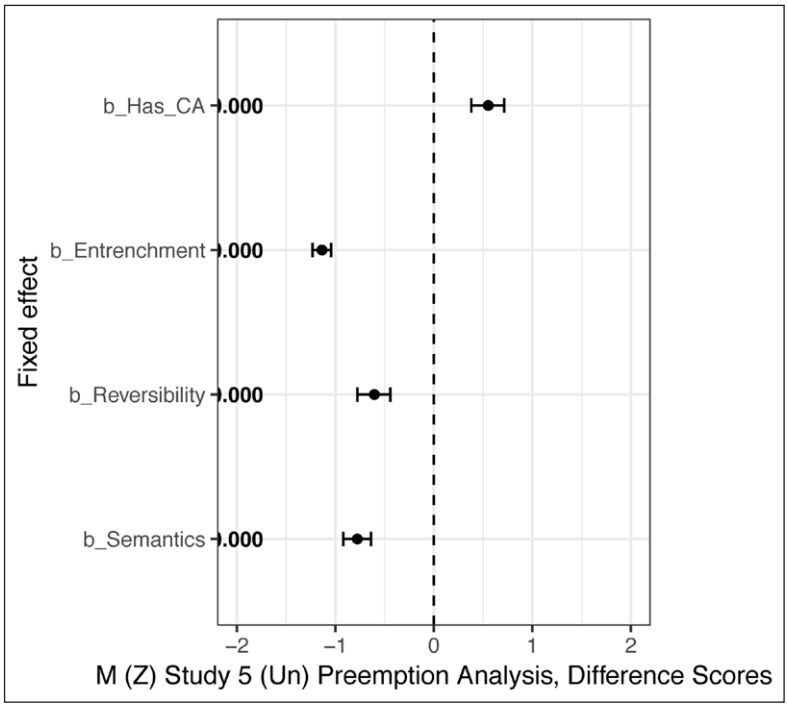

Figure 43: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Fixed effects (each from a separate regression model) for participants' difference scores (bare minus un- forms) and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

which had both CIs that did not overlap zero, and p_{MCMC} values of exactly zero. The model-comparison analysis (see Appendix Table A2) revealed that all five predictors explained unique variance, including – crucially – the has-CA preemption predictor. The difference-score analyses (**Figure 43** and Appendix Table A2) revealed

exactly the same pattern. **Figure 44** re-plots the has-CA preemption predictor from **Figure 38**, this time showing the individual verbs, as well as the mean ratings (+95% confidence intervals) for *un-* forms that do and do not have a potentially-preempting competing alternative. While *un-* forms that lack a competing alternating (shown in green) span the full range of acceptability ratings (as one would expect, given the importance of other predictors, such as semantics), *un-* forms that have a competing alternative (shown in red) – with only a handful of exceptions – receive acceptability ratings well below the mean. Thus, this analysis confirms from the main analysis an independent effect of preemption – or, in this case, at least of the existence versus nonexistence of potentially-preempting forms – above and beyond entrenchment (as well as semantics and all control predictors).

An additional test of the entrenchment hypothesis

The previous two analyses have already demonstrated an effect of entrenchment above and beyond preemption (and all other factors). However, perhaps the strongest possible test of the entrenchment hypothesis is whether this factor shows the predicted relationship with acceptability judgments looking only across *un-* forms that lack a potentially-preempting form: Because these *un-* forms lack a preempting form altogether, it would not be possible to argue away any observed entrenchment effect as a preemption effect "in disguise". That is, for the main analysis (§6.7.1), one could possibly argue that preemption and entrenchment are more highly correlated in the real world than our corpus-derived preemption and entrenchment predictors would suggest (after all, the corpus is only a rough approximation of the input language heard by our participants). As a result, the apparent independent effect of entrenchment observed in the main analysis might merely reflect a correlation with preemption which our measures are not capturing. Any entrenchment effect observed in the present, final analysis could not be observed in this way, since the analysis is restricted to verbs for which participants were unable to suggest a potentially preempting form (i.e., the 100 no-CA verbs). Thus, preemption was not included as a predictor, as per the following lme4 syntax:

> UnRating ~ (1 + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics | Participant) + (1|Verb) + Entrenchment + LogFreqUn + BareRating + Reversibility + Semantics)

For this verb set, correlations were again observed between entrenchment and reversibility (r = 0.58), entrenchment and semantics (r = 0.62) and reversibility and semantics (r = 0.55), again highlighting the importance of the model-comparison analysis.

Figure 45 shows, for this additional entrenchment analysis, the mean, 95% credible interval and (in bold) direction-corrected p_{MCMC} value for each single-predictor regression model. **Figures 46–49** plot against participants'

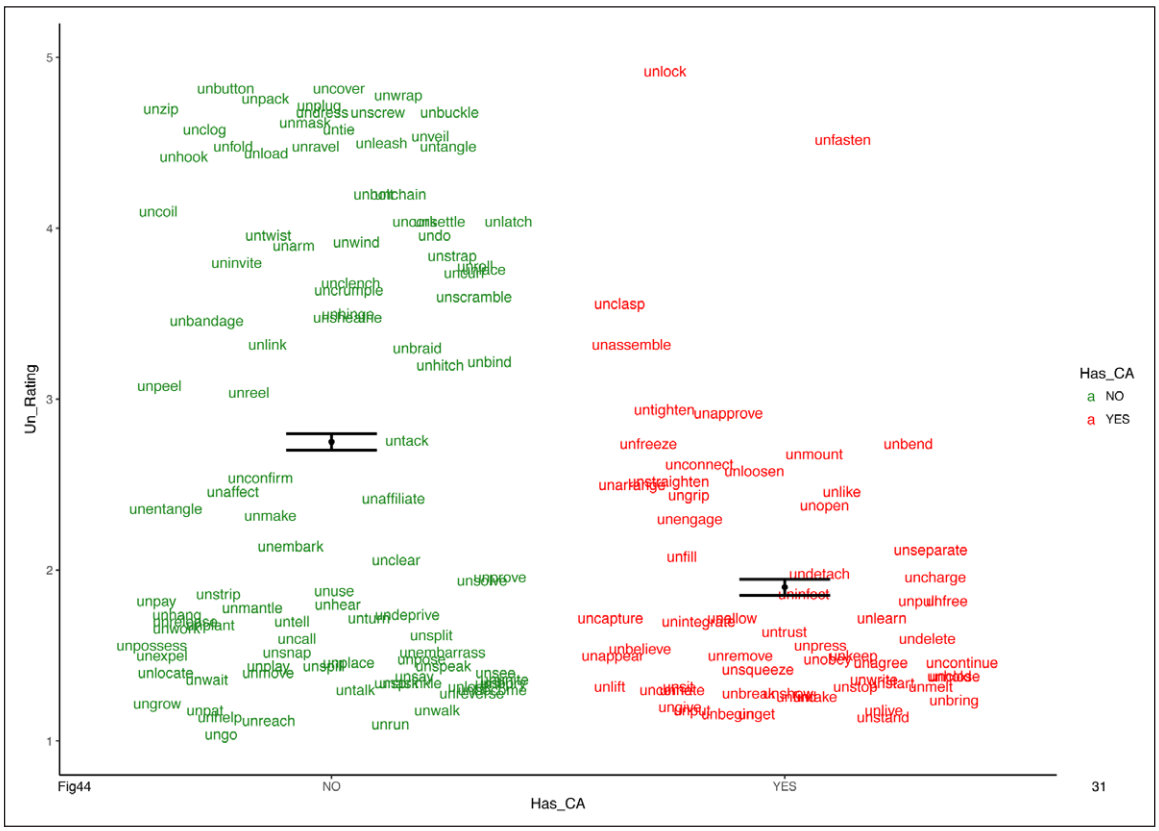

Figure 44: Study 5: New adult study of *un-* prefixation. Additional test of the preemption hypothesis. Mean acceptability ratings for *un-* forms with (YES) and without (NO) Competing Alternative forms (Has_CA). NB: These are the same data as Figure 38, but shown in an expanded format.

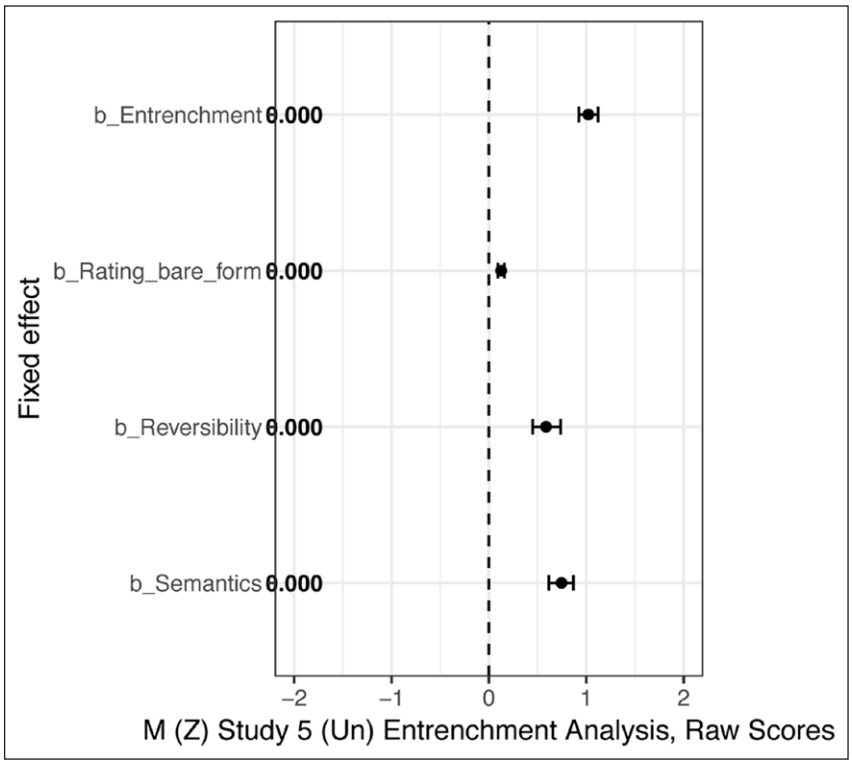

Figure 45: Study 5: New adult study of *un-* prefixation. Additional test of the entrenchment hypothesis. Fixed effects (each from a separate regression model) for participants' judgments of *un-* forms, and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

