The partial productivity of constructions as induction[1]

LAURA SUTTLE AND ADELE E. GOLDBERG

Abstract

Whether words can be coerced by constructions into new uses is determined in part by semantic sensicality and statistical preemption. But other factors are also at play. Experimental results reported here suggest that speakers are more confident that a target coinage is acceptable to the degree that attested instances cover the semantic space that includes the target coinage. The relevance of coverage is supported by combined effects of type frequency and variability of attested instances [Experiments 1a–1b], and an expected interaction between similarity and variability [Experiment 3]. Similarity to an attested instance is also found to play a role: speakers are more confident of a target coinage when the coinage is more similar to an attested instance [Experiment 3]. Experiment 2 provides a manipulation check that indicates that participants are in fact basing their confidence ratings on the perceived productivity of constructions. The results reported here lend support to the idea that the productivity of constructions depends on general properties of induction.

1. Introduction

Words can often be used in novel ways, allowing speakers to produce sentences that they have never heard before. At times this ability gives rise to noticeably novel phrases such as Dylan Thomas's *a grief ago*, or to an utterance such as *She sneezed the foam off the cappuccino* (Ahrens 1995). How is it that *a grief* can appear in a slot normally used for temporal phrases (e.g., *three years ago*) and a normally intransitive verb, *sneeze,* can appear in a frame prototypically used with verbs of caused-motion? The process involved is often referred to as *coercion* (Jackendoff 1997; Michaelis 2004; Pustejovsky 1995) or *accommodation* (Goldberg 1995: 159): a construction coerces the meaning of a word so that the word is construed to be compatible with the construction's function.

Linguistics 49–6 (2011), 1237–1269
DOI 10.1515/LING.2011.035
0024–3949/11/0049–1237

Intriguingly, coercion is not a fully general process. There are constraints on which words can appear in which constructions, even when the intended interpretation is completely clear. Thus we find the examples in (1)–(3) are ill formed.

(1) ?? *She explained him the news.*
(2) ?? *She wept herself to sleep.*
(3) ?? *She saw an afraid boy.*

Many researchers have discussed how complicated the general issue of constraining generalizations is. The language people hear does not come overtly marked with question marks or asterisks to indicate unacceptability, and we know speakers are not generally corrected for producing ill-formed utterances (Baker 1979; Bowerman 1988; Braine 1971; Brown and Hanlon 1970; Marcus 1993; Pinker 1989; but cf. Chouinard and Clark 2003).

The unacceptability of cases such as these (particularly 1 and 2) has been discussed with respect to the phenomenon of *partial productivity*: constructions are partially but not fully productive; i.e., they may be able to be extended for use with a limited range of items (Bowerman 1988; Pinker 1989; Goldberg 1995). This is the terminology we use here. Thus our use of the term *productivity* corresponds roughly to Barðdal (2008: 27)'s notion of *extensibility*.

It seems that productivity and *coercion* may both refer to aspects of the same phenomenon, at least when applied to phrasal constructions. A construction is considered to be productive to the extent that it can coerce new words to appear in it. If there is a difference, it is that *coercion* tends to be used when there is an intuition that a word is construed somewhat unusually in order to appear in the construction (i.e., the construction "coerces" the word). Also, researchers tend to use the term *coercion* when the resulting phrases are noticeably novel, whereas *productivity* is used when the resulting phrases are unremarkable. But the dividing line is not clear cut. The following each involve novel uses of the main verbs and so are productive instances of the constructions; insofar as each highlights different aspects of the main verbs than is apparent in other constructions, each requires the verb to be coerced by the construction.

(4) *She'd smiled herself an upgrade.*
(Douglas Adams, Hitchhiker's Guide to the Galaxy)
(5) *Sarah . . . winked her way through the debates.*
(http://pcneedtogo.blogspot.com)
(6) *Tim . . . sneezed the milk out of his nose.*
(http://www.zoackkennel.com/tims-story.html)
(7) *I actually had a moth go up my nose once. I . . . coughed him out of my mouth.*
(http://bikeforums.net/archive/index.php/t-292132)

We use the more neutral term productivity instead of coercion in the rest of the article, because we are not assuming that judgments are necessarily based on any unusual interpretation of the verbs involved.

There are two minimal criteria that must be met for a target coinage to be judged acceptable:

(i) The coinage must be *semantically sensical.*
(ii) The coinage *must not be preempted* by a conventional formulation with the same or more appropriate function.

Beyond these restrictions, there exist three additional, gradient factors that may be relevant: *type frequency, variability*, and *similarity*. This article provides experimental evidence that investigates whether speakers are more *confident* that a coinage is acceptable to the extent that:

(iii) The pattern has been witnessed with multiple instances (type frequency).
(iv) The pattern is relatively variable, being witnessed with a broad variety of instances.
(v) The potential coinage is relatively semantically similar to an attested instance.

In Section 5, we observe that the rather nuanced evidence gathered concerning type frequency, variability, and similarity combine to argue in favor of two general factors: *similarity* and *coverage.* Coinages are acceptable to the extent that they are similar to an existing attested instance, and coinages are acceptable to the extent that the semantic (and/or phonological) space is well covered by the smallest category that encompasses both the coinage and attested instances (Osherson et al. 1990; Goldberg 2006: 98). However, before we investigate these factors, we first briefly review evidence for the first two criteria: semantic sensicality and statistical preemption.

1.1. *Utterances must be semantically sensical*

The first criterion, that a coinage must be interpretable in context, is easy to take for granted. We do not produce utterances that make no sense because no one would understand us. Particularly relevant here is that the meaning of an utterance must be consistent with semantic constraints on the construction (Ambridge et al. 2009). This corresponds to Goldberg's (1995: 50–54) "Semantic Compatibility" constraint on how constructions and verbs can be combined.

Context can often ameliorate otherwise ill formed expressions if it serves to provide a sensical interpretation. For example, the [⟨time quantity⟩ *ago*]

construction requires that the first constituent be interpreted as a measurable time quantity. It is most commonly used with temporal phrases such as *three years, a decade,* or *one and half minutes*. *Ago* can also be used with complements that refer to events that can occur at specific intervals (e.g., *three games ago*). It can be extended to complements that refer to objects that metonymically refer to events that occur at specific intervals (*three stop lights ago*). We do not generally think of *grief* as something that recurs at regular intervals, nor as the type of bounded event that can be counted, but the expression *a grief ago* coerces just that interpretation. Other meanings may be more difficult to coerce, leading to infelicity: e.g., ?? *a future ago,* ?? *a past ago*. If contexts can be found to make sense of these phrases, they are immediately judged much improved.

1.2. *Statistical preemption*

A number of theorists have suggested that a process of *preemption* plays a role in speakers learning to avoid syntactic overgeneralizations (Clark 1987; Foraker et al. 2007; Goldberg 1993, 1995, 2006; Pinker 1981). Preemption can be viewed as a particular type of indirect negative evidence. It is an implicit inference speakers make from repeatedly hearing a formulation, B, in a context where one might have expected to hear a semantically and pragmatically related alternative formulation, A. The result is that speakers implicitly recognize that B is the appropriate formulation in such a context; this yields an implicit inference that A is not appropriate.

Preemption (or blocking) is already familiar from morphology: *did* preempts *do-ed, feet* preempts *foots,* and *go* preempts *went* (Aronoff 1976; Kiparsky 1982). The way speakers learn to say *went* instead of *goed* is that they repeatedly and consistently hear *went* in contexts in which *goed* would otherwise have been appropriate.

The idea that preemption is found between two phrasal forms requires discussion, since expressions formed from distinct phrasal constructions are virtually never semantically and pragmatically identical, and thus it is not clear that an instance of one phrasal pattern could preempt the use of another (Bowerman 1996; Pinker 1989). For example, the ditransitive construction in (8a) is distinct, at least in terms of its information structure, from the prepositional paraphrase (8b) (e.g., Goldberg 1995; Green 1974; Rappaport Hovav and Levin 2005; Bresnan et al. 2007). Thus, knowledge that the prepositional paraphrase is licensed as in (8b) should not in any simple way preempt the use of the ditransitive (8a). And, in fact, a large number of verbs do freely appear in both constructions (e.g., *tell* as in 9a–b).

(8) a. ?? *She explained me the story.*
b. *She explained the story to me.*

(9) a. *She told me the story.*
b. *She told the story to me.*

But preemption can be seen to play an important role in learning to avoid expressions such as (8a), once a speaker's expectations are taken into account in the following way. Learners may witness repeated situations in which the ditransitive might be expected because the relevant information structure suits the ditransitive at least as well as the prepositional paraphrase. If, in these situations, the prepositional alternative is systematically witnessed instead, the learner can infer that the ditransitive is not after all appropriate (Goldberg 1993, 1995, 2006, 2011; Marcotte 2005).

As Goldberg (2006: 94–98) has emphasized, the process is necessarily statistical, because a single use of the alternative formulation could be due to some subtle difference in the context that actually favors the alternative formulation. Or a single use may simply be due to an error by the speaker. But if an alternative formulation is consistently heard, a process of statistical preemption predicts that speakers will learn to use the alternative.

