Modelling Verb Selection within Argument Structure Constructions

THIS IS AN AUTHOR'S COPY OF THE MANUSCRIPT WHICH IS NOW PUBLISHED IN *LANGUAGE, COGNITION AND NEUROSCIENCE*. DOI: 10.1080/23273798.2016.1200732

Yevgen Matusevych
Tilburg Center for Cognition and Communication (TiCC); Department of Culture Studies, Tilburg University, the Netherlands

Afra Alishahi
Tilburg Center for Cognition and Communication (TiCC), Tilburg University, the Netherlands

Ad Backus
Department of Culture Studies, Tilburg University, the Netherlands

Modelling Verb Selection within Argument Structure Constructions

This article looks into the nature of cognitive associations between verbs and argument structure constructions (ASCs). Existing research has shown that distributional and semantic factors affect speakers' choice of verbs in ASCs. A formal account of this theory has been proposed by N. C. Ellis, O'Donnell, and Römer (2014a), who show that the frequency of production of verbs within an ASC can be predicted from joint verb–construction frequency, contingency of verb–construction mapping, and prototypicality of verb meaning. We simulate the verb production task using a computational model of ASC learning, and compare its performance to the available human data. To account for individual variation between speakers and for order of verb preference, we carry out two additional analyses. We then compare a number of prediction models with different variables, and propose a refined account of verb selection within ASCs: overall verb frequency is an additional factor affecting verb selection, while the effects of joint frequency and contingency may be combined rather than independent.

Keywords: verb selection; argument structure construction; input frequency; semantic prototypicality; computational model

Introduction

Speakers' language use is conditional on the linguistic means they possess. In a way, an individual's language use provides us with a "window to the mind" (Gilquin, 2010): linguistic representations are studied through language use (see a review by Clahsen, 2007). At the same time, one of the tenets of cognitive linguistics is that linguistic knowledge is directly grounded in previous usage events (e.g., Kemmer & Barlow, 2000). Such events include both language production and comprehension, thus an individual's language use depends to a certain extent on the properties of the input (s)he has been exposed to. Indeed, it is known that input-related (e.g., distributional) properties of a linguistic unit affect how this unit is used or processed (e.g., N. C. Ellis, 2002; Gor & Long, 2009; Hoff & Naigles, 2002). But to determine the importance of various input-related factors, we need formal models predicting language use from multiple factors at once.

In the present article, we study the processing of argument structure constructions through a verb production task. In the traditional view of argument structure, the term describes how the arguments of a predicate (typically a verb) are realised: the verb *eat* involves two participants, hence two arguments; importantly, the verb is believed to predict its structure (Haegeman, 1994). In constructionist accounts, in particular Goldberg's construction grammar (Goldberg, 1995, 2006; Goldberg, Casenhiser, & Sethuraman, 2004), argument structures obtain properties independent of particular verbs through the emergence of abstract argument structure *constructions*, a particular type of linguistic constructions (or form–meaning pairings) that "provide the means of clausal expression" (Goldberg, 1995, p. 3): for example, the verb *eat* often participates in a transitive construction, which has the form SUBJ VERB OBJ and the meaning X *acts on* Y. Such constructions slowly emerge in a learner's mind as (s)he categorises individual verb instances. Although this is a simplistic description, argument structures can be seen as verb-centred mental categories (Goldberg et al., 2004; Goldberg, Casenhiser, & Sethuraman, 2005), where a variety of verbs may occupy the central slot in each construction.

The studies mentioned above investigate, among other things, the role of individual verbs and their properties in formation of argument structure constructions, considering their abstract nature.