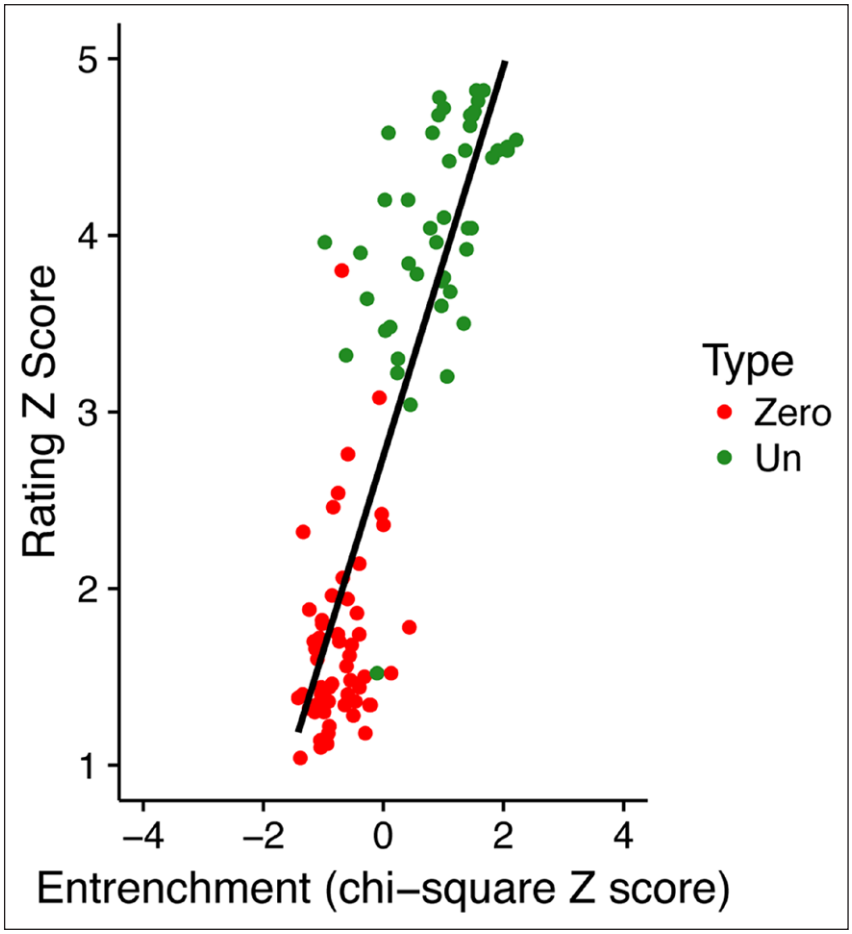

Figure 46: Study 5: New adult study of *un-* prefixation. Additional test of the entrenchment hypothesis. Entrenchment predictor.

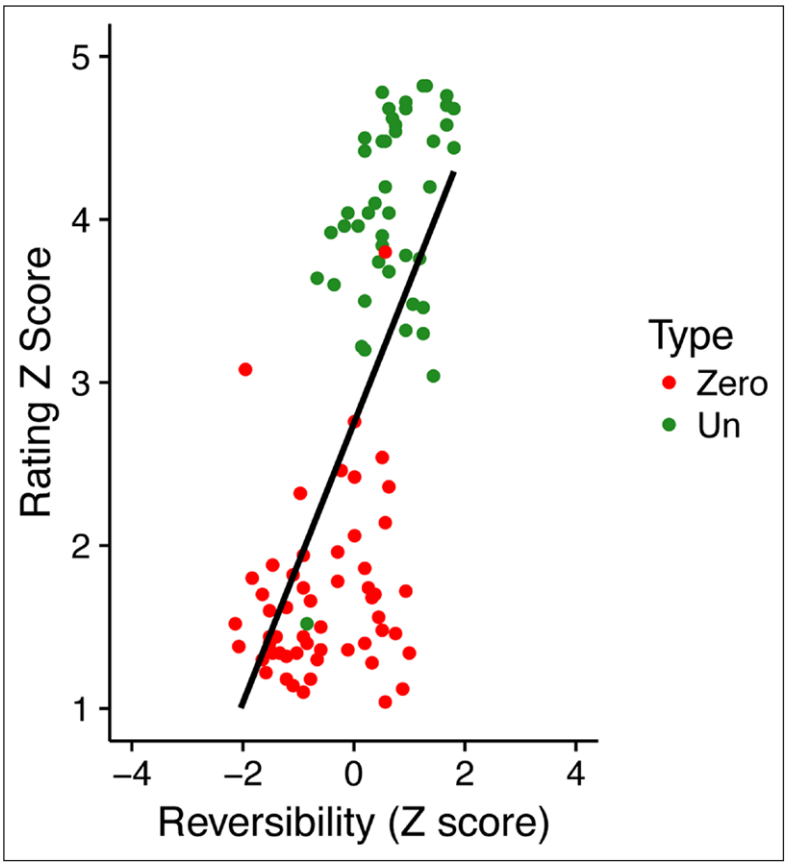

Figure 47: Study 5: New adult study of *un-* prefixation. Additional test of the entrenchment hypothesis. Reversibility predictor.

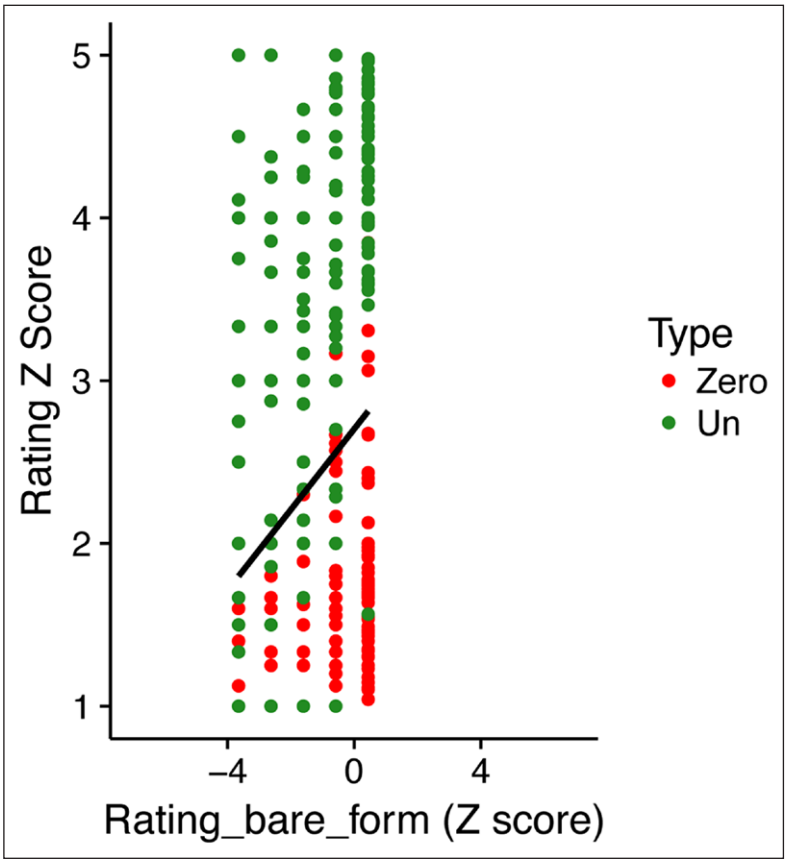

Figure 48: Study 5: New adult study of *un-* prefixation. Additional test of the entrenchment hypothesis. Bare-form-rating predictor.

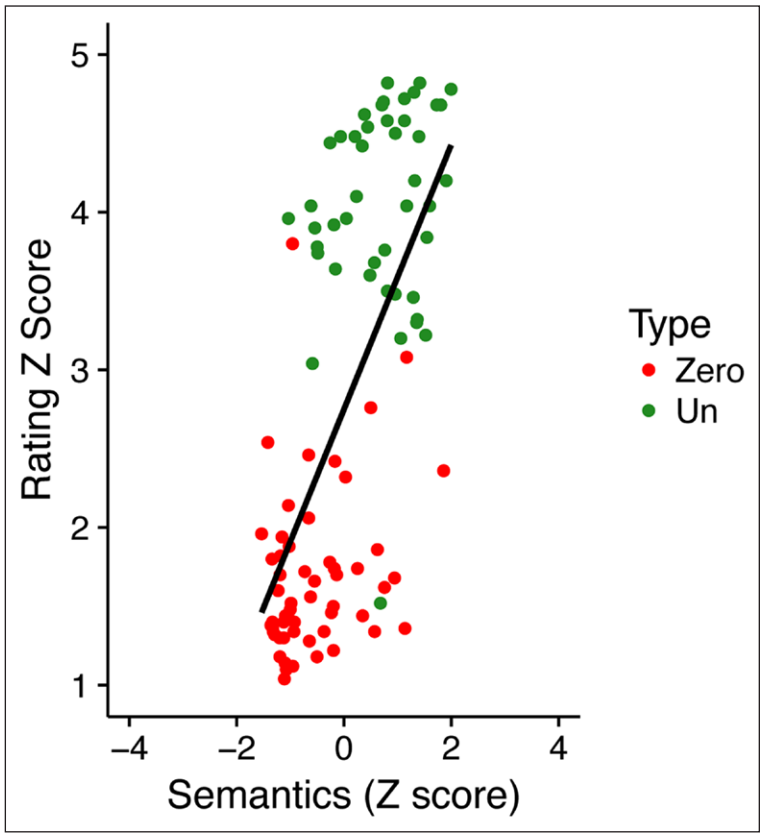

Figure 49: Study 5: New adult study of *un-* prefixation. Additional test of the entrenchment hypothesis. Semantics predictor.

judgments (Y axis), all four predictors –Entrenchment, Bare-form rating, Reversibility and Semantics – all of which had both CIs that did not overlap zero, and p_{MCMC} values of exactly zero. The model-comparison analysis (see Appendix Table A2) revealed that all four predictors explained unique variance, including – crucially – the entrenchment predictor. The difference-scores analysis (see **Figure 50** and Appendix Table A2) confirmed this pattern, including the significant effect of entrenchment by model comparison, except that the effect of semantics was no longer significant ($p = 0.09$). Because the *un-* forms included in this analysis were exclusively those for which participants could suggest no potentially-pre-empting competing alternative form, these findings constitute, in our view, the best evidence yet for an independent effect of entrenchment above and beyond preemption (as well as semantics and the control predictors).

Meta-analytic synthesis

Taken as a group, the studies reported above are inconclusive as to whether independent effects of preemption and entrenchment are observed. In order to answer this question, and to investigate whether any observed effect differs across age-groups and constructions, we conducted a meta-analytic synthesis[12] of the data from all five studies (for Study 5, the main analysis; not the additional analyses that have no equivalent in Studies 1–4). Semantic predictors were not included, as these are highly heterogeneous across studies (i.e., participants rated entirely different semantic properties in each).

Because the aim of the synthesis was to investigate whether an effect of preemption is observed *above and beyond entrenchment* – and vice versa – it would not have been appropriate to base our estimates of effect size on nonpartial correlations between each predictor and the dependent variable (as is conventional in meta-analysis, since – in general – most studies do not contain correlated predictor variables). We therefore used the commercial software package Comprehensive Meta Analysis (www.meta-analysis.com) to generate, as our measure of effect size, an *r* value based on the *p* value from the likelihood ratio tests reported for each study (i.e., the tests that compared a full model to a model without either (a) preemption or (b) entrenchment). Thus, these scores constitute a standardized measure of the effect of (a) preemption or (b) entrenchment, above and beyond all other predictors (including the various semantic and control predictors found across the studies). These *r* scores were converted into Fisher's *Z* values, and an estimate of variance calculated on the basis of the sample size (in participants), as per the following equations:

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

$$V_z = \frac{1}{N-3}$$

where *r* is the correlation coefficient and *N* is the number of participants (Schmidt & Hunter, 2015).