Statistical preemption has not received a great deal of attention in the experimental literature, except in a few notable articles (Brooks and Tomasello 1999; Brooks and Zizak 2002) and in some recent work by Boyd and Goldberg (2011). Brooks and colleagues demonstrated that seeing novel intransitive verbs in periphrastic causative constructions significantly preempts six-year-old children's use of the verbs in simple transitives (Brooks and Tomasello 1999; Brooks and Zizak 2002). Boyd and Goldberg (2011) investigated how speakers could learn to avoid using certain adjectives with an initial schwa sound, such as *afraid,* before nouns. Note that example (10) sounds decidedly odd:

(10) ?? *the afraid boy*
(11) *the scared boy*

They demonstrate that speakers avoid using even novel schwa-initial adjectives such as *afek* prenominally to some extent (Boyd and Goldberg 2011, Experiment 1), whereas novel non-a-adjectives (e.g., *chammy*) readily appear prenominally. If novel adjectives are witnessed in the preemptive context of a relative clause (e.g., *The cow that was afek moved to the star*), the novel adjectives behave indistinguishably from familiar a-adjectives in resisting appearance before nouns (2011, Experiment 2). Moreover, speakers generalized evidence gleaned from statistical preemption to other members of the nonsense schwa-initial category such as *ablim*; i.e., they avoided using *ablim* prenominally even though *ablim* was never witnessed in a preemptive context. Thus speakers make use of preemptive contexts and are even capable of generalizing the restriction to other members of a category.

These two factors, semantic sensicality and statistical preemption, combine to minimally allow and constrain the creative use of words in constructions. In the studies described below, we investigate three additional possible factors that may influence speakers' confidence levels in using verbs in creative ways: type frequency, variability, and similarity. Each of these factors is introduced below briefly before we turn to the experiments. Again, in section 5, the range of findings is discussed in terms of the more general criteria of similarity and coverage.

2. Additional gradient factors

2.1. *High type frequency*

Type frequency refers simply to the number of different (head) words that are witnessed in a given construction. Many have suggested that, *ceteris paribus*, the higher the type frequency of a pattern, the higher the productivity (Barðdal 2008; Bybee 1985, 1995; Clausner and Croft 1997; Goldberg 1995; Tomasello 2003). For example, argument structure constructions that have been witnessed with many different verbs are more likely to be extended to appear with additional verbs. To some extent, this observation has to be correct: learners consider a pattern extendable if they have witnessed the pattern being extended.

The role of type frequency has been established quite clearly in the morphological domain (e.g., Aronoff 1983; Bybee 1985). For example, irregular past tense patterns are only extended with any regularity at all if the type frequency of the pattern reaches half a dozen or so instances. For example, the pattern that involves /-id/ → /-ɛd/ in the past tense is attested in *read*/*read, lead*/*led, bleed*/*bled, feed*/*fed, speed*/*sped*. Albright and Hayes (2003) found that 23% of speakers productively suggested /*glɛd*/ as the past tense of /*glid*/.[2] On the other hand, the pattern of /-ɛl/ → /-old/ is only attested in two lexemes: *tell-told* and *sell-sold;* correspondingly, no respondents offered *grold* as the past tense of *grɛl.* Type frequency generally correlates well with the degree of productivity.

2.2. *Variability*

The degree of *variability* of a construction corresponds to the range of attested instances.[3] We hypothesize that the more variable a pattern is, the more likely it is to be extended; i.e., all other things being equal, constructions that have been heard used with a wide range of verbs are more likely to be extended than constructions that have been heard used with a semantically or phonologically

circumscribed set of verbs (for evidence of this in the lexical and morphological domains see Bowerman and Choi 2001; Bybee 1995; Janda 1990).[4]

Type frequency and degree of variability are often confounded in real language samples, since the degree of variability is likely to be higher the more attested types there are that appear in a given construction (cf. also Barðdal 2008). Experiments #1a-b aim to tease apart these two potentially important factors.

2.3. *Similarity*

With many others, we hypothesize that a new verb may be extended with greater confidence when that new verb is relevantly similar to one or more verbs that have already been witnessed in an argument structure construction (cf. also Barðdal 2008, 2011; Cruse and Croft 2004; Langacker 1987; Zeschel and Bildhauer 2009). In fact, similarity to attested instances has been argued to be the most relevant factor in licensing coinages (Bybee and Eddington 2006; Kalyan 2011).

Different researchers have primarily used two different ways of calculating similarity. *Summed* similarity involves comparing the coinage to all attested instances and summing the totals. *Maximum* similarity involves only comparing the coinage to the instance with which it is the most similar (e.g., Osherson et al. 1990).

If summed similarity were the relevant measure, it would follow that overall similarity would monotonically increase with type frequency, as long as the similarity is non-zero. This leads to the counterintuitive idea that a coinage would be ten times more acceptable if 100 instances that are similar to one another but relatively dissimilar to the coinage have been witnessed as compared with a situation in which 10 instances that are similar to one another but dissimilar to the coinage have been witnessed. Intuitively, witnessing 100 instances that are relatively alike and distinct from the coinage might well give rise to the inference that *only* instances of the same general type as the 100 instances are allowed. We therefore put aside the potential influence of summed similarity and focus instead on maximum similarity. Experiment #3 investigates potential roles of maximum similarity and variability in novel coinages.

3. Experiments 1a–1b: Type frequency and variability

Experiment 1a was designed to determine whether variability and/or type frequency lead to generalization of grammatical constructions and whether the

two factors potentially interact. Participants were given sets of sentences of a fictitious language, "Zargotian," and asked to determine how likely it was that a final target sentence was also a legitimate sentence in Zargotian. Type frequency (number of instances) and variability (degree of semantic similarity among instances) were manipulated such that each participant judged cases that had type frequency that was low (1 instance) vs. medium (3 instances) vs. high (6 instances) crossed with low vs. high variability. An example stimuli set is given in (12):

(12) Example set (involving medium type frequency (3 instances); high variability (*toast, crease* and *slap* are not very quantitatively semantically related):
Assume you can say these sentences:
The *zask* the *nop* toast-*pe.*
The *vash* the *yerd* crease-*pe.*
The *blib* the *nalf* slap-*pe.*
How likely is it that you can also say:
The *isp* the *bliz* clip-*pe.* ? Answer: __%

The definite article was used to indicate that novel words were nouns; familiar verbs were used along with various nonsense sentence-final particles (e.g., *pe* in the example above).

In effect, we are investigating the extent to which speakers are willing to extend a pattern by measuring how likely they are to judge the extension acceptable. We use the term *construction* because the form is given, and the two NPs imply some sort of two-argument semantics. Each construction was differentiated by the order of the words as well as the sentence particle used. The semantics was intentionally underspecified: no glosses were given and the nonsense nouns provided no content.

The design is represented schematically in Figure 1 below; target coinages are represented by a grey square in each condition, and attested instances are represented by black circles. The low type frequency case (type frequency of one), which is by necessity only low variability, is not pictured.

In Experiment 1, we controlled for maximum similarity: i.e., the similarity of the verb class of the target coinage to the verb class of its closest attested neighbor. For example, if the two closest neighbors in one condition came from the "bend" and "cut" classes, then the closest neighbors in every condition came from these same two classes. Semantic similarity of verbs was determined using Latent Semantic Analysis, which determines similarity on the basis of co-occurrence information in large corpora (Landauer and Dumais 1997). The verbs used in the study are provided in Table 1 in the methods section.

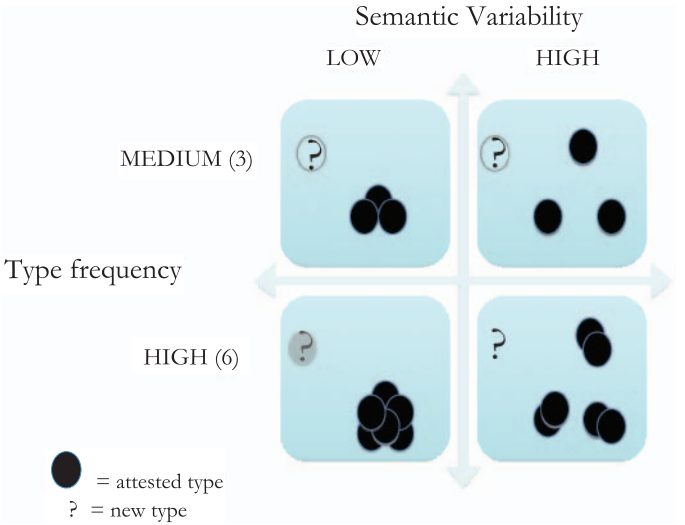

Figure 1. *Schematic representation of four experimental conditions (A fifth condition: Low type frequency|low variability not pictured). Target coinage represented by gray square. Attested instances are represented by black circles.*

3.1. *Method*

3.1.1. *Participants.* Fifty-five participants were paid $.75 to fill out a 5–10 minute online questionnaire on Amazon Mechanical Turk (www.mturk.com). Results obtained using Mechanical Turk have been shown to be reliable in other work (Paolacci et al. 2010; Buhrmester et al. 2011). The low wage is consistent with Mechanical Turk's compensation scale. Five participants were automatically excluded for using a single value as explained below. The data from the remaining 50 participants were analyzed.

3.1.2. *Procedure.* Stimuli for all experiments reported here were created using the nine verb classes provided in Table 1 (cf. Levin 1993). Experiment 3 included one additional verb class (verbs of cognition) described in section 5.

There were five conditions and three items in each condition for a total of fifteen stimuli sets in each questionnaire. Each set of stimuli involved the same construction (same word order and same sentence final particle); each different set of stimuli involved a different construction.