Within a given construction, speakers prefer some verbs over others. In particular, some verbs within a construction are produced more frequently than others, they come to mind first, and they are learned earlier (e.g., N. C. Ellis & Ferreira-Junior, 2009; Goldberg et al., 2004; Naigles & Hoff-Ginsberg, 1998; Ninio, 1999b; Theakston, Lieven, Pine, & Rowland, 2004): e.g., the SUBJECT VERB LOCATION construction attracts such verbs as *go*, *come*, and *get*, while *sleep* and *telephone* are rather rare (data from N. C. Ellis & Ferreira-Junior, 2009). Two groups of factors have been considered to predict verb preference: distributional and semantic factors, yet there is no conclusive evidence on the exact contribution of each factor. At the same time, it is important to reveal their exact contributions, in order to better understand the underlying nature of links between verbs and constructions in speakers' minds. Understanding which input properties enable individual verbs to group into constructions would contribute to our knowledge about the mental grammar, or "constructicon".

Our goal in this article is to evaluate the role of specific distributional and semantic factors. As a methodological tool, we use a computational model of construction learning. Computational models enable us to overcome some of the methodological limitations imposed by studying human subjects and, as a result, make informed predictions about the role of some of the proposed factors. Ultimately, our study endeavours to propose a refined prediction model explaining verb selection in argument structure constructions. This will help us to understand which factors are responsible for the emergence of links between verbs and constructions in the minds of language users.

The article is organised as follows. In the next section we review some existing studies on the issue (*Predicting verb selection*), motivate our focus on particular studies (N. C. Ellis et al., 2014a; N. C. Ellis, O'Donnell, & Römer, 2014b), and expose two methodological issues that we plan to address. We also introduce distributional and semantic factors considered in the article, and explain why these factors may be important (*Factors affecting verb selection*). This is followed in the section *Material and methods* by the description of the setup of our study: computational model, input data, test stimuli, and the exact predictor variables representing the distributional and semantic factors under consideration. The *Simulations and results* section consists of three studies: the first one is intended to simulate the original experiments: we demonstrate a reasonable performance of our model in the target task, and fit a regression explaining this performance as a function of the predictor variables. The second study addresses two methodological issues: we show how the regression coefficients change when each of the issues is resolved. In the final study (*Refining the prediction model*) we consider alternative combinations of predictor variables that may better explain the model's performance in the target task. *General discussion* summarises the article, and is followed by a short *Conclusion*.

Theoretical overview

Predicting verb selection

N. C. Ellis et al. (2014a, 2014b), henceforth EOR, provided native and non-native English speakers[1] with a set of stimuli, which schematically represented argument structure constructions with a verb missing: *it* __ *about the...*, *s/he* __ *across the...*, *it* __ *as the...*, etc. Each stimulus was presented both with an animate (*he* or *she*) and with an inanimate (*it*) pronoun. Participants had to spend a minute to produce verbs fitting the slot. Note that EOR's stimuli have a very weak semantic component: they are, in fact, form-based patterns, and participants are free in their interpretations of the arguments' thematic roles. Römer, O'Donnell, and Ellis (2015) motivate such an approach by the fact that they analyse semantic associations between verbs and constructions, and therefore it is "important to initially define the forms that will be analysed in a semantics-free, bottom-up manner" (p. 45). Although this is a controversial point (and we return to it in the discussion), in this study we follow their approach.

Importantly, this task is used to investigate the acquired associations between verbs and constructions, and it is not suitable for studying language production as such. In production speakers start from the intended meaning, and then encode this meaning using some of the suitable forms (words, grammatical patterns, etc.). In contrast, EOR's participants are cued with a pattern with little semantic information and have to select a verb (that is, a form and a meaning at the same time) that fits the pattern. In this capacity, the task is similar to other psycholinguistic tasks often used for studying human memory, implicit knowledge of words, and mental grammar: the fill-in-the-blank (cloze) task, the free word association task, and the cued recall task (see Shaoul, Baayen, & Westbury, 2014, for a review).

Following the task, the cumulative frequency of production of each verb in each construction was calculated. Statistical analyses revealed that the cumulative production frequency could be predicted from three input variables—verb frequency in the construction, contingency of verb–construction mapping, and prototypicality of verb meaning—with an independent contribution of each variable. Here we only briefly define the variables, more information on each of them is given below (see *Factors affecting verb selection*).