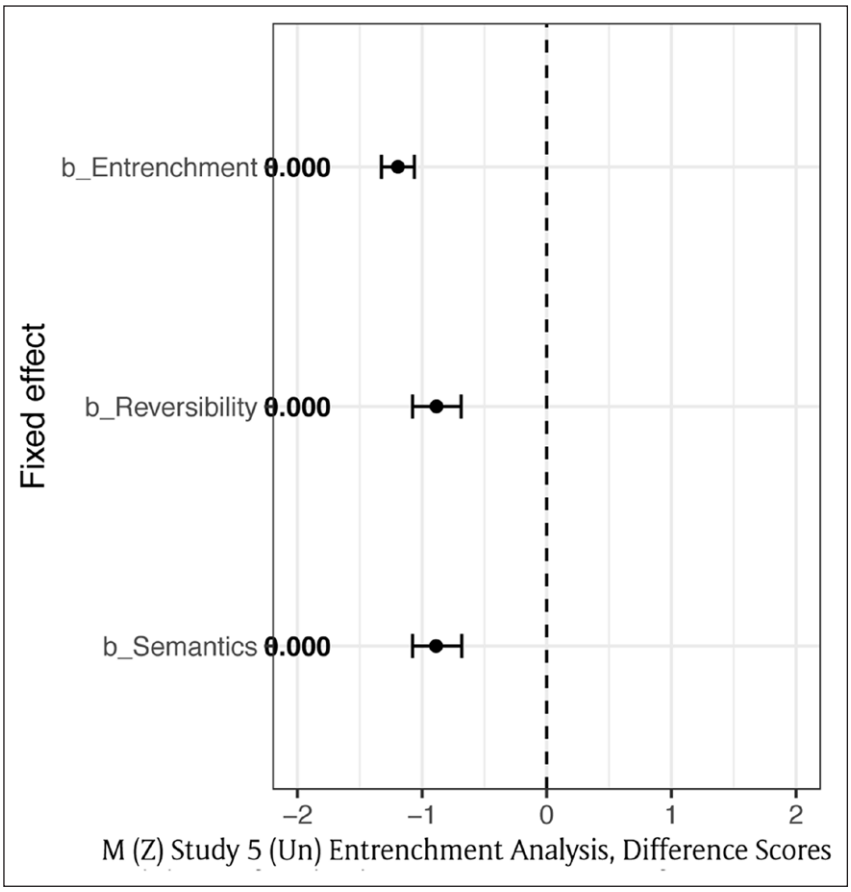

Figure 50: Study 5: New adult study of *un-* prefixation. Additional test of the entrenchment hypothesis. Fixed effects (each from a separate regression model) for participants' difference scores (bare minus un- forms) and accompanying P_{MCMC} values. Fixed effects are shown in standard deviation units (Z scores).

This procedure was used to generate an estimate of effect size (Z) and variance for each combination of – for the raw-score analyses – construction and age-group (19 independent estimates for each of preemption and entrenchment) and – for the difference-score analyses – study and age-group (13 independent estimates for each of preemption and entrenchment). For the various-constructions study (Study 3), "Various" was treated as a construction in its own right, as there were too few observations per construction to treat each of the eight constructions (see **Table 8**) as a separate construction with its own effect size. Due to the nested structure of our data (effect sizes within studies), we used multilevel meta-analysis. The random-effect was the study while the fixed effects were construction and age group included as potential moderators. Models were built for (a) Preemption, raw scores; (b) Entrenchment, raw scores; (c) Preemption, difference scores; (d) Entrenchment difference scores. All models were built using the metafor package (Viechtbauer, 2010) for R, as per the following example syntax:

```
Model_a = rma.mv(z_chi, z_var_chi, random = ~ 1
| Studyid, data = dat_pre)
```

The models are summarized as forest plots in **Figures 51–54**. In the raw-scores analysis (see **Figures 51–52**), both preemption (Z = 0.38, SE = 0.13, k = 19, p < 0.01) and entrenchment (Z = 0.23, SE = 0.04, k = 19, p < 0.0001) were significant. Examination of Q scores revealed that heterogeneity was a potential concern for the preemption model (Q = 78.34, p < 0.0001), but not the entrenchment model (Q = 18.95, p = 0.39, *n.s.*). Neither construction nor age group were found to be significant moderators of preemption (b = –0.05, p = 0.25, *n.s.*; b = 0.09, p = 0.06, *n.s.*) or entrenchment (b = –0.02, p = 0.31, *n.s.*; b = 0.05, p = 0.23, *n.s.*).

In the difference-scores analysis (see **Figures 53–54**), both preemption (Z = 0.33, SE = 0.06, k = 13, p < 0.0001) and entrenchment (Z = 0.29, SE = 0.07, k = 13, p < 0.0001) were again significant. Examination of Q scores revealed that heterogeneity was not a concern for either the preemption (Q = 16.66, p = 0.16, n.s.) or entrenchment model (Q = 10.26, p = 0.59, n.s.). Construction could not be investigated as a moderator, as the difference score combines scores from the two constructions rated within each study (except for Study 3, as explained above). Age group was found to be a significant moderator of preemption (b = 0.13, p = 0.01), but not entrenchment (b = 0.03, p = 0.60, n.s.). Inspection of **Figure 54** reveals that age moderates the effect of preemption, such that the magnitude of this effect increases with age. However, it is impossible to know whether this is because knowledge of preempting alternatives increases with age, or simply because children's judgment data are noisier than adults'.

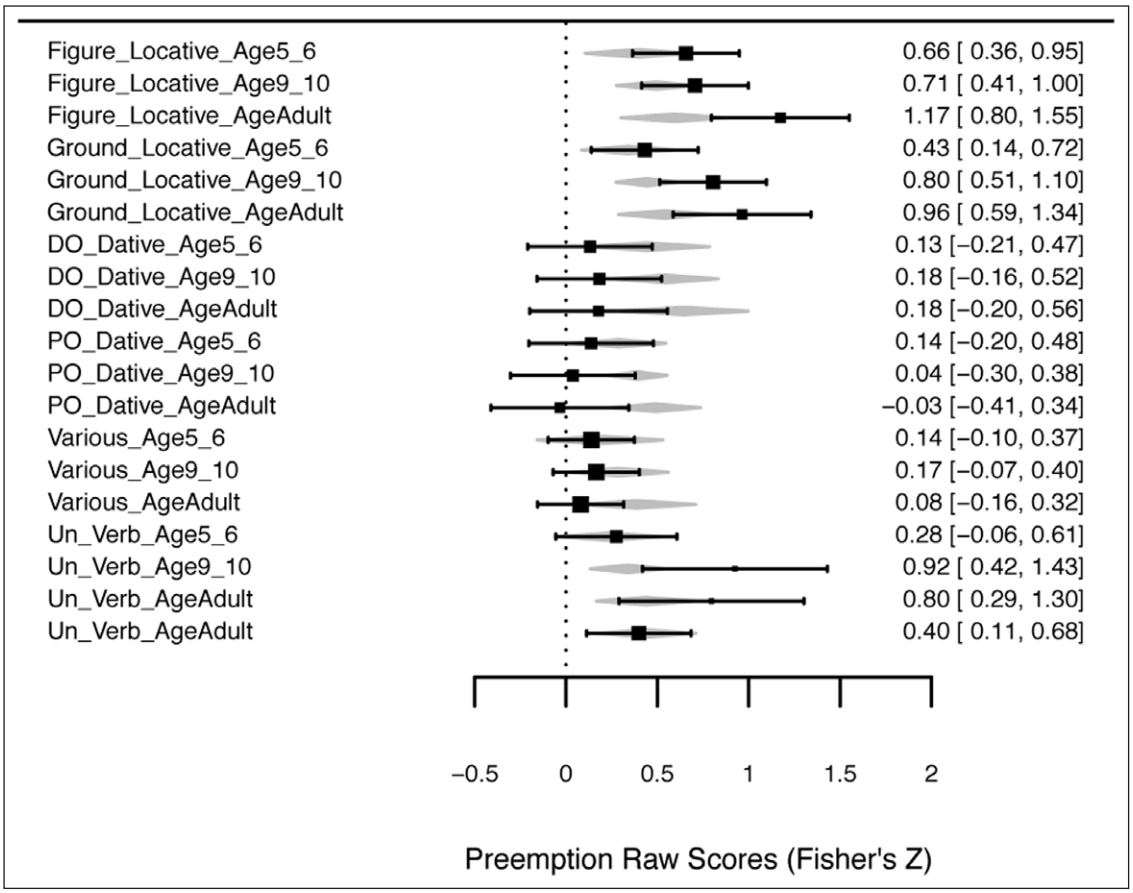

Figure 51: Meta-analytic synthesis for Preemption: Raw scores.

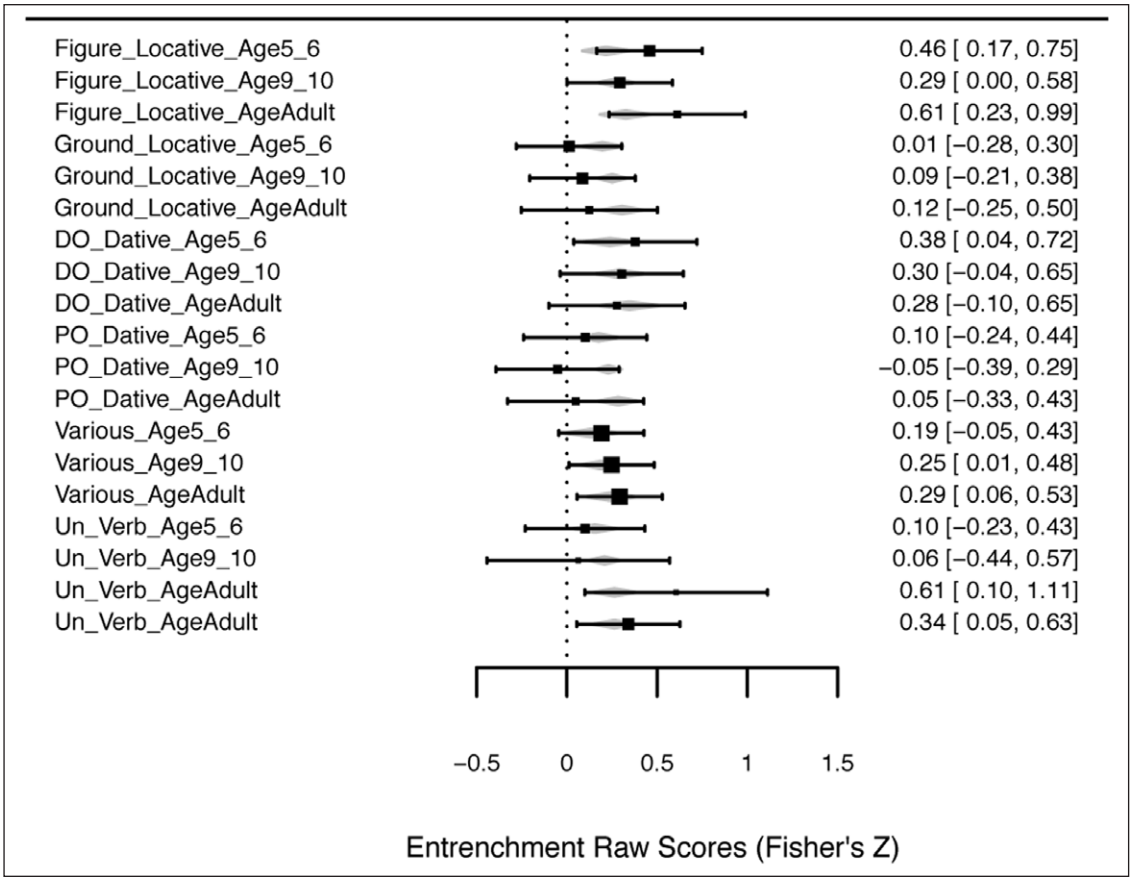

Figure 52: Meta-analytic synthesis for Entrenchment: Raw scores.