Four different lists were created using various items from each verb class, in an effort to avoid item effects. Within each list, all verb classes occurred at least once. Targets were created so that each verb class was represented as a

Table 1. *Verb classes used in all experiments (cf. Levin 1993)*

Break	break, crush, tear, smash, rip
Load	load, pile, heap, cram, pack, crowd
Bend	bend, crinkle, fold, crumple, crease, rumple
Cook	bake, cook, heat, fry, toast, boil
Cut	cut, snip, scratch, slash, scrape, clip
Get	acquire, buy, order, grab, get, borrow
Throw	throw, flip, pass, pitch, hurl, fling
Hit	smack, hit, kick, slap, knock, bump
Hold	hold, clasp, grip, grasp, clutch, handle

target in each list at least once. Within each list, items were presented in randomized order with the constraint that no verb appeared in two adjacent items.

Verbs in the high variability condition were drawn from three different verb classes; in the case of type frequency of six, two verbs were drawn from each of the three classes. Semantic similarity among the three classes chosen for each of the stimuli sets (the degree of variability) was controlled for, using Latent Semantic Analysis to determine semantic distances between verb classes. In Experiments 1a–1b and 2, the target coinage was drawn from a verb class that was distinct from those used as attested instances for that stimuli set; the semantic distance between each target and its closest neighbor was held constant.

Verbs in the low variability condition were all drawn from the same semantic class, with the particular class varying across stimuli sets such that each verb class was represented at least once.

No individual verb appeared more than twice within each list; most only occurred once.

3.1.3. *Results.* Participants were free to choose any range of values between 1 and 100 for their likeliness scores. Because participants used various ranges (e.g., 40–60, or 20–80, or 80–100), the analysis is based on z-scores for each subject. Z-scores allow us to compare participants' relative confidence of the grammaticality of particular coinages by calculating the standard deviation of each score from that person's mean. Because variation in scores is required to calculate z-scores, five participants who responded using a single value for all stimuli were dropped. This had the advantage of removing data from those few participants who did not appear to attempt to differentiate among possible responses. Table 2 shows the mean z-scores (with standard errors in parentheses) for the five experimental conditions.

Since it is not possible to have a single instance with high variability, we first performed a 2 × 2 repeated measures ANOVA using the other four cells (3

Table 2. *Results (n = 50) of Experiment 1a: Z-score means of how likely the target sentence is judged to be acceptable, given 1, 3 or 6 attested instances of low or high variability. Standard errors are in parentheses.*

	Low Variability	High Variability
1 instance	−.28 (.09)	N/A
3 instances	−.05 (.06)	.03 (.10)
6 instances	−.14 (.08)	.44 (.06)

and 6 instances × low and high variability). Results demonstrate a main effect of variability, $F(1,47) = 12.94$ ($p = .001$), and a marginal effect of type frequency, $F(1,47) = 3.37$ ($p = .07$), with a significant interaction, $F(1,47) = 9.21$ ($p = .004$), due to the fact that variability had a stronger effect in the high type frequency condition. The interaction suggests that type frequency and variability may be more than additive. However, we will see that the interaction is not replicated in Experiment 1b, so we leave that aspect of the findings aside.

Overall, stimuli sets that included more variable verbs (more "open" sets) were judged as acceptable with more confidence than those with low variability. Also, stimuli sets that included six examples (high type frequency) were judged to be acceptable with marginally more confidence than those with three instances (low type frequency).

To further explore the effect of type frequency, we included the one instance condition in the analysis by averaging the low and high variability conditions. A three level ANOVA reveals a clear main effect of type frequency, $F(2,94) = 7.81$ ($p = .001$). There is also a significant linear trend across the three type frequency conditions, $F(1,47) = 13.04$ ($p = .001$). Thus higher type frequency increases participants' confidence that a construction can be used productively.

3.2. *Experiment 1b: Replication*

Experiment 1b is a simple replication of Experiment 1a with a new group of 51 participants. We felt a replication was needed because we were unable to supervise participants due to the online nature of the survey. We also wanted to determine whether the interaction between type frequency and variability would replicate.

3.2.1. *Participants.* 51 new participants were paid $.75 to fill out the online questionnaire on Amazon Mechanical Turk. Two participants were automatically excluded for using only a single value as discussed above. The data from the remaining 49 participants were analyzed.

3.2.2. *Procedure.* Experiment 1b used the same survey as Experiment 1a, and results were analyzed the same way. We were able to ensure with reasonable confidence that participants had not taken part in Experiment 1a by comparing their network IDs with those of participants that took part in the 1a.

3.2.3. *Results.* The results of this replication are reported in Table 3.

Table 3. *Results (n = 51) of Experiment 1b: Z-score means of how likely the target sentence is judged to be acceptable, given 1, 3 or 6 attested instances of low or high variability. Standard errors are in parentheses.*

	Low Variability	High Variability
1 instance	−.43 (.07)	N/A
3 instances	−.03 (.06)	.12 (.07)
6 instances	.04 (.08)	.31 (.07)

The 2 × 2 repeated measures ANOVA, again using the four cells involving 3 and 6 instances, replicates the main effect of variability, $F(1,45) = 5.49$, ($p = .02$), with p value of .11 for type frequency, $F(1,45) = 2.63$. This time, there is no significant interaction between the two, $F(1,45) = .54$, ($p = .47$). We therefore do not attempt to further explain the interaction in Experiment 1a, but leave the finding aside for future research.

Including the one-instance condition in the analysis by averaging the low and high variability conditions, a three level ANOVA again reveals a clear main effect of type frequency, $F(2,90) = 25.76$, ($p < .001$). And, as in Experiment 1a, there is a significant linear trend across the three type frequency conditions, $F(1,45) = 38.11$, ($p < 001$).

Thus, the main effects of Experiment 1a were straightforwardly replicated in Experiment 1b. They provide clear evidence for a positive effect of both higher variability and higher type frequency. Both factors encourage constructional productivity, with a possible positive interaction between them (Experiment 1a only).

4. Experiment 2: Manipulation check

The variability factor was designed to investigate whether speakers use judgments about semantic spread (or "variability") in deciding how confidently they can extend a construction. But we were concerned that participants might interpret the task as requiring a simple judgment relating the target verb to a set of other verbs without relevance to the productivity of grammatical construc-

tions. As a manipulation check to determine whether judgments were based on the extensibility of grammatical constructions as intended, we tested a new group of participants on a modified design. In Experiment 2, each target sentence involved a different construction than the attested instances. An example stimulus set is given in (13); note in particular that the NP NP V-*pe* construction in the premises is distinct from the V-*to* NP NP construction of the target sentence.[4] If speakers are ignoring the constructional aspect of the experiment, we would expect them to perform as they did in Experiments 1a–1b in Experiment 2. If, however, participants made their judgments as we had intended in Experiments 1a–1b, then they would have no particular basis (or perhaps many different bases) for making judgments in Experiment 2. We would therefore *not* expect to replicate the findings of Experiments 1a-b in this experiment.

(13) Assume you can say these sentences [same as in experimental condition]:
The *zask* the *nop* toast-*pe.*
The *vash* the *yerd* crease-*pe.*
The *blib* the *nalf* slap-*pe.*
How likely is it that you can also say: Clip-*to* the *isp* the *bliz.* ?
Answer: __%

4.1. *Participants*

A total of 57 new participants completed the survey. Seven participants were automatically removed for using a single value for their likelihood estimates, as in Experiments 1a–1b. The remaining 50 participants were included in the analysis.

4.2. *Results*

Results are provided in Table 4.

Table 4. *Z-score means from manipulation check involving target constructions that are distinct from attested instances. Standard errors are in parentheses.*

	Low Variability	High Variability
1 instance	−.02 (.10)	N/A
3 instances	.03 (.07)	.06 (.09)
6 instances	−.01 (.07)	−.06 (.09)

Reassuringly, results across conditions are not significantly different from one another. A 2 × 2 repeated measures ANOVA found no main effect for either variability, $F(1,46) = .02$ ($p = .90$), or type frequency, $F(1,46) = 1.01$ ($p = .32$), and no significant interaction between the two ($p = .61$). A second three level ANOVA that includes type frequency also found no effect of type frequency, $F(2,75) = .30$, ($p = .70$.). These null results are what we expect if, as intended, speakers were in fact providing productivity judgments in Experiments 1a-b, but were unable to do so in Experiment 2, since the target construction was always different from the construction suggested by the instances. The manipulation check provides evidence that participants did not simply use a similarity heuristic without regard to the issue of grammatical productivity.

Results also demonstrated that the overall non-normed mean in the second experiment ($M = 43.4$) was significantly lower than that of the first experiment ($M = 56.2$), $F(1,90) = 7.71$ ($p = .007$). This is expected if participants in fact had less confidence overall in judging target constructions that involved a different construction from that used in the attested instances.

To summarize, the results of Experiment 2 suggest that speakers did not judge the likelihood of target coinages in any systematic way. They recognized that the construction used in the target coinage was not the same as that in the attested instances. Therefore they had no basis for deciding the likelihood of the coinage. This contrasts with the findings in both Experiment 1a and 1b, in which the target coinage and the attested instances were instances of the same general construction (same word order and same verbal morpheme). These facts lend credence to the claim that participants in Experiments 1a and 1b were indeed providing likelihood ratings about the productivity of *constructions*.