- Verb frequency in the construction: how frequently a verb appears within a specific construction in the linguistic input.

- Contingency of verb–construction mapping: to what extent the use of a specific construction is indicative of a particular verb, compared to other constructions/verbs.

- Prototypicality of verb meaning: how representative the verb meaning is for the general semantics of a construction.

Some of these findings are in line with some existing studies in language acquisition, which look at verb production by children. In particular, the verb frequency effect has been also found by Naigles and Hoff-Ginsberg (1998), Ninio (1999a), and Theakston et al. (2004). However, Ninio (1999a) suggests that the effects of frequency and prototypicality are not independent, and Theakston et al. (2004) find no effect of prototypicality after the frequency is accounted for.

Additionally, there is a number of studies carried out by Ambridge and colleagues, who investigate whether distributional and semantic factors help children and L2 learners to learn restrictions for the verb use in various argument structure constructions (Ambridge, Bidgood, Twomey, et al., 2015; Ambridge & Brandt, 2013; Ambridge, Pine, & Rowland, 2012; Ambridge, Pine, Rowland, Freudenthal, & Chang, 2014, etc.). Although these studies mostly use grammaticality judgements, a production experiment has been reported as well (Blything, Ambridge, & Lieven, 2014). This line of research demonstrates the role of both distributional and semantic factors in construction learning. Their results in terms of the role of distributional factors are consistent with other studies mentioned above. As for the role of semantics, Ambridge and colleagues in their studies use a very different interpretation of verb semantics, focusing on fine-grained discriminative features of the verb meaning, which are based on Pinker's (2013) verb classes (we return to this issue in the final discussion). This makes it difficult to compare their findings in terms of verb semantics to what other studies report.

In short, there is no conclusive evidence about the exact contribution of each specific factor to explaining the verb use within argument structure constructions. We focus on the studies of EOR, because they investigate both groups of factors on a large set of constructional patterns.

Methodological issues

There are two potential methodological issues in EOR's analyses, which may have some implications for the ecological validity of their studies. The first issue relates to how the values of the predictor variables (in particular, frequency and contingency) are obtained. All input estimates are based on the British National Corpus (BNC). Although the use of large corpora for approximating language input to learners is rather common and well justified overall, the method has certain shortcomings when it comes to accounting for the individual variation between speakers (e.g., Blumenthal-Dramé, 2012). The variation in individual experiences with a language may lead to the formation of different linguistic representations in learners (Dąbrowska, 2012; Misyak & Christiansen, 2012). The variation is even higher among L2 learners, whose learning trajectories may vary greatly (e.g., Grosjean, 2010). In EOR's case, verb production data obtained from multiple individuals are predicted by input-related measures computed from a corpus, which is, again, generated by a language community. This way, EOR demonstrate that their model predicts verb selection on the population level. But cognition is individual, and for making informed claims about cognitive representations we need to test the selection model on the input to individual speakers and individuals' production data. This is a challenging task for studies with human subjects, because it is nearly impossible to account for the whole learning history of an individual.

Another issue we focus on relates to the use of cumulative frequency of verb production. Calculating the total number of times each verb has been produced by all the speakers in a specific construction results in losing the information about the order of production. Yet, the order of verb listing must also be taken into account. For example, the verb position in a produced list has been shown to correlate with the frequency of production of this verb in a category-listing task (Plant, Webster, & Whitworth, 2011). Similarly, studies on sentence production show that, all things being equal, the more accessible (prototypical, frequent) word in a word pair tends to be placed earlier in a sentence than the less accessible one (e.g., Bock, 1982; Onishi, Murphy, & Bock, 2008). These findings suggest it is important to account for the order of verb production in the experimental task described above. In fact, EOR briefly mention this issue among the limitations of their study.