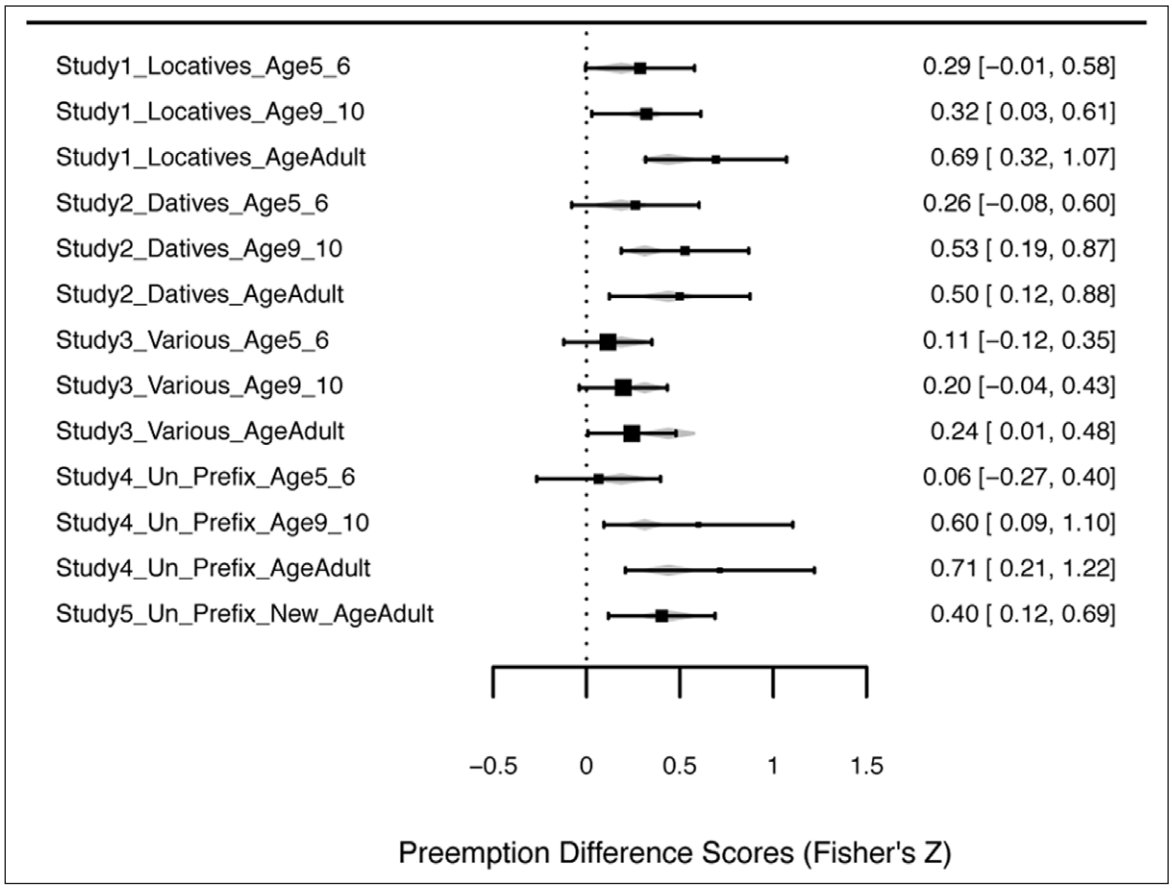

Figure 53: Meta-analytic synthesis for Preemption: Difference scores.

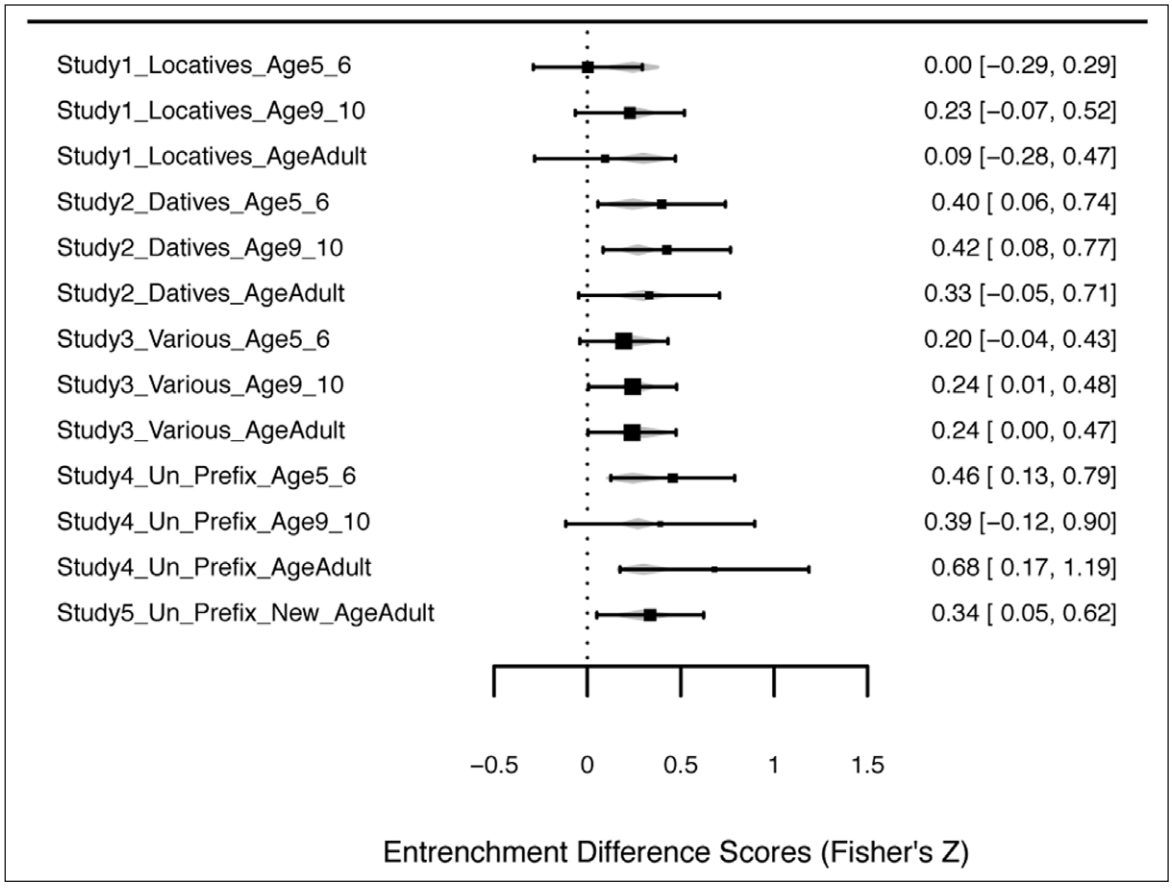

Figure 54: Meta-analytic synthesis for Entrenchment: Difference scores.

In summary, the meta-analytic synthesis of the five studies reported above revealed evidence for both preemption and entrenchment, independent of each other, and of the various semantic and control predictors found across the studies. Interestingly, the synthesis found no evidence that either effect varies according to the construction under investigation.

General Discussion

The question that the present studies, and the original studies reanalysed here, aimed to address was the following: Do speakers know that (for example) using *pour* in the ground locative construction (e.g., **She poured the glass with water*) is relatively unacceptable because:

(a) in situations where *pour* might have been used in the ground locative construction (given the speaker's intended meaning), it consistently appeared in the figure locative construction (e.g., *She poured water into the glass*) instead (***preemption***)? OR
(b) speakers have witnessed *pour* with high frequency regardless of construction, including in semantically-unrelated expressions such as *It's pouring outside*, leading them to implicitly conclude that if the ground locative use of *pour* were acceptable it would have been witnessed by now (above and beyond simply the odd "slip of the tongue") (***entrenchment***)?

The previous studies were inconclusive with regard to the relative contributions of preemption and entrenchment, and even whether one or both effects are observed. However, the reanalyses, new study, and meta-analytic synthesis reported here suggest that, when the two factors are carefully dissociated by means of model-comparison, both are observed. Although construction was not a significant moderator in the meta-analytic synthesis, conceptually, the distinct effects are particularly clear in the final two studies – both looking at *un-* prefixation – in which adults without exception showed effects of both preemption and entrenchment across a range of different verb sets and analysis types. Although the nonsignificant moderating effect of construction does not constitute positive evidence for no by-construction differences (Altman & Bland, 1995; Dienes, 2014), a reasonable default assumption (e.g., Croft & Cruse, 2004) is that the same restriction processes operate across all morphological constructions (here, *un-* prefixation) and verb argument structure constructions.

Indeed, if anything, *un-* prefixation constitutes a stronger test case for entrenchment than does verb argument structure. As noted above, ungrammatical *un-* forms such as **uncome* or **unsit* do not compete semantically with their bare forms (e.g., *come, sit*), unless they do so extremely indirectly (e.g., *he sat down, and then I told him* ***not to sit*** *there anymore*). Thus, it is difficult to explain, at least in terms of competition for meaning, why the availability of such bare forms (as measured by entrenchment) explains variance in the (un)acceptability of these un- forms (e.g., **uncome* or **unsit*). The fact that such an effect is observed suggests the existence of relatively "pure" form of entrenchment as an inference from absence, rather than an effect that occurs as a result of more indirect semantic competition.

Of course, all of the studies reported here were conducted on English, and it remains to be seen if comparable effects are observed for morphology and/or verb argument structure in other languages. If, in the meantime, we proceed on the tentative assumption that effects of both preemption and entrenchment are observed regardless of the particular language and construction under investigation, this raises the question of how this is to be explained theoretically. In particular, on the assumption that learners are not literally calculating chi-square statistics, a successful account is likely to be one that yields preemption and entrenchment as *effects* that fall naturally out of the learner's attempts to communicate meaning, rather than one that treats these effects as *mechanisms* in their own right. In the following section, we consider a number of current theoretical accounts that, potentially, have exactly this property; though it is important to note that the present findings do not provide a basis for deciding between them.

Theoretical Accounts

In order to explain the data presented in this and other papers, we need a theoretical account that can not only explain the present findings of effects of verb semantics, preemption and entrenchment, but can do so in a way that involves learning *graded* preferences, not simply that some uses are "ungrammatical". Three of the most promising approaches are outlined below.

FIT account

Under this account (see Ambridge & Blything, 2016, for a review), all constructions in the speakers' inventory (or, in practice, all that pass some threshold for *relevance*, as defined below) compete for the right to express the speakers' intended message, on the basis of four factors, illustrated here for the example message "MARGE CAUSED HOMER TO HAVE THE BOX BY PULLING THE BOX TO HOMER" (example adapted from Ambridge & Blything, 2016).