5. Experiment 3: Similarity of target coinage to attested instances and variability

In Experiments 1a–1b, we varied how similar the members of a set of verbs were to each other, but we held constant the similarity between the newly coined target instance and its closest neighbor. In Experiments 1a–1b, we found that increased variability served to increase the confidence that a construction could be generalized. However, one might imagine that increased variability may not always benefit generalizations. In particular, if increased variability served to make attested instances less similar to the target coinage, it might have the effect of dampening generalization. Moreover, it seems likely that there may be limits to the effect of variability. If instances are variable but nonetheless still very different from the target coinage, the variability may play

less of a role. Experiment 3 investigates the roles of variability and *similarity* in judgments of productivity.

In a within-subjects 3 × 2 design, we included three levels of similarity of coinage to the closest attested neighbor and two levels of variability, as illustrated in Figure 2:

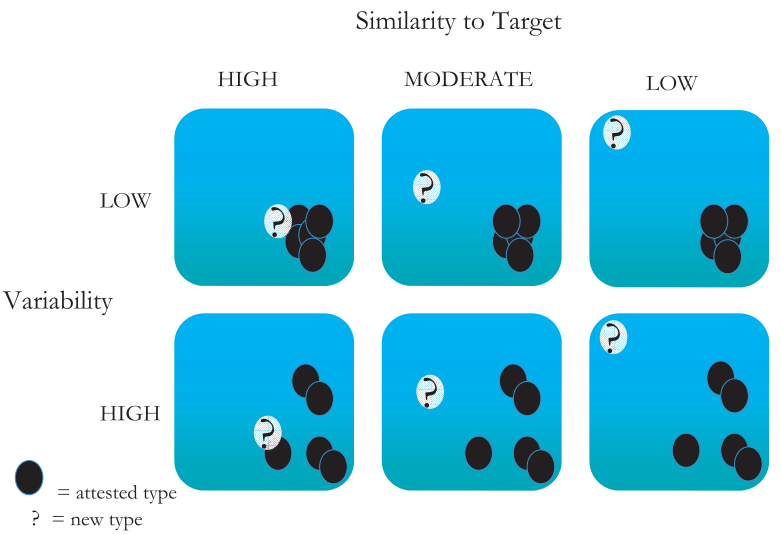

Figure 2. *Schematic representation of the six conditions in Experiment 2. Type frequency was held constant while maximum similarity and degree of variability were manipulated.*

5.1. *Participants*

A total of 55 new participants completed the survey. Five participants were automatically removed for using a single value for their likelihood estimates. The remaining 50 participants were included in the analysis.

5.2. *Procedure*

There were six conditions and three items in each condition for a total of 18 stimuli sets used in each questionnaire. As in Experiments 1a-b, each set of stimuli involved the same construction (same word order and same sentence final particle); each different set of stimuli involved a different construction. Also as in Experiments 1a–1b, four different lists were created, in an effort to

avoid item effects. Each verb class appeared once in each similarity condition across one of the four lists.

For the attested instances, verbs for each stimuli set were randomly chosen from a subset of the same nine semantic classes that were used in Experiment 1a–1b. The target coinages were chosen as follows. High similarity target verbs came from the same verb class as *one* of the attested instances in the high variability condition, and from the same verb class as *all* attested instances in the low variability condition (since all five attested instances came from the same verb class in the latter condition). The targets were created by randomly selecting one of the verbs from the high type frequency stimuli sets from Experiments 1a–1b.

Medium similarity target verbs came from the same verb classes used in Experiments 1a–1b for target instances. All attested and target verbs designated concrete actions (see Table 1 for classes); the target verbs came from some verb class distinct from all other attested instances, as in Experiment 1a–1b.

Low similarity target verbs were drawn from a new class of verbs: verbs cognition, including *admire, fear, love, despise, appreciate,* and *pity* (Levin 1993: 191). This class was determined to be maximally dissimilar to all other classes, according to Latent Semantic Analysis measures.

5.3. *Results*

A 3 × 2 repeated measures ANOVA between maximum similarity and variability revealed a significant main effect of similarity, $F(2,98) = 26.81$ ($p < .001$). That is, coinages that were more similar to an attested instance were judged more likely to be acceptable.

There was no main effect for variability, $F(1,49) = .37$ ($p = .55$), but results demonstrated an interaction between variability and similarity, $F(2,98) = 13.46$ ($p < .001$). In the medium similarity condition (the middle column in Figure 2), participants were significantly more confident that the target instances were acceptable if attested instances were more varied, $F(1,49) = 9.87$ ($p = .001$). This result replicates a finding in Experiments 1a-b: when the target instance is from a different, but not wholly unrelated verb class than the attested instances, participants are more confident in a coinage when the construction has been witnessed with a greater variety of verbs. Intriguingly, we find an effect in the opposite direction when the target instance is from the same verb class as its closest neighbor (see the first column in Figure 2). In particular, the high similarity condition was rated as *less* acceptable when the attested instances had high variability, $F(1,49) = 20.13$ ($p < .001$). In the low similarity conditions (the third column in Figure 2), variability was not a significant factor, $F(1,49) = .03$ ($p = .86$).

Condition means are provided in Figure 3:

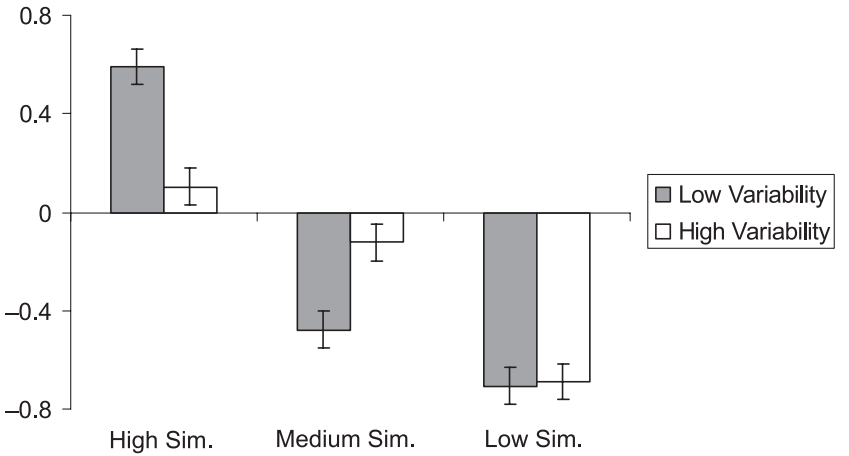

Figure 3. *Condition z-score means for Experiment 2 comparing high and low variability with low, medium or high similarity between target coinage and attested instances.*

The numerical data are provided in Table 5:

Table 5. *Z-score mean likelihood estimates when maximum similarity and variability are manipulated.*

	High Similarity	Medium Similarity	Low Similarity
High Variability	.10 (.07)	−.12 (.07)	−.68 (.07)
Low Variability	.59 (.07)	−.48 (.07)	−.71 (.08)

Let us consider each of the three levels of semantic similarity depicted as columns in Table 3. The medium similarity condition — in which the target is not too close and not too far from any attested instances — used the same degree of similarity as in Experiments 1a and 1b; the finding that variability serves to encourage generalization was replicated. In the high similarity case, the target coinage is already highly similar to at least one of the attested instances, since it is drawn from the *same* semantic class. If the pattern is highly variable, other attested instances are necessarily less similar to the coinage, since they are drawn from other semantic classes. If the pattern is not variable, the target coinage is from the same semantic class as *all* of the attested instances; clearly there is an advantage to that situation, and we correspondingly find a significant advantage for low variability when similarity to at least one of the instances is high.

On the other side of the similarity scale, there is no effect of variability, i.e., if the coinage is sufficiently dissimilar to all of the attested items, the variability among attested items does not have a significant effect on confidence ratings, perhaps because the attested instances are judged to be irrelevant to the target coinage (Medin et al. 2003).

6. Coverage

A factor of COVERAGE can be used to make sense of the variability and type frequency findings of Experiments 1a–1b, as well as the interaction of variability and similarity in Experiment 3. Coverage is defined as the degree to which attested instances "cover" the category determined jointly by attested instances together with the target coinage; coverage is high if the target coinage falls within a dense neighborhood; coverage is low if the space defined by the coinage together with the attested instances is mostly empty. Coverage is a gradient factor and is determined by the degree to which the attested instances are similar to members of the lowest level category that includes both the coinage and the attested instances.

The factor of coverage has the advantage that it has been demonstrated to be relevant to categorical induction more generally (Osherson et al. 1990; Sloman 1993). In the case of language, the degree to which previously attested instances fill the semantic or phonological space that includes the potential target instance, the more confident speakers will be in using the target instance (Goldberg 2006: 98).[6] Coverage is related to the density of a neighborhood (cf. also Zeschel 2010; Zeschel and Bildhauer 2009), but takes into account the relationship between the coinage and the neighborhood.

Coverage is thus, importantly, not a direct function of either type frequency of attested instances or of variability of attested instances in isolation. Rather it is determined by the relationship *between* the coinage and attested instances.

Osherson and colleagues demonstrate several properties that follow from coverage. For example, speakers are more confident of the soundness of the conclusion in (A), that rabbits have some given property ("X"), than they are of the same conclusion in (B):[7]

(A) Assumption 1: Lions have property X
Assumption 2: Giraffes have property X
Conclusion: Rabbits have property X

(B) Assumption 1: Lions have property X
Assumption 2: Tigers have property X
Conclusion: Rabbits have property X

Intuitively this is because the assumptions in (B), that lions and tigers have some property X, tells us something only about large felines and says nothing at all about rabbits, whereas the assumptions in (A) lead us to suspect that the generalization may hold of all mammals, in which case there is more reason to believe the property holds also of rabbits as an instance of the category mammal. This property corresponds to variability (their "premise diversity").