One objective of the current study is to simulate EOR's experiments using the computational model of argument structure construction learning (Alishahi & Stevenson, 2008; Matusevych, Alishahi, & Backus, 2015b). The second objective is to test whether the findings of EOR still hold after addressing the two methodological issues described above; the computational model is particularly helpful in this respect. First, it provides us with control over the input to each simulated learner, and eliminates other possible sources of individual variation, related to learners' cognitive abilities, propensities, etc. (R. Ellis, 2004). Second, the model generates the probability of production of each verb, which makes it easy to account for the order of verb preference (see *Test data and elicited production* below).

Our final objective relates to the original prediction model, which uses frequency, contingency and prototypicality to explain verb selection. Based on some theoretical premises presented in the next section, we propose a refined prediction model in the current study, and show that it may have a higher explanatory power than EOR's original model. We proceed with a critical overview of the three variables used in the original experiments.

Factors affecting verb selection

Input frequency

Language learners are sensitive to frequencies of occurrence of linguistic units in the input. Frequency effects have been demonstrated in many domains of language processing and language use (see overviews by Ambridge, Kidd, Rowland, & Theakston, 2015; Diessel, 2007; Divjak & Caldwell-Harris, 2015; Lieven, 2010). Frequencies also relate to the concept of entrenchment in cognitive linguistics: more frequent words (in this case, verbs) get entrenched stronger in learners' minds, which makes them more accessible (Bybee, 2006; Langacker, 1987; Schmid, in press). Although the existence of frequency effects is commonly recognised in cognitive linguistics, it is unclear yet which frequencies count (N. C. Ellis, 2012): of a particular word form (*goes*), of a lemma (all occurrences of *go*, *went*, etc.), of a form used in a specific function (*go* as an imperative), of an abstract meaning alone, etc. The frequency effect may also depend on the level of granularity of the examined units (Lieven, 2010). The complexity of the issue is reflected in the number of different kinds of frequencies discussed in the literature:

- Token vs. type frequency (Bybee & Thompson, 1997): the number of occurrences (tokens) of a specific lexical unit in a corpus vs. the number of various specific units (types) in a corpus matching a given abstract pattern.

- Absolute vs. relative frequency (Divjak, 2008; Schmid, 2010): the absolute measure denotes the independent frequency of a unit (e.g., the verb *go* has been produced 25 times in the construction *he/she/it* VERB *across* NOUN), while the relative measure relates the frequency of the target unit to the frequencies of competitor units, capturing this way paradigmatic relations of the units (e.g., the verb *go* takes a 10 percent share of all the verb tokens produced in the construction *he/she/it* VERB *across* NOUN). This difference between the measures has to do with the notion of contingency (association strength), discussed in more detail in the next section. It is useful to visualise it using a verb–construction frequency (or contingency) table (see Table 1): the absolute verb frequency is expressed as $a + b$, while the relative frequency must relate this value to the frequency of competing verbs, $c + d$.

- Marginal vs. joint frequency: unlike the previous pair, this distinction concerns the syntagmatic relations of two units. A unit's marginal frequency is its overall frequency in a corpus (e.g., the verb *go* occurs in the BNC approximately 86,000 times); also sometimes referred to as "raw frequency". In Table 1, the marginal frequency of the target verb is denoted as $a + b$, and the marginal frequency of the target construction is $a + c$. The joint frequency a, on the other hand, denotes how frequently the target verb occurs in the target construction (e.g., the verb *go* in the construction SUBJ VERB *across* LOC occurs in the BNC approximately 120 times).