- **Verb-in-construction frequency.** The verb in the message (here *pull*) activates each construction in proportion to the frequency with which it has appeared in that construction in input sentences. This factor yields preemption effects because every input occurrence of pull in a PO-dative boosts the activation of this construction, at the expense of the DO-dative construction, in production. In principle, this factor can also yield "entrenchment" type effects (without an inference-from-absence entrenchment mechanism *per se*), simply because every input occurrence of *pull* in any other construction (e.g., a simple transitive such as *He pulled the string*) boosts the activation of this construction at the expense of the DO-dative.
- **Relevance.** A "relevant" construction is one that contains a semantically-suitable slot for every item in the speaker's message (such that, on a global level, the

semantics of the construction – e.g., transfer – match that of the speaker's message). So, for the present example, both the PO-dative (yielding *Marge pulled the box to Homer*) and the DO-dative *(*Marge pulled Homer the box*) are more relevant than, for example, the transitive (*Marge pulled the box*). The notion of relevance captures the intuition of the preemption hypothesis that the PO- and DO-dative are better competitors for one another than are other constructions such as the transitive. As we will see shortly, relevance is therefore crucial for simulating the primacy of preemption over entrenchment, as observed in the present study.

- **Fit.** The third factor is the compatibility (or fit) between the semantic properties of each item in the message (e.g., the verb) and the relevant slot in each candidate construction. The semantics of each slot are a frequency-weighted average of the semantics of each item which appeared in that position in the input utterances that gave rise to the construction. This factor is designed to capture the finding that ratings of the extent to which verbs exhibit semantic properties to do with "causing to have" and "causing to go" predict acceptability in the DO- and PO-dative respectively (Ambridge et al, 2014; and the present reanalysis, see **Tables 7–8**).
- A fourth factor, **overall construction frequency**, may also be important. That is, all else being equal, a speaker is more likely to select a higher frequency construction (e.g., an active transitive) than a lower frequency alternative (e.g., the passive). This factor may be necessary to explain by-construction differences.

Under this account, the extent to which previous experience of a verb in a particular construction "preempts" usage of the verb in a different construction is dependent on the degree of competition between the two constructions. On this view, what is traditionally classed as preemption is competition from witnessed forms that are *highly synonymous* with the overgeneralization error in question. This is likely universal across morphological and verb argument structure constructions alike (though, in practice, it may be difficult to detect when it is highly correlated with usage in other more distant constructions). However, effects of competition from more distantly competing forms – traditionally seen as "entrenchment" – will occur too, particularly when the relevant preempting form is of low frequency, but will generally be smaller and more sporadic.

A problem facing this account is that, as noted above, conceptualizing entrenchment as more distant semantic competition is probably a stretch too far in the case of *un-* prefixation; the construction for which entrenchment has been most unambiguously demonstrated. Given a particular error (e.g., *unsqueeze*), the form that entrenches away from this error (e.g., *squeeze*) is not really competing with the *un-* form for the same meaning (except possibly from very indirect deverbal formulations like *I stopped/reversed/undid the squeezing*).

A second problem facing this account is that, as a verbal model, it does not make quantitative predictions that can be tested experimentally. Ultimately, then, a successful account of this phenomenon will almost certainly have to take the form of a computational model. One preliminary attempt is the connectionist model of Ambridge and Blything (2016), which is based directly on the FIT account. However, although this model was able to simulate the pattern of human judgments found for the DO-dative (Ambridge et al, 2014), it failed to do so for the PO-dative.

Another limitation of this model is that, because it represents temporal order only at the utterance level, it cannot account for possible effects of temporal ambiguity as sentences are produced or comprehended word-by-word (or morpheme-by-morpheme) in real time. For example, for the dative constructions (§3.3), it is possible that real-time parsing could lead to competition from constructions that are not globally synonymous, but that are temporarily ambiguous with the target construction. That is, if constructions X and Y are temporarily ambiguous, greater experience of a verb in construction X might lead to initial misparsing when it is encountered in construction Y, yielding a garden-path effect which could contribute to a sense of ungrammaticality. This could be the case for DO datives, where frequently encountering a verb in the transitive (e.g. *carry* as in *He carried his brother*) might lead to an initial mis-parse when encountering this verb in a DO dative (e.g. *Lisa carried Marge…. the shopping*). Similarly, frequently encountering a verb in the intransitive (e.g. *giggle* as in *She giggled loudly*) could lead to a strong bias to parse this verb as preceded by an ACTOR rather than an AGENTIVE CAUSER, creating a garden path effect if is followed by a direct object (**Bart giggled…Marge*). Again, the greater the exposure to the verb in intransitive constructions, the greater the magnitude of this garden path effect, potentially reducing the grammaticality of the re-parsed sentence.).[13] Although this specific possibility is mere speculation on our part, the more general point that both comprehension and production are sensitive to temporal ambiguities that arise in real time (which are beyond the scope of Ambridge & Blything's, 2016, model) is well established in both adults and children (e.g., Frazier & Rayner, 1982; Trusewell et al. 1993; Phillips & Ehrenhofer, 2015).

CENCE ME account and Incremental Bayesian clustering

Goldberg (in press) sets out an account that is similar in many respects to the FIT account, but is both broader and more detailed, and places more emphasis on error-driven learning. Goldberg summarises the key principles of the CENCE ME (pronounced 'sense me') account as follows:

A) Speakers balance the need to be **E**xpressive and **E**fficient while obeying the **N**ormative conventions of their speech community.
B) Our **M**emory is vast; new information is related to old information. Representations are partially abstract (lossy).
C) Lossy memories are aligned when they share relevant aspects of form and function, resulting

in emergent clusters of representations: **C**onstructions

D) Multiple constructions are activated to the degree that they are suitable to express the intended message, and **C**ompete with one another for expression.

E) Mismatches between what is expected and what is witnessed fine-tune our network of learned constructions via **E**rror-driven learning.

Like the FIT account, the CENCE ME account, as a verbal model, does not yet make precise quantitative predictions that can be tested experimentally. Again, however, preliminary steps have been taken in this direction. Barak, Goldberg and Stevenson (2016) set out a Bayesian clustering model based on an older model (Alishahi & Stevenson, 2008), which itself simulates many aspects of the retreat from overgeneralization. This model was shown to be more successful than that of Ambridge and Blything (2016) in simulating the dative data from Ambridge et al (2014), yielding a significant correlation with all three measures of the dative alternation (DO-dative, PO dative, difference scores). This correlation is explained by the model's ability to capture gradient degrees of relevance and fit of semantic-syntactic pairings together with the distributional properties of the verbs. Unlike the connectionist model that predicts a syntactic choice based on the value of each semantic dimension, the Bayesian model creates multiple clusters which represent an association of related semantic vectors (across multiple dimensions) with a syntactic pattern. Importantly, the analysis of Barak et al (2016) is also in line with our current findings on the role of semantic properties in this learning process. However, this analysis suggests the need for additional semantic properties beyond those used in the current studies, in order to fully capture the factors of relevance and fit across different classes of verbs.

This model shares a shortcoming with that of Ambridge and Blything (2016). Because it does not produce sentences in word-by-word fashion, it cannot explain possible effects of temporal ambiguity that occur in real-time sentence production or processing. Another shortcoming is that, like the fit account, it struggles to explain the observed effects of entrenchment in the domain of *un-* prefixation, which do not seem to rely on semantic competition or clustering.

Discriminative Learning

A third possible account is based on discriminative learning;[14] a concept that originates in the animal learning literature. The key feature of discriminative-learning models is that learning is a process by which prediction error is used to discriminate uninformative versus informative cues. Thus, such models weight cue strength from both *cue-outcome* pairings that are observed, and *cue-outcome* pairings that are predicted, but not observed. For example, suppose that rat learns to associate a tone (cue) with a shock (outcome), and so freezes in anticipation of a shock whenever the tone is heard. In an otherwise-identical setup with additional tones that are not followed by a shock, learning is attenuated. Indeed, the likelihood of the rat freezing in response to the done decreases in proportion to the *background rate* of tones that are not followed by a shock (Rescorla, 1968). Discriminative-learning models are also designed to explain behaviour in situations with multiple cues. For example, if a rat has already learned to associate a tone (cue) with a shock (outcome), its ability to learn that another cue (say a buzzer) also predicts this shock is reduced; a phenomenon known as *blocking* (Kamin, 1969; though see Maes et al., 2016, for 15 failures to replicate). Cues can combine as well as compete. For example, a rat can learn that a buzzer predicts a shock, but only if a tone is also present; or – alternatively – only if a tone is absent (phenomena known respectively as *positive-* and *negative occasion-setting*; Holland, 1983). More generally, Bellingham, Gillette-Bellingham and Kehoe (1985) demonstrated that multiple cues can combine in a nonlinear fashion.

A number of different discriminative-learning algorithms have been proposed, but all share the assumption of learning via prediction error, a characteristic also displayed by connectionist models (indeed, in the limit, the two are formally equivalent). If a cue predicts that an outcome will occur, and it does so, the association between the cue and the outcome is boosted. Crucially, if a cue predicts that an outcome will occur, and it does *not*, the association between the cue and that outcome is weakened. In this way, the model provides exactly the type of negative evidence that would be useful for language learners (but that they are usually assumed to lack; e.g., Bowerman, 1988; Pinker, 1989). In the domain of language acquisition, discriminative learning has usually been formalized using the Rescorla-Wager (1972) learning rule: an algorithm that can model animals' behaviour in the learning scenarios outlined above (and many others). In the discussion below, we therefore focus on this rule as an example of a discriminative-learning algorithm that can potentially explain the current findings. However, we are not claiming that particular properties of the Rescorla-Wagner rule make it uniquely well-suited to the problem under investigation here; alternative discriminative-learning algorithms would likely fare similarly.

In a series of studies with children, Ramscar and colleagues have demonstrated that the Rescorla-Wagner learning rule provides an excellent fit to language learning in a number of different domains, including word-learning and morpho-syntax.[15] Most relevant for the present work, Ramscar, Dye and McCauley (2013) investigated English-speaking children's noun plural *–s* over-regularization errors (e.g., **mouses*). The learning situation was formalized as a task in which children learn the predictive value of real-world semantic cues (e.g., *multiple items, single item, multiple mouse items, single mouse item, mousiness, stuff*) for particular linguistic outcomes or events (e.g., *dog+s, dog+0, mouse+0, mice+0, mouse+s*). Key to the model's success is its use of error-based learning: When the semantic cues (e.g., *mousiness, multiple items*), as instantiated by a picture of several mice, are strongly predictive of an overgeneralized form (e.g., **mouse+s*), the violation of this expectation (i.e., encountering *mice*)

is highly informative. Thus, when trained on a realistic distribution of input forms, the model – exactly like children – initially produces overgeneralization errors (e.g., because *mousiness* and *multiple items* are strongly predictive of **mouse+s*). Later in development, when learning of the predictive value of *multiple items* for *+s* has reached asymptote, the model – exactly like children – continues to learn that *multiple mouse items* (and the combination of *mousiness* and *multiple items*) is predictive of *mice*, and errors cease.

In addition to simulating overgeneralization-then-retreat (and, indeed, U-shaped learning), the Rescorla-Wagner model makes a counterintuitive prediction. Early in development, presentation of regular plurals (e.g., *dog+s*) will increase the rate of overregularization (e.g., **mouse+s*), by boosting the predictive value of *multiple items* for +s. However, later in development, when this association has reached asymptote, presentation of regular plurals (e.g., *dog+s*) will decrease the rate of overregularization (e.g., **mouse+s*), by boosting the predictive value of *multiple items*, **in the absence of *mousiness or multiple mouse items***, for *+s*. This prediction was confirmed in an elicited-production training study with children (Ramscar et al, 2013; see also Ramscar & Yarlett, 2007).