Osherson et al. also observe that more inclusive sets of premises lead to more secure inductive inferences, as long as the additional premises are drawn from the same category defined by the original instances together with the conclusion (their "premise monotonicity"). Thus (D) below can be seen to be a stronger argument than (C). In (C), the smallest category that includes foxes, pigs, and gorillas is animal, and this category clearly includes wolves. This observation corresponds to the type frequency factor varied in the present Experiments 1a–1b.

(C) Assumption 1: Foxes have property Y
Assumption 2: Pigs have property Y
Conclusion: Gorillas have property Y

(D) Assumption 1: Foxes have property Y
Assumption 2: Pigs have property Y
Assumption 3: Wolves have property Y
Conclusion: Gorillas have property Y

At the same time, if new premises are added that *extend* the category created by the previous premises and the conclusion, then the new premises can serve to weaken a conclusion. Osherson et al. note that the first judgment below is stronger than the second one (1990: 188, exs. 14a, 14b), since flies, moths, and bees in (E) determine a category of insects, whereas the additional premise in (F) about wolves requires the category be extended more broadly with a resultant loss in coverage.

(E) Assumption 1: Flies have property Z
Assumption 2: Moths have property Z
Conclusion: Bees have property Z

(F) Assumption 1: Flies have property Z
Assumption 2: Moths have property Z
Assumption 3: Wolves have property Z
Conclusion: Bees have property Z

This idea can be used to make sense of the fact that variability decreased confidence ratings of coinages in the high similarity condition in Experiment 3. In that case, variability served to extend the required category well beyond what would otherwise be needed (if all instances and the coinage came from the same semantic class). The result was that attested instances covered the category less well in the high similarity-high variability condition. In other words, if the coinage is from the same category as all of the attested instances (high similarity), coverage is already maximized; higher variability can only serve to spread the category out, reducing coverage.

Recall that there was a lack of a significant difference between high variability and low variability when similarity between the attested instances and the coinage was low. In terms of coverage, the coinage was at some remove from the attested items, which were relatively alike (to more or less degree). Therefore, the category that contained all attested items and the target coinage would necessarily be mostly empty. Correspondingly, whether the cluster of attested instances was tight (less variable) or less tight (more variable) did not have an effect. No difference is what we would predict if the lowest level category that included both the attested instance and novel coinage was not increased by higher variability. It seems that cognition verbs were sufficiently distinct from each of the verb classes independently, so that including instances from more than one verb class did not significantly increase overall coverage.

Coverage can thus explain the interaction between variability and similarity in Experiment 3. In the medium similarity condition, and only in this condition, the category created by the combination of attested instances and target coinage covers the semantic space better when the attested instances are varied.

To summarize, the present work provides compelling evidence in favor of an expected interaction between variability and similarity, as well a main effect of type frequency. We have argued that these findings can be combined if we recognize them as lending supportive evidence to the notion of *coverage* as an important factor that affects a construction's productivity. Judgments about productivity, on this view, are akin to non-linguistic induction.

7. Similarity and induction

Recall that a clear main effect was found in Experiment 3 when similarity was systematically manipulated: the closer the coinage was to an attested instance, the more secure participants' judgments were found to be. Thus similarity to an attested instance appears to be an additional factor beyond coverage (cf. also Osherson et al. 1990).

Similarity is not entirely independent of coverage: if a category has sufficiently high coverage, there will be some instance that *is* similar to the coinage. However, the converse is not necessarily true: it is possible that a coinage will be similar to one attested instance and yet the overall category may have low coverage (because other attested instances are too few and/or too spread out).

Following Osherson et al. (1990), we assume that similarity and coverage are complementary variables; some speakers may well weight one relatively more than the other. The present experiments serve only as a starting point; clearly more work investigating these two factors is needed.

Similarity is also related to statistical preemption. In the limiting case, increased similarity becomes identity. In this case, in which a coinage has an *identical* function to that of an attested item, productivity is actually not encouraged but inhibited, due to statistical preemption. Speakers generally formulate their utterances in conventional ways (see Section 1.2).

Before leaving the topic of similarity, it behooves us to mention the fact that difficulties arise when one attempts to measure similarity objectively. It is important to keep in mind that in real world contexts, similarity is largely determined by context and by the comparison group involved (Kalyan 2011; Medin et al. 1993; Tversky 1977).

In determining whether a new word can fill a slot in a construction, the function of the entire construction is taken into account, not just the similarity of the words in isolation. An illustrative case comes from a Hebrew construction, discussed by Meir (in preparation), involving what was originally the fixed idiom provided in (14):

(14)	*'ibed*	*'et*	*'acmo*	*la-da'at*
	to destroy(biblical)/to lose(present)	ACC	self	to knowledge
	"to commit suicide"			

Meir observes that the verb slot can nowadays be filled by many different verbs, with glosses like:

(15) "The Haaretz newspaper recycles itself to knowledge"
(= destroying itself by using the same material over and over again)
(16) "The Supreme Court humiliated itself to knowledge."
(17) "[Some reporter] documented himself to knowledge."
(18) "The prime minister chattered/babbled himself to knowledge."

The implication in all of the examples (15)–(18) is that some action is performed despite an awareness of the harm it will bring to oneself. While "recycle," "humiliate," "document," and "chatter" cannot be said to be particularly similar to the original verb "destroy/lose" or to each other, the overall meaning

of the construction V_X*'et 'acmo la-da'at,* "to intentionally harm oneself by doing X" provides the relevant dimension of similarity.

In the present studies, participants only knew that the verbs in question appeared in a two-argument construction. The construction was not given a gloss and the nonsense nouns did not provide any clues, other than the basic fact that the construction has two arguments. That is why judgments will depend on the similarity involved between the verbs of the attested instances. But it is important to bear in mind that ultimately similarity must be determined with respect to the construction's meaning, and not the verbs' meaning in isolation.[8]

8. Conclusion

We began by acknowledging that semantic sensicality and statistical preemption play critical roles in determining whether constructions are productive; i.e., whether words can be coerced by constructions. In particular:

(a) A novel coinage must be semantically interpretable.
(b) A novel coinage cannot be statistically preempted by some other formulation that is semantically and pragmatically equivalent or preferred in the context of use.

The results reported here provide experimental evidence for two additional factors that affect speakers' confidence levels in accepting a target coinage: similarity and coverage. Tellingly, these same two factors have been argued to be relevant to induction more generally (Osherson et al. 1990; Sloman 1993). We take this to lend support to the idea that linguistic categories are categories, and that productivity depends on general properties of induction. More specifically we found that:

(c)–(d) Speakers are more confident of a target coinage when the category formed by relating attested instances and the target coinage to the extent that the coinage and the attested instances jointly *cover* the semantic (or phonological) space that includes them. *Coverage* was supported by combined effects of type frequency and variability (variability of attested instances) [Experiments 1a–1b] and a predicted interaction between similarity and variability [Experiment 3].
(e) Speakers are more confident of a target coinage when the coinage is more similar to an attested instance. [Experiment 3].

Experiment 2, moreover, demonstrated that participants were performing judgments about grammatical constructions, since all effects disappeared if different constructions were used in the attested instances on the one hand and the target coinage on the other. This goes some way toward ensuring that the task is appropriate to the topic we set out to explore: grammatical productivity.

Inductive inferences often involve general knowledge and causal inferences; people do not reason in a vacuum or in a "neutral" context (Heit 2000; Lopez et al. 1997; Medin et al. 2003). The present experimental work focused on similarity and coverage by manipulating variability, similarity, and type frequency. General knowledge and context effects were minimized by restricting knowledge about the construction and the fictitious "Zargotian" language; ultimately, such context effects will need to be taken into account.

In future work, it will also be worth exploring online measures instead of the explicit judgment task used in the present experiments. Moreover, it should be possible to quantify the contribution of various factors. The present work is only a beginning. However, it does lend support to the idea that morphological productivity and grammatical productivity are closely related, since the same factors recognized in morphology have been shown to play a role in the productivity of constructions (cf. also Barðdal 2008). This work also lends support to the growing trend toward treating knowledge of language as knowledge (Goldberg 1995; Lakoff 1987), wherein general principles of categorization and induction play central roles (cf. also e.g., Ambridge et al. 2006; Boyd et al. 2009; Casenhiser and Goldberg 2005; Langacker 1987; Tomasello 2003).

Received 2 April 2010 *Princeton University*
Revised version received 5 January 2011

Appendix 1. Mixed Linear Model Analyses of Results

At the request of one of the reviewers, we conducted mixed linear model analyses in addition to the ANOVAs reported in the main body of the text. While ANOVAs are more familiar to many (including us), mixed models are emerging as the preferred statistic in certain quarters. The findings are essentially comparable. We used raw scores instead of z-scores, treating subjects and items as random effects. We also included the variation in factors by subject when appropriate (see Experiment 1). Recall that items involve sets of sentences of "Zargotian" of varying types, together with a target sentence. We provide the best-fit models for each experiment in turn.