This last distinction requires further attention here. EOR in their analysis always employ the joint verb–construction frequency as one of the predictors. This measure has been considered in

Table 1: A verb–construction contingency table

	Target construction	**Other constructions**	**Total**
Target verb	a	b	$a + b$
Other verbs	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

studies of some linguistic behaviours, such as acceptability judgements (e.g., Divjak, 2008), as well as in language acquisition (e.g., Theakston et al., 2004). However, these studies also take into account the marginal verb frequency. In particular, Ambridge, Kidd, et al. (2015) argue that both types of frequencies affect child language learning. Talking about production in particular, Blything et al. (2014) carried out a production experiment with children, and used, among others, measures called "entrenchment" and "preemption" to predict the probability of verb production. Both measures were based on the overall frequency of a verb (or verbs) in the BNC, and their observed effects also support the idea that the marginal verb frequency is important. This idea is also in line with the theoretical account of units' entrenchment in the cognitive system, proposed by Schmid (2010); Schmid and Küchenhoff (2013). They distinguish between cotext-free and cotextual entrenchment: while cotext-free entrenchment is related to the marginal item frequency, cotextual entrenchment captures syntagmatic associations between items, just as the joint frequency of two items does.[2] For measuring the syntagmatic association strength, various association measures have been proposed, which we discuss in the next section.

At this point it is important to note that the verb selection model of EOR does not take into account the marginal verb frequency, and we believe that including this variable in the model could improve it. EOR motivate their exclusion of the marginal verb frequency ("raw", in their terminology) by the fact that verb selection in their test correlates better with the joint verb–construction frequency than with the marginal verb frequency. But assuming the potentially independent effects of the two kinds of frequencies, the inclusion of the marginal verb frequency into the model may be justified.

Contingency of mapping

The second factor in EOR's model is contingency, or the reliability of verb–construction mapping. Although EOR use a particular measure explained below, contingency is an umbrella term for multiple measures of the association strength between a particular verb and a particular construction. The notion of contingency comes from the paradigm of human contingency learning, focusing on learning associations between stimuli, which are often described in terms of cues and outcomes. The term is rarely used in linguistic studies, which prefer talking about association strength, or about "contextualised" frequency measures (Divjak & Caldwell-Harris, 2015). Joint verb–construction frequency is the simplest example of such a measure, while other measures represent more sophisticated ways to quantify how well a verb and a construction go together. Therefore, we argue that the simultaneous use of two contingency measures within the same model may be redundant.

In various disciplines, the impact of contingency has been shown to be independent from that of frequency. In particular, some classical models of memory recall implement the effects of frequency and association strength independently of one another (Anderson, 1983; Gillund & Shiffrin, 1984). Studies on item- versus association-memory in word retrieval also indicate that these two types of memories are independent of each other (e.g., Hockley & Cristi, 1996; Madan, Glaholt, & Caplan, 2010). However, these studies talk about the marginal item frequency, which, as we have mentioned, deals with an item in isolation. Therefore, the mentioned studies can hardly be used as an argument in favour of the independent effects of *joint* frequency and contingency within the same model.

The second issue related to contingency has to do with the ongoing discussion in cognitive linguistics about which contextualised measure has a higher predictive power (Bybee, 2010; Divjak, 2008; Gries, 2013, 2015; Küchenhoff & Schmid, 2015; Schmid & Küchenhoff, 2013; Stefanowitsch & Gries, 2003). Just as in the previous section, these measures are commonly presented using a contingency table (see Table 1). Despite a great number of proposed association measures (see overviews by Evert, 2005; Pecina, 2010; Wiechmann, 2008), we can make a simple distinction between three types, based on how many of the table cells *a–d* the measure takes into account (e.g., Divjak, 2008; Divjak & Caldwell-Harris, 2015):

1. Raw joint frequency (cell *a*) is the most intuitive way to measure how well a verb and a construction go together: the verb *go* in the construction SUBJ VERB *across* LOC occurs in the BNC approximately 120 times.

2. Conditional probabilities relate the joint frequency to the marginal token frequency of either a construction ($Attraction = \frac{a}{a+c}$) or a verb ($Reliance = \frac{a}{a+b}$). Such normalisation of the raw joint frequency is useful when, for example, multiple constructions with different frequencies are studied: the same number of 120 occurrences of a particular verb may account for 90 percent of all verb usages in one construction, but only for 10 percent in another one.