An advantage of the Rescorla-Wagner model (or a similar discriminative-learning model) is that it is both more precise than verbal models such as the FIT and CENCE ME accounts, and simpler than the computational models of Ambridge and Blything (2016) or Barak et al (2016). Another advantage is the considerable support that such models already enjoy in the domains of both child language acquisition and human and animal learning more generally (e.g., Rescorla, 1988; Gureckis & Love, 2010; Arnon & Ramscar, 2012). Indeed, its grounding in the human and animal learning literature renders discriminative learning psychologically plausible as a *model of human language learning*. In this respect, it contrasts with formal Bayesian rational-learner models that can also explain entrenchment and preemption effects as types of inference from absence, but that operate at a higher level of abstraction (e.g., Hahn & Oaksford, 2008; Hsu & Chater, 2010; Perfors, Tenenbaum & Wonnacott, 2010).

In principle, a discrimination-learning algorithm such as the Rescorla-Wagner model can be applied to the present domain in much the same way as it was applied (by Ramscar et al, 2013) to the domain of plural –s overgeneralization (see **Table 12**). Consider, for example, overgeneralization errors (or at least dispreferred forms) in which *drag* is used in a *DO-dative* construction (e.g., **Marge dragged Homer the box*). The learning situation can again be formalized such that children learn the predictive value of real-world semantic cues (e.g., *transfer event, nontransfer event, transferring by dragging, dragging but not transferring, dragging*) for particular linguistic outcomes: the occurrence of a verb in a particular construction (*drag+PO dative, drag+nondative, give+DO dative, drag+DO dative*). Early in development, the model predicts overgeneralization errors, because the cues

Table 12: Discrimination learning (e.g., the Rescorla-Wagner leaning model), as applied to overgeneralization errors in the domain of English noun plural formation (Ramscar & Yarlett, 2007; Ramscar, Dye & McCauley, 2013), and English datives (the present Study 2).

Unconditioned Stimulus (UCS) or Cue	**Conditioned Stimulus (CS) or Outcome**			
	Correct Target	**Overgeneralized form**	**Target lexical item in non-target form**	**Form that undergoes relevant generalization**
	Mice+0	***Mouse+s***	***Mouse+0***	***Dog+s***
Multiple items	**1**	1	0	1
Single item	**0**	0	1	0
Multiple mouse items	**1**	1	0	0
Single mouse item	**0**	0	1	0
Mousiness	**1**	1	1	0
Stuff	**1**	1	1	1
Unconditioned Stimulus (UCS) or Cue	**Conditioned Stimulus (CS) or Outcome**			
	drag+PO dative	***drag+DO dative***	***drag+nondative***	***give+DO dative***
Transfer event	**1**	1	0	1
Nontransfer event	**0**	0	1	0
Transferring by dragging	**1**	1	0	0
Dragging but not transferring	**0**	0	1	0
Dragging	**1**	1	1	0
Event	**1**	1	1	1

transfer event and *dragging* are independently strongly predictive of *drag+DO dative*. Later in development, when learning of the predictive value of *transfer event* for *DO-dative* has reached asymptote, the model continues to learn that the combination of *transfer event* and *dragging* (and *transferring by dragging*) is predictive of *drag+PO dative*, and errors cease.

Importantly, this model subsumes preemption and entrenchment: Preemption occurs when the semantic cues (e.g., *dragging, transfer event*) are strongly predictive of *drag+DO dative*, but this expectation is violated, and *drag+PO dative* occurs instead. Entrenchment occurs when the semantic cues are strongly (e.g., *dragging, transfer event*) or more weakly (e.g., *dragging*) predictive of *drag+DO dative*, but this expectation is violated, and – for example – *drag+nondative* occurs instead. Thus, entrenchment effects reflect the background rate at which a particular event (e.g., *dragging*) occurs without the outcome of interest (e.g., *drag+DO dative*). In this scenario, the entrenchment effects observed in the present study arise from the fact that the corpus frequency of, for example, *drag*, serves as a proxy for the frequency of *dragging* events, with the latter the cue that is predictive of the linguistic outcome *drag+PO* (but not *drag+DO*). This parallels the finding that the learning that (for example) a tone predicts a shock is attenuated if the tone is additionally presented without a shock on some trials, thus increasing its background rate (Rescorla, 1968).

Furthermore, rather than being treated as a different kind of approach altogether, the semantic factors evidenced in the present study could straightforwardly be instantiated in a more detailed discriminative-learning model. For example, the cue of *dragging* could be replaced by the five verb-level (morpho)semantic factors from Study 2 (*Speech, Mailing, Bequeathing, Motion* and *Latinate*), with the cue strength of each determined on the basis of the semantic ratings produced by participants in the original study. Similarly, the cue of transfer event could be replaced by the three subtly different types of *transfer event* (referred to in Study 2 as PO-dative semantics and DO-dative semantics 1 and 2). This would yield a model that is similar in some ways to the connectionist model of Ambridge & Blything (2016), but that is simpler and more transparent. It is also possible to build versions of a discriminative-learning model that learn about temporally presented sequences of cues and outcomes (e.g., Gureckis & Love, 2010, for the Rescorla-Wagner model) and that therefore could potentially explain any effects of real-time processing subsequently observed in this domain.

Finally, if set up in just the right way, a discriminative learning model – unlike the FIT and CENCE-ME accounts – naturally explains the effect of entrenchment in the domain of *un-* prefixation. For example, the combination of the semantic cues *squeezing+reversal* is strongly predictive of the form **unsqueeze*. So, when – in the presence of this semantic-cue combination – this predicted form fails to occur, the predictive value of *squeezing+reversal* for **unsqueeze* is reduced, *even if nothing appears in its place*. Again, this parallels the finding that the learning of a tone → shock pairing is attenuated if the tone is additionally presented without a shock on some trials, thus increasing its background rate (Rescorla, 1968).

In our view, then, a discriminative learning model along the lines of the Rescorla-Wagner model would seem to hold the greatest promise of a truly unitary account of learners' acquisition of verbs' restrictions; and one that is well-grounded in the human and animal learning literature.

Methodological considerations

Although we have focussed here on the theoretical contributions made by the present paper, we wish to highlight a number of important methodological considerations that, in our view, could be profitably applied to many different types of study, particularly those that use corpus data to derive predictions to be tested experimentally. First, the method used for **obtaining corpus counts** – combining automatic extraction and hand-coding – should prove useful for any study that requires counts of verbs in particular constructions (whether experimental, or entirely corpus based). Second, the chi-square method used across all studies for operationalizing entrenchment and preemption as **measures of contingency** (based on Stefanowitsch, 2008) is applicable to any study that requires a measure of the relative strength of competing forms (e.g., past versus non-past verb forms; Tatsumi, Ambridge, & Pine, 2017) that is sensitive to **both proportional and absolute frequency. Third, our use of both nonpartial regression models and model comparison** proved crucial for discovering whether particular predictors show the predicted relationship with the dependent measure, and whether the effect holds when controlling for other predictors. We urge our colleagues to learn from a mistake made in many of the original papers (and earlier versions of the present paper), and to check for collinearity between predictor variables before running simultaneous regression models. Our use of a **Bayesian statistical models** (for the single-predictor analyses) allowed us not only to build conservative maximal models (Barr et al., 2013) without convergence failure, but also to present *p(MCMC)* values and credible intervals that – unlike frequentist *p* values and confidence intervals – yield intuitive interpretations. McElreath's (2016) *rethinking* R package played a crucial role here in translating data and syntax formulated for lme4 (Bates et al, 2015) into a format suitable for Bayesian analysis, with very little effort or knowledge required on our part. We are also late converts to the brms package (Bürkner, 2016) which we used to explore – though ultimately reject as too computationally intensive – a Bayesian model-comparison procedure based on leave-one-out cross validation (Vehtari, Gelman & Gabry, 2017). Finally, our use of **meta-analytic techniques** was useful for confirming that – when taken together – the studies reviewed here strongly suggest evidence for both preemption and entrenchment.

Conclusion

In conclusion, although disentangling entrenchment and preemption remains difficult, the present findings suggest that – to the extent that this is possible for English – both

effects are observed. We therefore hope that the new approaches outlined here to operationalizing and testing statistically for effects of these variables will provide a firmer methodological grounding for future work that aims either to disentangle these factors or – better still – to describe a unitary learning mechanism that yields these effects, and that is psychologically plausible.

Data Accessibility Statement

All data and analysis scripts are available at https://osf.io/7rvsq/.

Additional Files

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

- **Appendices.** Doi: https://doi.org/10.1525/collabra.133.s1

Notes

[1] We owe the reader a definition of *construction*, a term that we use frequently throughout this paper. Essentially, we adopt Goldberg's (1995: 4) definition that a construction is a "form–meaning pair such that some aspect of the form or some aspect of the function is not strictly predictable from the construction's component parts, or from other previously established constructions". The verb-argument structure (Studies 1–3) and morphological constructions (Studies 4–5) investigated here meet this definition because each of these constructions – patterns of abstract slots such as [NP] [VERB] [NP] [NP] – contributes some meaning in and of itself. For example, the DO-dative construction [NP] [VERB] [NP] [NP] investigated in Study 2 contributes to the utterance a meaning of literal or metaphorical transfer (e.g., *She told him a story*) that is not strictly predictable from the meaning of any of these individual words, or other constructions (e.g., [NP] [VERB] [NP], as in *She read the story*).

[2] We take these constructions to be instances of higher level, more general constructions: the *caused-motion* and *causative* constructions respectively (e.g., *Bart sent a parcel to Lisa; Bart broke the plate with the hammer*). However, we retain the terms *figure-* and *ground-locative*, partly because these were the terms used in the original studies, but more importantly because it is these lower-level, more specific constructions that are used when estimating the total number of uses of each construction found in the relevant corpus (see Study 1 Methods section).

[3] We follow here the linguistic convention of indicating "ungrammatical" utterances with an asterisk (*). However, a major finding of the present studies (and the original studies whose data we reanalyse) is that acceptability is graded rather than binary; an important point that our informal use of asterisks is not intended to obscure. The presence or absence of an asterisk was determined either on the basis of the intuitions of the authors whose work we consulted when creating the original stimulus sets (Pinker, 1989; Levin, 1993) or – in the case of *un-*prefixation (Studies 4–5), the absence or presence of the *un-* form in the *British National Corpus*. Thus, an asterisk is intended as nothing more than an *aide-memoire*, indicating a hunch that the relevant form is probably at least somewhat less than fully acceptable. But nothing hinges on these classifications, which were not used as a factor in any statistical analysis. On our view, the grammatical acceptability of a particular form can be determined only in a large-scale rating study of the type reanalysed here (and, for what it is worth, these data largely confirm our hunches). Even then, each rating is specific to a particular context (i.e., as a description of the particular picture sequence or animation with which it was paired in the study). For example, as *a previous* Action Editor (Max Coltheart) pointed out, **unsqueeze* is relatively acceptable in the context of reformatting a video image from 4:3 to 16:9 aspect ratio. But it is relatively unacceptable in the context in which it was rated in the present Studies 4–5 (*Lisa squeezed the sponge and then she unsqueezed it*, which received a mean adult rating of 1.5/5). This may well be part of a broader phenomenon whereby less-than-fully-grammatical forms are used knowingly for jocular, poetic or other special *effect* (e.g., Pinker, 1989; Goldberg's (in press) book *Explain me this*). For example, **unbreak* (though not amongst the forms rated in Study 4) sounds to us relatively unacceptable in a mundane sentence such as *Lisa broke the cup and then she unbroke it*, but is used to poetic effect in the Toni Braxton song *Un-break my heart*. The hyphen in the title suggest that the song's writer, Diane Warren, did not consider *unbreak* to be an everyday English word like *unchain* (the title of another of her songs).