Experiment 1: Type frequency and Variability of attested instances

We first compared Type frequency and Variability for all 100 participants of Experiments 1A and 1B (recall 1B was simply a replication of 1A). As before, the first analyses exclude the 1-item type frequency condition, since the 3- and 6-item conditions had both low and high variability, but it is necessary to have low variability with a type frequency of 1. Table 6 reports the model for independent effects of Type frequency and Variability. The interaction between the two did not significantly contribute to the model and so was excluded.

Table 6. *Mixed linear models analysis for variability and type frequency in Experiment 1.*

	Fixed Effects				
	Estimate	Std. Error	*t* value	*p* value	Sig.
Intercept	26.45	1.766	14.98	<.001	
Variability	2.305	0.934	2.467	0.01	**
Type Frequency	0.313	0.284	1.102	0.27	

Random Effects	
Groups	Std. Dev.
Subject	10.62
Subject × Variability	3.851

Thus we find, as before, that variability significantly increases the likelihood scores ($t = 2.467$, $p = .01$), and there is no interaction between type frequency and variability. There was no significant random effect of Item, suggesting that subjects treated items of the same type essentially the same. There was a significant random effect of Subject, and a significant random interaction of Variability by Subject. These effects suggest that each subject used the probability estimation space somewhat differently. This is to be expected since we offered a range of 0–100; it is the reason we used z-scores in the ANOVA analyses.

We do not find an effect of type frequency when comparing items of 3 and 6 exemplars ($p = .27$). But as we did in the ANOVA analysis, we further investigated a possible effect of type frequency by including the type frequency of 1 condition in the analysis, along type frequencies of 3 and 6 (averaging low and high variability). As shown in Table 7, there was a significant effect of type frequency, $t = 3.596$ ($p = .0003$). This time, both random effects of Item and Subject were significant contributors to the model.

Table 7. *Mixed linear models analysis for type frequency in Experiment 1, comparing effects of type frequency 1, and average scores for type frequency 3 and 6.*

	Fixed Effects				
	Estimate	Std. Error	*t* value	*p* value	Sig.
Intercept	50.743	2.93	17.319	<.0001	
Type Frequency	2.168	0.603	3.596	0.0003	***

Random Effects	
Groups	Std. Dev.
Subject	17.10
Item	6.216

Experiment 2: Control

A mixed linear model is presented in Table 8 for the data from the control experiment. As expected, and as was found with the ANOVA analysis, there were no significant effects. A model not including the interaction term is not shown below, but it also showed no significant effects. Note that there were still significant random effects, suggesting that subjects responded to the control condition in differing ways as is to be expected.

Table 8. *Mixed linear models analysis for variability and type frequency in Experiment 2*

	Fixed Effects			
	Estimate	Std. Error	*t* value	*p* value
Intercept	47.31	4.390	8.804	<.0001
Type Frequency	−0.944	0.599	−1.58	0.12
Variability	1.093	1.796	0.609	0.54

Random Effects	
Groups	Std. Dev.
Subject	22.81
Subject × Variability	0.0001

Experiment 3: Variability and similarity to target item

In the final experiment, recall that type frequency was held constant while Variability and Similarity were manipulated. Because Similarity is a three-way categorical variable between low, medium, and high similarity, the data was coded so that low similarity was the baseline and the effects of medium and high similarity were dummy coded as separated variables. Any effects of Similarity then are based on comparing medium or high similarity to low similarity.

There was a significant increase in likelihood scores from low to medium similarity and from low to high similarity, both significant effects ($p = .03$ and $p < .0001$, respectively). The difference between low and high variability was not significant ($p = .65$); however, the interaction between Variability and Similarity was significant at both levels of Similarity. This is the same interactive effect we saw using ANOVAs (recall Figure 3).

Table 9. *Mixed linear models analysis for similarity and variability in Experiment 3*

	Fixed effects				
	Estimate	Std. Error	*t* value	*p* value	Sig.
Intercept	46.82	3.578	13.09	<.0001	
Variability	−1.629	3.642	−0.447	0.65	
Similarity Medium	7.771	3.656	2.125	0.03	*
Similarity High	35.54	3.656	9.720	<.0001	***
Variability × Similarity Medium	8.475	5.159	1.643	0.1	+
Variability × Similarity High	−12.21	5.159	−2.367	0.02	*

Random effects	
Group	Std. Dev.
Item	4.9072
Subject	17.0933

In short, the mixed model analyses confirm the ANOVA analyses in the main text.

Appendix 2. Example stimuli for the experiments

Example stimuli for Experiments 1a–1b

(1) The *tob* grip-*bo* the *drak*.
The *olf* tear-*bo* the *jark*.
The *solk* rumple-*bo* the *quix*.

The *grib* pack-*bo* the *orf*.

(2) The *alk* fling-*ti* the *burd*.
The *kwap* hurl-*ti* the *tirp*.
The *ralb* pass-*ti* the *clib*.

The *knib* fold-*ti* the *gorp*.

(3) The *loof* hit-*nu* the *morb*.
The *feg* grab-*nu* the *telk*.
The *rit* bake-*nu* the *ilp*.
The *erg* acquire-*nu* the *lin*.
The *siff* heat-*nu* the *glick*.
The *crof* bump-*nu* the *pok*.

The *siln* rip-*nu* the *urm*.

(4) The *hast* the *farg* fold-*ga*.

The *ulb* the *flim* crowd-*ga*.

(5) The *irb* the *frum* buy-*lu*.
The *mab* the *shap* order-*lu*.
The *jirm* the *talm* borrow-*lu*.

The *craz* the *trum* smash-*lu*.

(6) The *frim* the *poft* crease-*ze*.
The *ferb* the *malk* rip-*ze*.
The *vew* the *phox* clasp-*ze*.
The *smek* the *zask* bend-*ze*.
The *flir* the *nirm* grasp-*ze*.
The *wilk* the *rab* break-*ze*.

The *nep* the *sleb* pile-*ze*.

(7) Pass-*ko* the *zot* the *shum*.
Crease-*ko* the *hom* the *dirf*.

(8) The *lerm* the *pret* grab-*xa*.
The *holm* the *terk* buy-*xa*.
The *dax* the *reg* acquire-*xa*.
The *kasp* the *wib* get-*xa*.
The *hap* the *rew* order-*xa*.
The *larn* the *wuft* borrow-*xa*.

The *merb* the *shrop* bend-*xa*.

(9) The varg the *kod* clip-*ma*.
The *arb* the *folp* throw-*ma*.
The *rup* the *yorg* pile-*ma*.

The *yalp* the *blib* crumple-*ma*.

(10) Crinkle-*pe* the *vash* the *yik*.
Fold-*pe* the *yol* the *laf*.
Crumple-*pe* the *grap* the *vip*.

Cram-*pe* the *tak* the *kep*.

(11) The *rull* pitch-*ju* the *yig*.
The *nalf* fling-*ju* the *woz*.
The *yerd* hurl-*ju* the *zast*.
The *elp* flip-*ju* the *drib*.
The *borb* throw-*ju* the *quelp*.
The *urd* pass-*ju* the *rark*.

The *jorp* crinkle-*ju* the *yeb*.

(12) The *skoo* buy-*de* the *zib*.

The *neld* crush-*de* the *bup*.

(13) Flip-*wi* the *hib* the *bix*.
Load-*wi* the *fomp* the *yeep*.
Snip-*wi* the *zolk* the *nop*.
Pitch-*wi* the *vork* the *krig*.
Scrape-*wi* the *pulg* the *geed*.
Pack-*wi* the *isp* the *yib*.

Rumple-*wi* the *yob* the *bamp*.

(14) Get-*fo* the *slom* the *ern*.
Fry-*fo* the *erst* the *voft*.
Knock-*fo* the *clat* the *herl*.

Shatter-*fo* the *cloop* the *voy*.

(15) Crumple-*yi* the *smoll* the *vilm*.
Fold-*yi* the *bliz* the *ank*.
Crease-*yi* the *lom* the *pid*.
Rumple-*yi* the *knorf* the *moop*.
Crinkle-*yi* the *bolg* the *bish*.
Bend-*yi* the *grap* the *rull*.

Heap-*yi* the *caf* the *elt*.

Example stimuli for Experiment 2

(1) The *rolm* the *erst* borrow-*xe*.
The *xazz* the *yeep* order-*xe*.
The *zug* the *nell* buy-*xe*.

Cram-*zo* the *larn* the *jure*.

(2) The *girt* pass-*sha* the *pid*.
The *solk* hurl-*sha* the *zast*.
The *kep* flip-*sha* the *nop*.
The *voft* fling-*sha* the *feg*.
The *zot* pitch-*sha* the *kiw*.
The *isp* throw-*sha* the *lerm*.

The *hiff* the *poft* crinkle-*fu*.

(3) The *rew* the *pret* fold-*chu*.
The *hom* crush-*ve* the *gop*.

(4) Fry-*ko* the *mab* the *tak*.
Knock-*ko* the *ast* the *slim*.
Get-*ko* the *xalt* the *jex*.

The *jark* the *yunk* pack-*tu*.

(5) The *geed* hurl-*de* the *nome*.
The *grib* fling-*de* the *pok*.
The *ulb* pass-*de* the *holm*.

The *yaft* the *ruge* fold-*nu*.

(6) The *quard* tear-*yi* the *fuft*.
The *saft* rumple-*yi* the *oll*.
The *rup* grip-*yi* the *clib*.

Shatter-*pe* the *yig* the *rark*.