3. Complex associative measures take into account all the four cells *a–d*. An example of such a measure is $\Delta P_{Attraction}$, or $\Delta P(construction \rightarrow word) = \frac{a}{a+c} - \frac{b}{b+d}$, which is used in the original studies of EOR. Other popular measures include, e.g., Minimum Sensitivity (Wiechmann, 2008) and the *p*-value of Fisher–Yates exact test (Stefanowitsch & Gries, 2003). The use of such measures can be motivated by the need to capture the competition between the verbs and the constructions at the same time, in particular to address the problem of hapax legomena. For example, in a study of *as*-predicative (Gries, Hampe, & Schönefeld, 2005) the unrepresentative verb *catapult* scored highest in Reliance among many other verbs, only because it never occurred in other constructions in the corpus. The use of a complex measure solved the problem in their case. At the same time, other researchers (e.g., Blumenthal-Dramé, 2012; Divjak, 2008; Schmid & Küchenhoff, 2013) suggest that complex measures may have little advantage over the conditional probabilities (type 2 above).

To summarise, we think that including both joint frequency and ΔP (or any other contingency measure) into the model, as in EOR's studies, may not be well justified. We suggest that only one such measure should be considered in the analysis, while the other is redundant. In the current study we consider one measure of each type specified above, as well as their combinations, to test which of them predicts verb selection better.

Semantic prototypicality

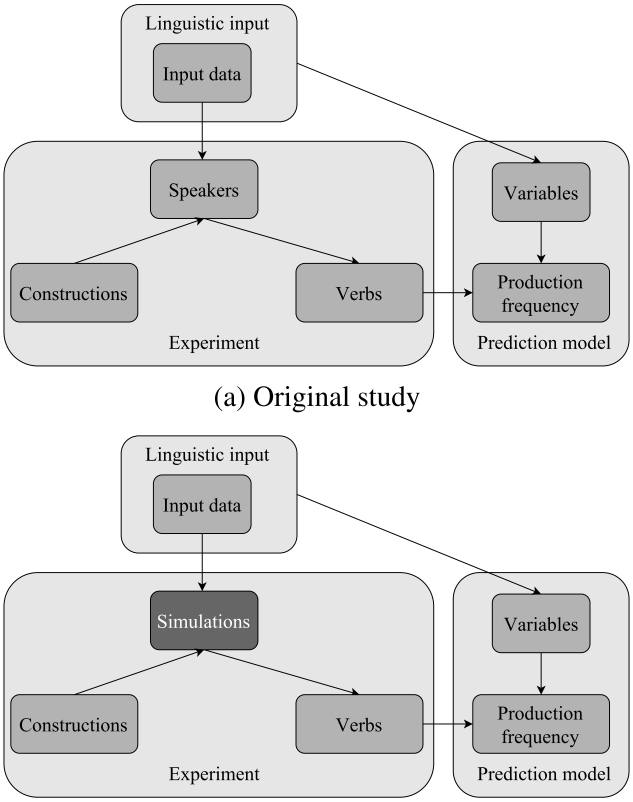
Semantic prototypicality is a concept borrowed from studies on category structure; it is also known under alternative names, such as "family resemblance" (Rosch & Mervis, 1975), "goodness-of-example" (Mervis, Catlin, & Rosch, 1976), "typicality", "goodness of membership" (Onishi et al., 2008), etc. It is common in cognitive science to estimate the typicality of concepts within a semantic category using so-called category norms—ranked lists of items based on human production data (e.g., Kelly, Bock, & Keil, 1986; Plant et al., 2011). EOR, however, do not use this approach, as it would lead to circular reasoning: prototypicality is used to predict the production data, and thus can not be computed based on other production data. Instead, for each considered construction (e.g., *he/she/it* VERB *across* NOUN) they build a semantic network of verbs participating in this construction (*go*, *move*, *face*, *put*, etc.). This network is organised according to the similarity of verb meanings, as informed by WordNet (Miller, 1995). Using a network for a particular construction, they compute a measure called betweenness centrality, which indicates the centrality of each verb's meaning in this construction. This way, the most general verbs in the construction (in this case, *go* and *move*) tend to obtain higher prototypicality values (see Gries & Ellis, 2015; Römer et al., 2015, for more detail). In this sense, "semantic generality" would be a more suitable term, however we follow EOR and other studies mentioned next in using the word "prototypicality". An additional advantage of EOR's method to compute prototypicality is that the resulting values are independent of the corpus-based frequency and contingency measures.