[4] We thank an anonymous reviewer for this suggestion.

[5] Evidence for preemption, though not in every case over-and-above entrenchment, comes from other corpus-based work (Goldberg 2011; Robenalt & Goldberg 2015; 2016), and elicited-production studies (e.g., Brooks & Tomasello 1999; Brooks & Zizak, 2002; Brooks, et al., 1999; Boyd & Goldberg 2011; Perek & Goldberg, 2017. Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition, 168*, 276–293).

[6] Note that this problem does not arise for the *un-*prefixation studies of Ambridge (2013) and Blything et al (2014). For example, for a target ungrammatical form such as **unsqueeze*, the preemption measure (here, the frequency of *release* and *loosen*) and the entrenchment measure (here, the frequency of *squeeze*) are independent, at least in principle.

[7] This problem could be avoided (see Robenalt & Goldberg, 2015) by deliberately selecting verbs of high *overall* frequency (e.g., *sneeze, laugh*), but relatively low frequency in the relevant constructions (e.g., the locative constructions); a point to which we return in the General Discussion. However, the current goal is to reanalyze existing data sets (which did not contain such verbs), using more appropriate measures and analyses.

[8] That is, the preemption measure is (as predicted) negatively related to judgments in a single-predictor model, but flips sign in a two-predictor model. A tempting, but incorrect, interpretation of this pattern is that having taken into account the relationship between a verb's overall frequency (entrenchment) and participants' judgments, occurring in a particular construction (preemption) actually seems to *increase* the acceptability of uses in the "pre-empting" construction (opposite to the prediction of this account). However, this interpretation is incorrect. In fact, as Wurm and Fisicaro (2014) point out "the variable that changes sign…does not relate to the DV in the way theorized, but operates 'as a measure of the sources of error' in the other predictor (Darlington, 1990, p. 155), whose effect is stronger. Put another way, the predictor whose sign has changed accounts for (or suppresses) a portion of the variance in the other predictor that is unrelated to the DV (Pandey & Elliott, 2010)".

[9] In order to explore the feasibility of this approach, we used the LOO functionality of brms (Bürkner, 2016) to run a subset of the 8,000 possible models required for the final study reported in this paper. On a 4-core 2.9 GHz i7 machine, a subset of 200 models took approximately 72 hours.

[10] Only transitive uses were extracted as only transitive uses are candidates for classification as figure-/ground-locative uses. Non-transitive uses (e.g., *It's pouring*) are captured by the overall count.

[11] A reviewer (Mike Ramscar) and the Action Editor (Max Colthear) suggested that many of these asterisked *un-* forms are indeed grammatical. As we noted in Footnote 3, context is everything, and we therefore accept that it is possible to imagine sentences in which the acceptability of these forms would be much improved. At the same time, visual inspection of Figures 31–32 confirms that the vast majority of *un-* forms which we informally describe as "errors" are indeed rated as somewhat less than fully acceptable by adults (again, we stress that acceptability is a graded phenomenon). All except one of these "ungrammatical" forms (*unbend*) have a mean rating below the midpoint of the scale (3). Conversely, all but two of the "grammatical" forms (*unlatch* and *undelete*) have a mean rating above the midpoint of the scale. Thus, while nothing hinges on our informal classification of *un-* forms into correct and errors, it seems to be broadly in line with the judgment data from adult participants.

[12] We use the term *meta-analytic synthesis* rather than *meta-analysis*, because we do not include all of the steps required by a true meta-analysis, in particular a literature search (though since the present article presents new operationalisations of the relevant predictors, we are confident that no directly comparable studies have been omitted).

[13] Note that on this account, hearing a verb in the intransitive construction would NOT block its usage in transitive sentences for which no such re-parse was necessary. For example, Goldberg (1995) points out that the sentence *He sneezed the napkin off the table* appears to be grammaticality acceptable, despite the high frequency of *sneeze* in the intransitive construction, speaking against a role for entrenchment. We suggest this could be partly due to the fact that there is no garden path here: The role of the subject (*he*) is unchanged from its role in an intransitive.

[14] We thank Mike Ramscar for an extremely helpful review that highlighted the relevance of the Rescorla-Wagner model, and the animal learning literature more generally, to the present domain.

[15] For example, Ramscar, Dye & Klein (2013) showed that the performance of 2-year-olds in an ambiguous word-learning task conformed to the predictions of the Rescorla-Wagner model, but not those of an account based on *mutual exclusivity* (the principle that if all but one of the objects present have a known label, a new label must refer to the remaining object). Similarly, Arnon and Ramscar (2011) demonstrated that one reason why second language learners struggle with grammatical gender (e.g., *le chat*) is that the well-learned relationship between the semantic cue (a cat) and the noun (*chat*) blocks the ability to learn the relationship between *le* and *chat*; a finding echoed in the animal-learning literature on *blocking* (e.g., Kamin, 1969), and accounted for by the Rescorla-Wagner model.

Funding Information

Ben Ambridge is Professor in the International Centre for Language and Communicative Development (LuCiD) at The University of Liverpool. The support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no 681296: CLASS).

Competing Interests

The authors have no competing interests to declare.

Author Contributions

- BA wrote the first draft of the paper, conducted the statistical analyses, and designed and ran the new experimental study (Study 5)
- LB conducted the new corpus analysis for Study 1, and gave comments and corrections on the manuscript
- EW advised on data analysis and gave comments and corrections on the manuscript
- CB advised on data analysis, wrote part of the analysis code, and gave comments and corrections on the manuscript
- GS performed the meta-analytic synthesis

References

Alishahi, A., & **Stevenson, S.** (2008). A computational model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789–834. DOI: https://doi.org/10.1080/03640210801929287

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149. DOI: https://doi.org/10.3758/BF03334492

Altman, D. G., & **Bland, J. M.** (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ*, *311*(7003), 485. DOI: https://doi.org/10.1136/bmj.311.7003.485

Ambridge, B. (2013). How do children restrict their linguistic generalizations?: an (un-)grammaticality judgment study. *Cognitive Science, 37*(3), 508–543. DOI: https://doi.org/10.1111/cogs.12018

Ambridge, B. (2017). Horses for courses: When acceptability judgments are more suitable than structural priming (and vice versa). *Behavioral and Brain Sciences, 40*, e284. DOI: https://doi.org/10.1017/S0140525X17000322

Ambridge, B., Bidgood, A., Twomey, E., Pine, J. M., Rowland, C. F., & **Freudenthal, D.** (2015). Preemption versus Entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS ONE, 10*(4), e0123723. DOI: https://doi.org/10.1371/journal.pone.0123723

Ambridge, B., Pine, J. M., & **Rowland, C. F.** (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition, 123*(2), 260–279. DOI: https://doi.org/10.1016/j.cognition.2012.01.002

Ambridge, B., Pine, J. M., Rowland, C. F., Freudenthal, D., & **Chang, F.** (2014). Avoiding dative overgeneralization errors: semantics, statistics or both? *Language, Cognition and Neuroscience, 29*(2), 218–243. DOI: https://doi.org/10.1080/01690965.2012.738300

Ambridge, B., Pine, J. M., Rowland, C. F., Jones, R. L., & **Clark, V.** (2009). A semantics-based approach to the 'no negative-evidence' problem. *Cognitive Science, 33*(7), 1301–1316. DOI: https://doi.org/10.1111/j.1551-6709.2009.01055.x

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*(1), 87–129. DOI: https://doi.org/10.1016/j.cognition.2006.12.015

Ambridge, B., Rowland, C. F., Theakston, A. L., & **Kidd, E. J.** (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*(2), 239–73. DOI: https://doi.org/10.1017/S030500091400049X

Arnon, I., & **Ramscar, M.** (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*(3), 292–305. DOI: https://doi.org/10.1016/j.cognition.2011.10.009

Baayen, R. H., Milin, P., & **Ramscar, M.** (2016). Frequency in lexical processing. *Aphasiology, 30*(11), 1174–1220. DOI: https://doi.org/10.1080/02687038.2016.1147767

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Enquiry, 10*, 533–581.

Barak, L., Goldberg, A. E., & **Stevenson, S.** (2016). Comparing computational cognitive models of generalization in a language acquisition task. *Proceedings of the 2016 conference on Empirical Methods in Natural Language Processing*, 96–106. DOI: https://doi.org/10.18653/v1/D16-1010

Barr, D. J., Levy, R., Scheepers, C., & **Tily, H. J.** (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language, 68*(3), 255–278. DOI: https://doi.org/10.1016/j.jml.2012.11.001

Bates, M., Maechler, M., Bolker, B., & **Walker, S.** (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Bellingham, W. P., Gillette-Bellingham, K., & **Kehoe, E. J.** (1985). Summation and configuration in patterning schedules with the rat and rabbit. *Learning & Behavior, 13*(2), 152–164. DOI: https://doi.org/10.3758/BF03199268

Bidgood, A., Ambridge, B., Pine, J. M., & **Rowland, C. F.** (2014). The retreat from locative overgeneralisation errors: A novel verb grammaticality judgment study. *PLoS ONE, 9*(5), e97634. DOI: https://doi.org/10.1371/journal.pone.0097634

Blything, R. P., Ambridge, B., & **Lieven, E. V. M.** (2014). Children use statistics and semantics in the retreat from overgeneralization. *PLoS ONE, 9*(10), e110009. DOI: https://doi.org/10.1371/journal.pone.0110009

Bowerman, M. (1988). The "no negative evidence" problem: how do children avoid constructing an overly general grammar? In: Hawkins, J. A. (Ed.), *Explaining language universals*, 73–101. Oxford: Blackwell.

Boyd, J. K., & **Goldberg, A. E.** (2011). Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language, 87*(1), 55–83. DOI: https://doi.org/10.1353/lan.2011.0012

Braine, M. D. S. (1971). On two types of models of the internalization of grammars. In: Slobin, D. I. (Ed.), *The ontogenesis of grammar*, 153–186. New York: Academic Press.

Braine, M. D. S., & **Brooks, P. J.** (1995). Verb argument strucure 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*, 352–376. Hillsdale, NJ: Erlbaum.

Brooks, P. J., & **Tomasello, M.** (1999). How children constrain their argument structure constructions.