(7) The *herl* the *smoll* get-*lu*.
The *yib* the *lutt* borrow-*lu*.
The *jib* the *quig* order-*lu*.
The *moff* the *misp* buy-*lu*.
The *glick* the *dirf* grab-*lu*.
The *celk* the *xam* acquire-*lu*.

The *knib* heap-*hi* the *vip*.

(8) Fold-*qo* the *nud* the *hap*.
Crumple-*qo* the *burd* the *bliz*.
Crinkle-*qo* the *noof* the *bamp*.

The *krig* smash-*ri* the *cerl*.

(9) Pack-*ma* the *bolg* the *nug*.
Flip-*ma* the *bup* the *jace*.
Scrape-*ma* the *muce* the *vuss*.
Pitch-*ma* the *zep* the *ern*.
Load-*ma* the *rit* the *jorp*.
Snip-*ma* the *baff* the *woz*.

The *pax* the *hult* rumple-*fo*.

(10) The *tusp* buy-*sa* the *shum*.

The *lin* the *fuge* crowd-*ni*.

(11) Crumple-*xa* the *yeb* the *imp*.

Fold-*xa* the *nit* the *knorf*.
Bend-*xa* the *het* the *wiss*.
Crinkle-*xa* the *cherb* the *cimp*.
Crease-*xa* the *bix* the *yorg*.
Rumple-*xa* the *flir* the *cloop*.

The *smek* bend-*ze* the *hurft*.

(12) The *deet* the *yob* clip-*bo*.
The *olf* the *carm* pile-*bo*.
The *morb* the *cink* throw-*bo*.

Crumple-*ka* the *goff* the *vew*.

(13) The *neld* heat-*ga* the *wilk*.
The *belg* acquire-*ga* the *dut*.
The *pheb* bump-*ga* the *zib*.
The *erg* grab-*ga* the *fazz*.
The *tirp* bake-*ga* the *dulk*.
The *irb* hit-*ga* the *nam*.

Pile-*tho* the *skoo* the *gare*.

(14) Pass-*wi* the *terk* the *welf*.

The *folp* crease-*ba* the *nirm*.

(15) The *suce* the *quelp* bend-*ju*.
The *paff* the *deg* rip-*ju*.
The *alk* the *urd* grasp-*ju*.
The *pum* the *bish* crease-*ju*.
The *hib* the *kasp* break-*ju*.
The *zolk* the *vash* clasp-*ju*.

Rip-*re* the *fomp* the *pham*.

Example stimuli for Experiment 3

(1) The *gare* the *merb* crumple-*nu*.
The *yeb* the *gulst* pass-*nu*.
The *jorp* the *smek* rumple-*nu*.
The *nalf* the *solk* borrow-*nu*.
The *peln* the *jark* buy-*nu*.

The *knib* the *nop* throw-*nu*.

(2) Hurl-*ko* the *mab* the *roff*.
Tear-*ko* the *het* the *celk*.
Heat-*ko* the *berz* the *zast*.
Smash-*ko* the *feg* the *sleb*.
Flip-*ko* the *leff* the *shap*.

Admire-*ko* the *suce* the *oll*.

(3) The *cit* the *arg* break-*lu*.
The *vilm* the *pret* grasp-*lu*.
The *paff* the *tak* bend-*lu*.
The *urm* the *gop* crease-*lu*.
The *yob* the *goff* clasp-*lu*.

The *siln* the *nud* appreciate-*lu*.

(4) Bump-*xe* the *dop* the *hap*.
Grab-*xe* the *jeg* the *nop*.
Hit-*xe* the *pum* the *rew*.
Bake-*xe* the *yol* the *whep*.
Acquire-*xe* the *pok* the *yold*.

Rip-*xe* the *urm* the *wib*.

(5) The *jirm* the *yib* cut-*yi*.
The *tirp* the *pid* crumple-*yi*.
The *cherb* the *knib* grasp-*yi*.
The *rull* the *kig* grip-*yi*.
The *shum* the *hib* fold-*yi*.

The *dirf* the *ferb* snip-*yi.*

(6) Scrape-*ga* the *warl* the *vuss*.
Kick-*ga* the *larn* the *bolg*.
Knock-*ga* the *nud* the *tusp*.
Shatter-*ga* the *orf* the *bup*.
Smash-*ga* the *leff* the *frum*.

Pity-*ga* the *geed* the *hult*.

(7) The *yeep* clutch-*ma* the *welf*.
The *pax* handle-*ma* the *suce*.
The *xam* hold-*ma* the *nep*.
The *roff* grip-*ma* the *dulk*.
The *hiff* grasp-*ma* the *vip*.

The *moff* heap-*ma* the *grap*.

(8) The *voft* the *urd* pitch-*ve*.
The *hink* the *quard* throw-*ve*.
The *irb* the *borb* hurl-*ve*.
The *hult* the *libe* fling-*ve*.
The *quix* the *dulk* flip-*ve*.

The *bamp* the *solk* crinkle-*ve*.

(9) The *flim* buy-*sha* the *voy*.
The *nome* pack-*sha* the *phox*.
The *ilp* smack-*sha* the *kwap*.
The *hink* load-*sha* the *cimp*.
The *wiss* order-*sha* the *ilp*.

The *nirm* bake-*sha* the *jib*.

(10) Clutch-*hi* the *misp* the *cholk*.
Grip-*hi* the *deg* the *ank*.
Fry-*hi* the *lom* the *kep*.
Cram-*hi* the *muce* the *moop*.
Pile-*hi* the *isp* the *voy*.

Clip-*hi* the *herl* the *vash*.

(11) The *fazz* crinkle-*ju* the *ank*.
The *nug* fold-*ju* the *rup*.
The *vork* bend-*ju* the *cug*.
The *dut* rumple-*ju* the *xalt*.
The *krig* crumple-*ju* the *nep*.

The *zast* crease-*ju* the *vash*.

(12) The *knorf* the *fuge* tear-*xa*.
The *noof* the *frim* break-*xa*.
The *poft* the *ruge* smash-*xa*.
The *junt* the *tak* rip-*xa*.
The *rew* the *yol* crush-*xa*.

The *jure* the *pid* shatter-*xa*.

(13) Pack-*de* the *josk* the *pum*.
Pile-*de* the *burd* the *bish*.
Cram-*de* the *pulg* the *holm*.
Load-*de* the *erst* the *junt*.
Crowd-*de* the *cloop* the *cherb*.

Despise-*de* the *vuss* the *wilk*.

(14) The *fuge* the *peln* order-*wi*.
The *zug* the *jell* acquire-*wi*.
The *yorg* the *ast* get-*wi*.
The *whep* the *dax* buy-*wi*.
The *arb* the *yaft* grab-*wi*.

The *xazz* the *smoll* fear-*wi*.

(15) The *ralb* knock-*qo* the *shrop*.
The *kiw* kick-*qo* the *ast*.
The *erst* slap-*qo* the *nug*.
The *clat* hit-*qo* the *yik*.
The *gorp* bump-*qo* the *shap*.

The *zep* smack-*qo* the *goff*.

(16) Get-*fo* the *lerm* the *kwap*.
Order-*fo* the *gulst* the *feg*.
Grab-*fo* the *nell* the *pham*.
Borrow-*fo* the *cink* the *jace*.
Buy-*fo* the *pid* the *gam*.

Love-*fo* the *berz* the *zug*.

(17) The *hobe* crease-*sa* the *drib*.
The *mung* bend-*sa* the *woz*.
The *darf* crinkle-*sa* the *knorf*.
The *deg* crumple-*sa* the *jorp*.
The *quig* fold-*sa* the *gop*.

The *lin* tear-*sa* the *hom*.

(18) The *sleb* flip-*pe* the *bup*.
The *fazz* snip-*pe* the *yunk*.
The *pid* pitch-*pe* the *elt*.
The *zot* scrape-*pe* the *tirp*.
The *hiff* load-*pe* the *poft*.

The *luce* pack-*pe* the *crof*.

Notes

1. We would like to thank Ben Ambridge, Johanna Barðdal, Joan Bybee, Andy Conway, Sam Glucksberg, Siva Kalyan, Phil Johnson Laird, Arne Zeschel and an anonymous reviewer for helpful feedback on an earlier version or a presentation of this work. Harald Baayen is owed a special debt of gratitude for his encouragement and his advice on the mixed model analyses. Correspondence Address: Laura Suttle, Psychology Department, Princeton University, Princeton, NJ 08540, USA. E-mail: lsuttle@princeton.edu.

2. A handful of other verbs violate the pattern, including *need, heed, weed, bead, cede,* and *seed*. These verbs may play a role in dampening the tendency to productively extend the pattern, given that 23% is less than the majority of speakers.
3. To some extent, variability corresponds inversely to Barðdal's (2008) "semantic coherence." That is, the greater the semantic variability among instances, the less "semantic coherence" a pattern may have. We prefer the term variability in order to allow ultimately for phonological variability as well as semantic variability (cf. also in fact Barðdal 2008: 27, who intends this as well, despite referring to semantic coherence). Also, it is important to distinguish the semantic coherence of a construction from the semantic coherence of the verbs that may appear in it; a construction may be highly semantically coherent and yet allow for variable range of main verbs (e.g., in the English *way* construction). See section 7 for more discussion of this issue. Finally, Barðdal argues that greater type frequency necessarily correlates with lower 'semantic coherence' (higher variability). While the two are certainly correlated in natural language samples, we disentangle these two factors in the present work.
4. Bybee (2010: 105) and Clausner and Croft (1997) essentially equate *schematicity* of a pattern with variability of exemplars. The present work aims to determine directly whether variability is a factor in schematicity (which can be defined as abstraction, as evidenced by productivity).