Semantic prototypicality has also been studied in language acquisition research: semantically general verbs have been suggested to be "pathbreaking" in child language use (e.g., Ninio, 1999a, 1999b). However, semantic generality is often confounded with input frequency: general verbs tend to be used most frequently (Goldberg et al., 2004; Ninio, 1999a), and the independent effect of semantic generality is not always found (Theakston et al., 2004). At the same time, EOR argue that the effect of semantic prototypicality is independent of frequency: while frequency relates to entrenchment, prototypicality has to do with the spreading activation in semantic memory (Anderson, 1983): if verbs within a construction form an interconnected network, then more central (general, prototypical) verbs in this network are more likely to be activated, and thus to be produced. To summarise, there is no conclusive evidence on whether the semantic prototypicality of a verb is a good predictor of its use.

Summary

This theoretical overview shows that the role of both the distributional (frequency, contingency) and the semantic factors (prototypicality) requires further research. In particular, it is unclear yet whether marginal verb frequency plays an independent role in predicting verb selection; which measures of contextual frequency should be included into a prediction model, and how many of such measures; finally, the role of semantic prototypicality is under discussion. We will address these issues in our study, but first we proceed with its methodological description.

Figure 1: Design of EOR’s study and its simulation; updated components are marked with a darker colour



(a) Original study

(b) Our initial study: computational simulations replace human speakers

Material and methods

Study overview

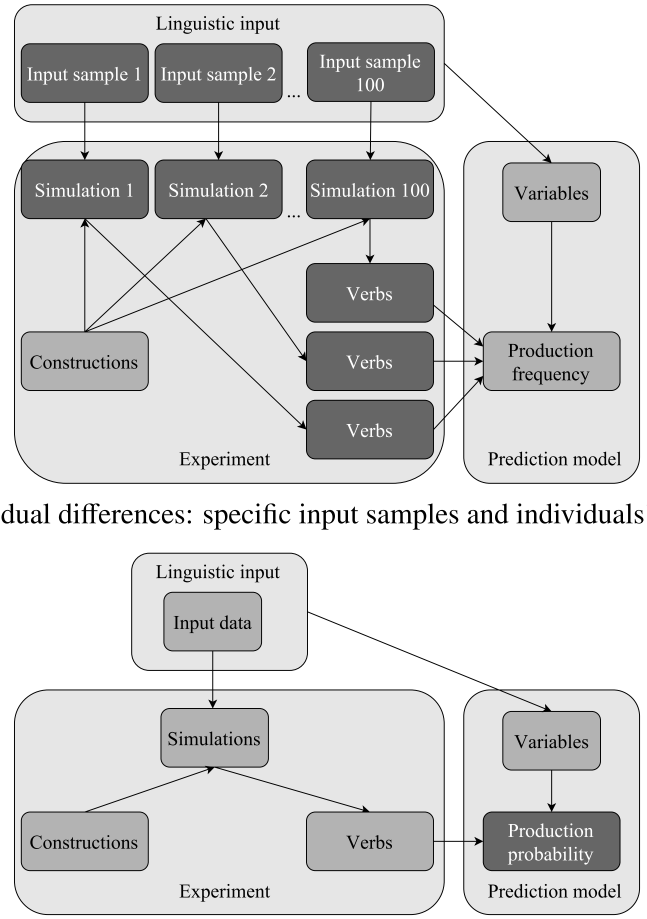
Figures 1–3 present a schematic overview of the design employed in the original studies and in the present study, the latter being divided into three main steps. Only a brief summary for each step is given here, while more detail can be found in the respective sections below.