Language, 75(4), 720–738. DOI: https://doi.org/10.2307/417731

Brooks, P. J., Tomasello, M., Dodson, K., & **Lewis, L. B.** (1999). Young children's overgeneralizations with fixed transitivity verbs. *Child Development, 70*(6), 1325–1337. DOI: https://doi.org/10.1111/1467-8624.00097

Brooks, P. J., & **Zizak, O.** (2002). Does preemption help children learn verb transitivity? *Journal of Child Language, 29*, 759–781. DOI: https://doi.org/10.1017/S0305000902005287

Bürkner, P. C. (2016). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton. DOI: https://doi.org/10.2307/412745

Clark, E. V., & **Clark, H. H.** (1979). When nouns surface as verbs. *Language*, 767–811. DOI: https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003. DOI: https://doi.org/10.1017/CBO9780511803864

Croft, W., & **Cruse, D. A.** (2004). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Darlington, R. B. (1990). Regression and linear models. New York: McGraw-Hill Publishing Company.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. DOI: https://doi.org/10.3389/fpsyg.2014.00781

Eager, C., & **Roy, J.** Mixed Effects Models are Sometimes Terrible. arXiv preprint. https://arxiv.org/abs/1701.04858.

Fisher, C., Gleitman, H., & **Gleitman, L. R.** (1991). On the semantic content of subcategorization frames. *Cognitive psychology, 23*(3), 331–392. DOI: https://doi.org/10.1016/0010-0285(91)90013-E

Frazier, L., & **Rayner, K.** (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology, 14*(2), 178–210. DOI: https://doi.org/10.1016/0010-0285(82)90008-1

Gallistel, C. R. (2003). Conditioning from an information processing perspective. *Behavioural Processes, 62*, 89–101. DOI: https://doi.org/10.1016/S0376-6357(03)00019-6

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & **Lotocky, M. A.** (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*(1), 58–93. DOI: https://doi.org/10.1006/jmla.1997.2512

Gelman, A., & **Stern, H.** (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328–331. DOI: https://doi.org/10.1198/000313006X152649

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. New York: Oxford University Press.

Goldberg, A. E. (2011). Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics, 22*(1), 131–153. DOI: https://doi.org/10.1515/cogl.2011.006

Goldberg, A. E. (in press). *Explain Me This*. Princeton: Princeton University Press.

Gries, S. Th. (2012). Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. *Studies in Language 36*(3). 477–510. DOI: https://doi.org/10.1075/sl.36.3.02gri

Gries, S. Th. (2015). More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff (2013). *Cognitive Linguistics*, *26*(3), 505–536. DOI: https://doi.org/10.1515/cog-2014-0092

Gropen, J., Pinker, S., Hollander, M., & **Goldberg, R.** (1991). Affectedness and Direct Objects – the Role of Lexical Semantics in the Acquisition of Verb Argument Structure. *Cognition, 41*(1–3), 153–195. DOI: https://doi.org/10.1016/0010-0277(91)90035-3

Gureckis, T. M., & **Love, B. C.** (2010). Direct associations or internal transformations? Exploring the mechanisms underlying sequential learning behavior. *Cognitive Science, 34*(1), 10–50. DOI: https://doi.org/10.1111/j.1551-6709.2009.01076.x

Hahn, U., & **Oaksford, M.** (2008). Inference from absence in language and thought. In: Chater, N., & Oaksford, M. (Eds.), *The probabilistic mind*, 107–112. Oxford, England: Oxford University Press. DOI: https://doi.org/10.1093/acprof:oso/9780199216093.003.0006

Harmon, Z., & **Kapatsinski, V.** (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology, 98*, 22–44. DOI: https://doi.org/10.1016/j.cogpsych.2017.08.002

Hughes, J. J. (in press). Corpus Linguistics and Event-Related Potentials In: Baker, P., & Egbert, J. (Eds.). Triangulating Corpus Linguistics with Other Methods. London: Routledge.

Holland, P. C. (1983). Occasion setting in Pavlovian feature positive discriminations. In: Commons, M. L., Herrnstein, R. J., & Wagner, A. R. (Eds.), *Quantitative analyses of behavior: Discrimination processes, 4*, 183–206. New York: Ballinger.

Hsu, A. S., & **Chater, N.** (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science, 34*(6), 972–1016. DOI: https://doi.org/10.1111/j.1551-6709.2010.01117.x

Humboldt, W. V. (1836). *On the Difference of Human Language Building*. Berlin: Claasen & Roether.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In: Cambell, B. A., & Church, R. M. (Eds.), *Punishment and Aversive Behavior*. New York: Appleton-Century-Crofts.

Klein, D., & **Manning, C. D.** (2003). Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics,* 423–430. DOI: https://doi.org/10.3115/1075096.1075150

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.

Li, P., & **MacWhinney, B.** (1996). Cryptotype, overgeneralization and competition: A connectionist model of the learning of English reversive prefixes. *Connection Science, 8*(1), 3–30. DOI: https://doi.org/10.1080/095400996116938

MacDonald, M. C., Pearlmutter, N. J., & **Seidenberg, M. S.** (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676. DOI: https://doi.org/10.1037/0033-295X.101.4.676

Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A. M., D'Hooge, R., De Houwer, J., & **Beckers, T.** (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General, 145*(9), e49. DOI: https://doi.org/10.1037/xge0000200

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & **Bates, D.** (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. DOI: https://doi.org/10.1016/j.jml.2017.01.001

McElreath, R. (2016). rethinking: Statistical Rethinking book package. R package version 1.59.

Pandey, S., & **Elliott, W.** (2010). Supressor variables in social work research: Ways to identify in multiple regression models. *Journal of the Society for Social Work and Research, 1*, 28–40. DOI: https://doi.org/10.5243/jsswr.2010.2

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407–418. DOI: https://doi.org/10.1037/h0022602

Perek, F., & **Goldberg, A. E.** (2017). Linguistic generalization on the basis of function and constraints on the basis of statistical preemption. *Cognition*, *168*, 276–293.

Perfors, A., Tenenbaum, J. B., & **Wonnacott, E.** (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of child language, 37*(3), 607–642. DOI: https://doi.org/10.1017/S0305000910000012

Perfors, A., & **Wonnacott, E.** (2011). Bayesian modeling of sources of constraint in language acquisition. In: Clark, E. V. (Ed.), *Experience, Variation, and Generalization: Learning a first language. Amsterdam: John Benjamins*, 277–294. DOI: https://doi.org/10.1075/tilar.7.16per

Phillips, C., & **Ehrenhofer, L.** (2015). The role of language processing in language acquisition. *Linguistic Approaches to Bilingualism, 5*(4), 409–453. DOI: https://doi.org/10.1075/lab.5.4.01phi

Pinker, S. (1989). *Learnability and Cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Ramscar, M., Dye, M., & **Klein, J.** (2013). Children value informativity over logic in word learning. *Psychological Science, 24*(6), 1017–1023. DOI: https://doi.org/10.1177/0956797612460691

Ramscar, M., Dye, M., & **McCauley, S. M.** (2013). Error and expectation in language learning: The curious absence of mouses in adult speech. *Language, 89*(4), 760–793. DOI: https://doi.org/10.1353/lan.2013.0068

Ramscar, M., Sun, C. C., Hendrix, P., & **Baayen, R. H.** (2017). The Mismeasurement of mind: Life-span changes in paired-associate-learning scores reflect the "cost" of learning, not cognitive decline. *Psychological Science, 28*(8): 1171–1179. DOI: https://doi.org/10.1177/0956797617706393

Ramscar, M., & **Yarlett, D.** (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science, 31*(6), 927–960. DOI: https://doi.org/10.1080/03640210701703576

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1–5. DOI: https://doi.org/10.1037/h0025984

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist, 43*, 151–60. DOI: https://doi.org/10.1037/0003-066X.43.3.151

Rescorla, R. A., & **Wagner, A. R.** (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A. H., & Prokasy, W. F. (Eds.), *Classical Conditioning II*, 64–99. New York, NY: Appleton-Century-Crofts.

Robenalt, C., & **Goldberg, A. E.** (2015). Judgment evidence for statistical preemption: It is relatively better to vanish than to disappear a rabbit, but a lifeguard can equally well backstroke or swim children to shore. *Cognitive Linguistics, 26*(3), 467–503. DOI: https://doi.org/10.1515/cog-2015-0004

Robenalt, C., & **Goldberg, A. E.** (2016). Nonnative speakers do not take competing alternative expressions into account the way native speakers do. *Language Learning, 66*(1), 60–93. DOI: https://doi.org/10.1111/lang.12149

Schmidt, F. L., & **Hunter, J. E.** (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Newbury Park, CA: Sage. DOI: https://doi.org/10.4135/9781483398105

Snedeker, J., & **Trueswell, J. C.** (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49*(3), 238–299. DOI: https://doi.org/10.1016/j.cogpsych.2004.03.001

Stan Development Team. (2015a). Stan: A C++ Library for Probability and Sampling, Version 2.10.0. URL: http://mc-stan.org/.

Stan Development Team. (2015b). Stan Modeling Language User's Guide and Reference Manual, Version 2.10.0. URL: http://mc-stan.org/.

Stefanowitsch, A. (2008). Negative evidence and preemption: A constructional approach to

ungrammaticality. *Cognitive Linguistics, 19*(3), 513–531.

Stefanowitsch, A., & **Gries, S. T.** (2003). Collostructions: Investigating the interaction of words and constructions. *International journal of corpus linguistics, 8*(2), 209–243. DOI: https://doi.org/10.1075/ijcl.8.2.03ste

Tatsumi, T., Ambridge, B., & **Pine, J. M.** (2017). Disentangling Effects of Input Frequency and Morphophonological Complexity on Children's Acquisition of Verb Inflection: An Elicited Production Study of Japanese. *Cognitive Science*.

Tatsumi, T., Ambridge, B., & **Pine, J. M.** (in press). Testing an input-based account of children's errors with inflectional morphology: An elicited production study of Japanese. *Journal of Child Language*. DOI: https://doi.org/10.1017/S0305000918000107

Theakston, A. L. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgement tasks. *Cognitive Development, 19*(1), 15–34. DOI: https://doi.org/10.1016/j.cogdev.2003.08.001

Trueswell, J. C., Tanenhaus, M. K., & **Kello, C.** (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 528. DOI: https://doi.org/10.1037/0278-7393.19.3.528

Twomey, K. E., Chang, F., & **Ambridge, B.** (2014). Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition. *Cognitive Psychology, 73*, 41–71. DOI: https://doi.org/10.1016/j.cogpsych.2014.05.001

Vehtari, A., Gelman, A., & **Gabry, J.** (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. DOI: https://doi.org/10.1007/s11222-016-9696-4

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. URL: http://www.jstatsoft.org/v36/i03/. DOI: https://doi.org/10.18637/jss.v036.i03

Westfall, J., & **Yarkoni, T.** (2016). Statistically controlling for confounding constructs is harder than you think. *PloS one, 11*(3), e0152719. DOI: https://doi.org/10.1371/journal.pone.0152719

Wonnacott, E., Newport, E. L., & **Tanenhaus, M. K.** (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology, 56*(3), 165–209. DOI: https://doi.org/10.1016/j.cogpsych.2007.04.002

Wurm, L. H., & **Fisicaro, S. A.** (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37–48. DOI: https://doi.org/10.1016/j.jml.2013.12.003

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society, 75*(6), 579–652. DOI: https://doi.org/10.2307/2340126

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.133.pr

How to cite this article: 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): 23. DOI: https://doi.org/10.1525/collabra.133

Senior Editor: Rolf Zwaan

Editor: Fernanda Ferreira

Submitted: 10 January 2018 **Accepted:** 22 May 2018 **Published:** 02 July 2018

UNIVERSITY of CALIFORNIA PRESS | Collabra

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

OPEN ACCESS