5. Word order of a given argument structure can typically vary in discourse conditioned contexts (e.g., topicalizations, questions, etc.). We therefore altered both the word order *and* the verbal suffix to ensure that speakers would treat the target sentence as an instance of a distinct construction.
6. Clausner and Croft (1997) appear to have a similar idea in mind when they refer to "the proportion of a schemas range that can be instantiated as constructions." Likewise, Barðdal (2008: 43–52, 77) notes that there is an expected "interrelation between type frequency and open schema."
7. The properties X, Y, and Z in the syllogisms that follow can be filled in by a "blank" predicate, i.e., a predicate of which participants have no knowledge, e.g., "have reddish zisopherous in their intestines."
8. This, too, has its analogue in nonlinguistic similarity judgments. Osherson et al. (1990) take pains to note that their evidence is based on "blank" predicates, which offer participants little or no prior domain knowledge (cf. Medin et al. 2003 for further discussion).

References

Ahrens, Kathleen. 1995. *The mental representation of Vverbs*. San Diego: University of California San Diego dissertation.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.

Ambridge, Ben, Anna L. Theakston, Elena V. M. Lieven & Michael Tomasello. 2006. The distributed learning effect for children's acquisition of an abstract syntactic construction. *Cognitive Development* 21(2). 174–193.

Ambridge, Ben, Julian M. Pine, Caroline F. Rowland, Rebecca L. Jones & Victoria Clark. 2009. A semantics-based approach to the "no negative evidence" problem. *Cognitive Science* 33(7). 1301–1316.

Aronoff, Mark. 1976. *Word formation in Generative Grammar*. Cambridge: MIT Press.

Aronoff, Mark. 1983. Potential words, actual words, productivity and frequency. *Proceedings of the 13th International Congress of Linguists*. 163–171.

Baker, C. L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10(4). 533–581.

Barðdal, Jóhanna. 2008. *Productivity: Evidence form case and argument structure in Icelandic*. Amsterdam: John Benjamins.

Barðdal, Jóhanna. 2011. The rise of the dative substitution in the history of Icelandic: A diachronic Construction Grammar approach. *Lingua* 121(1). 60–79.

Bowerman, Melissa. 1988. The "No negative evidence" problem: How do children avoid constructing an overly general grammar? In John A. Hawkins (ed.), *Explaining Language universals*, 73–101. Oxford: Blackwell.

Bowerman, Melissa. 1996. Argument Structure and Learnability: Is a Solution in Sight? *Proceedings of the Berkeley Linguistics Society* 22. 454–468.

Bowerman, Melissa & Soonja Choi. 2001. Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories. In Melissa Bowerman & Stephen C. Levinson (eds.), *Language acquisition and conceptual development*, 475–511. Cambridge: Cambridge University Press.

Boyd, Jeremy, Erin Gottschalk & Adele E. Goldberg. 2009. Linking rule acquisition in novel phrasal constructions. *Language Learning* 59(S1). 64–89.

Boyd, Jeremy & Adele E. Goldberg. 2011. Learning what NOT to say: The role of statistical preemption and categorization in *a*-adjective production. *Language* 87(1). 55–83.

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

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative Aaternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts, *Cognitive Foundations Of Interpretation,* 69–94. Chicago: University of Chicago Press.

Brooks, Patricia J. & Michael Tomasello. 1999. How children constrain their argument structure constructions. *Language* 75(4). 720–738.

Brooks, Patricia J. & Otto Zizak. 2002. Does preemption help children learn verb transitivity? *Journal of Child Language* 29(4). 759–781.

Brown, Roger & Camille Hanlon. 1970. Derivational complexity and order of acquisition in child speech. In John R. Hayes (ed.), *Cognition and the development of language*, 11–53. New York: Wiley.

Buhrmester, Michael D., Tracy Kwang, & Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? *Perspectives on Psychological Science* 6(1). 3–5.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Bybee, Joan L. 1995. Regular Morphology and the Lexicon. *Language and Cognitive Processes* 10(5). 425–455.

Bybee, Joan L. & David Eddington. 2006. A usage-based approach to spanish verbs of 'becoming'. *Language* 82(2). 323–355.

Bybee, Joan L. 2010. *Language, usage and cognition.* Cambridge: Cambridge University Press.

Casenhiser, Devin & Adele E. Goldberg. 2005. Fast mapping of a phrasal form and meaning. *Developmental Science* 8(6). 500–508.

Chouinard, Michelle & Eve V. Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language* 30(3). 637–669.

Clark, Eve V. 1987. The principle of contrast: A constraint on language acquisition. In Brian MacWhinney (ed.), *Mechanisms of language acquisition*, 1–33. Hillsdale, NJ: Lawrence Erlbaum.

Clausner, Timothy C. & William Croft. 1997. Productivity and schematicity in Metaphors. *Cognitive Science* 21(3). 247–282.

Cruse, D. Alan & William Croft. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Foraker, Stephani, Terry Regier, Naveen Khetarpal, Amy Perfors & Joshua B. Tenenbaum. 2007. Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science* 33(2). 287–300.

Goldberg, Adele E. 1993. Another Look at Some Learnability Paradoxes. *Proceedings of the 25th Annual Stanford Child Language Research Forum,* 60–75. Stanford: CSLI Publications.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar approach to argument structure.* Chicago: Chicago University Press.

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Goldberg, Adele E. 2011. Corpus evidence of the viability of statistical preemption. *Cognitive Linguistics* 22(1). 131–153.

Green, Georgia M. 1974. *Semantics and syntactic regularity*. Indiana University Press.

Heit, Evan. 2000. Properties of inductive reasoning. *Psychonomic Bulletin & Review* 7(4). 569–592.

Jackendoff, Ray. 1997. *Architecture of the language faculty*. Cambridge: MIT Press.

Janda, R. D. 1990. Frequency, markedness and morphological change: On predicting the spread of noun-plural *-s* in Modern High German and West Germanic. *Proceedings of the Eastern States Conference on Linguistics (ESCOL) 7*. 136–153.

Kalyan, Siva. to appear. Similarity in linguistic categorization. *Cognitive Linguistics*.
Kiparsky, Paul. 1982. Lexical morphology and phonology. *Linguistics in the morning calm: Selected papers from SICOL*. 3–91.
Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.
Landauer, Thomas K. & Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review* 104(2). 211–240.
Langacker, Ronald W. 1987. *Foundations of cognitive grammar: volume I*. Stanford: Stanford University Press.
Levin, Beth. 1993. *English verb classes and alternations*. Chicago: Chicago University Press.
Lopez, Alejandro, Scott Atran, John D. Coley, Douglas L. Medin & Edward E. Smith. 1997. The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology* 32(3). 251–295.
Marcotte, Jean-Philippe. 2005. Causative alternation errors and innate knowledge: Consequences of the 'no negative evidence' fallacy. In Eve V. Clark and Barbara F. Kelly (eds.), *Constructions in acquisition*, 205–232. Stanford: CSLI Publications.
Marcus, Gary F. 1993. Negative evidence in language acquisition. *Cognition* 46(1). 53–85.
Medin, Douglas L., Robert L. Goldstone & Dedre Gentner. 1993. Respects for similarity. *Psychological Review* 100(2). 254–278.
Medin, Douglas L., John D. Coley, Gert Storms & Brett K. Hayes. 2003. A relevance theory of induction. *Psychonomic Bulletin & Review* 10(3). 517–532.
Meir, Irit. in preparation. A Hebrew idiom: "To lose oneself in knowledge". Unpublished manuscript Haifa University.
Michaelis, Laura. 2004. Implicit and explicit type-shifting in Construction Grammar. *Cognitive Linguistics* 15(1). 1–67.
Osherson, Daniel N., Ormond Wilkie, Edward E. Smith, Alejandro Lopez & Eldar Shafir. 1990. Category based induction. *Psychological Review* 97(2). 185–200.
Paolacci, Gabriele, Jesse Chandler & Panagiotis Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and decision Making* 5(5). 411–419.
Pinker, Steven. 1981. Comments on the paper by Wexler. In Carl L. Baker & John J. McCarthy (eds.), *The Logical problem of language acquisition*. Cambridge, MA: Harvard University Press. 53–63.
Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, MA: MIT Press.
Rappaport Hovav, Malka & Beth Levin. 2005. Are dative verbs polysemous? Paper presented at the Linguistics Colloquium, Princeton University, 17 February.
Sloman, Steven A. 1993. Feature-based induction. *Cognitive Psychology* 25(2). 231–280.
Tomasello, Michael. 2003. *Constructing a language: A Usage-Based Theory of Language Acquisition*. Boston: Harvard University Press.
Tversky, Amos. 1977. Features of Similarity. *Psychological Review* 84(2). 327–352.
Zeschel, Arne & Felix Bildhauer. 2009. Islands of acceptability. Paper presented at the AfLiCO, Paris.
Zeschel, Arne. 2010. Exemplars and analogy: semantic extension in constructional networks. In Dylan Glynn & Kerstin Fischer (eds.), *Quantitative Methods in Cognitive Semantics*, 201–223. Berlin & New York: Mouton deGruyter.