There are three main blocks of the original study: (1) experiment, (2) linguistic input, and (3) prediction model (Figure 1a). During the experiment, L1 or L2 speakers are exposed to a set of constructions with the main verb missing, and produce a set of verbs. Three predictor variables are extracted from the BNC, under the assumption that this corpus provides an approximation of the linguistic input that participants have been exposed to in their lifetime. These variables are then used in the prediction model to explain the frequency of production of verbs within constructions.

The overall design of our first step (Figure 1b) is almost identical, except we use computational simulations instead of human speakers, and different data sets. The goal of this step is to check the validity of our computational model; that is, to see whether it selects verbs that fit the target constructions, and whether such selection can be explained by the same input-related features as in EOR’s experiments.

At step two we address the methodological issues described earlier (Figure 2). First, we dis-

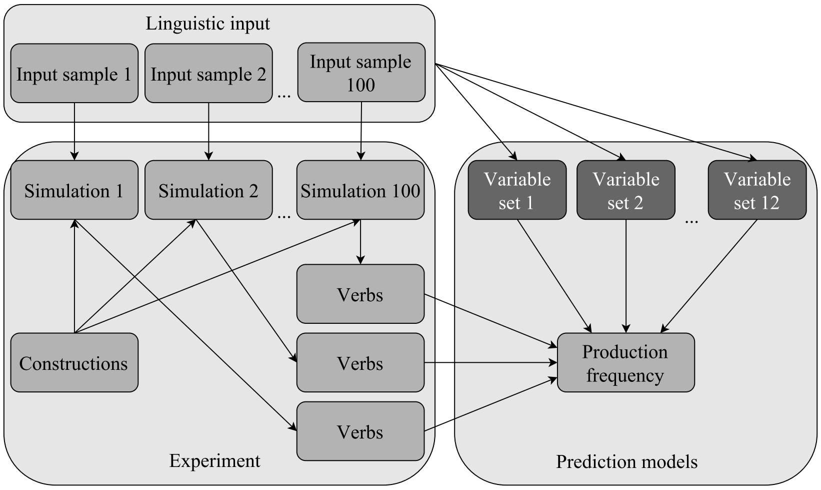
Figure 2: Analyses addressing methodological issues; updated components are marked with a darker colour



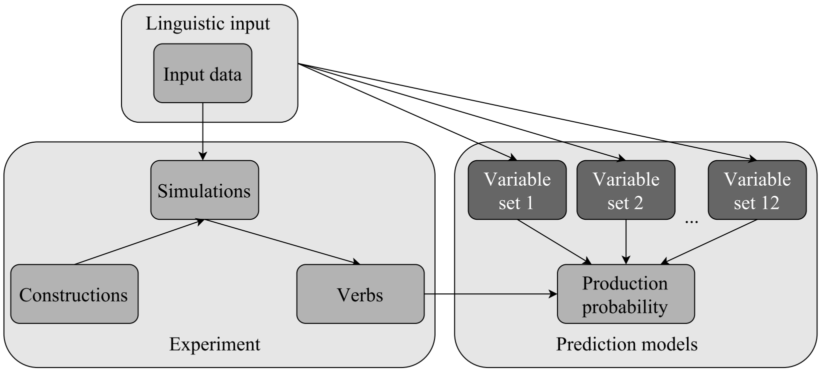
(a) Accounting for individual differences: specific input samples and individuals' production lists are used

(b) Accounting for order of preference: production probability replaces production frequency

Figure 3: Refining the prediction model; updated components are marked with a darker colour



(a) Models accounting for individual differences: alternative sets of predictors are considered, cf. Figure 2a



(b) Models accounting for order of preference: alternative sets of predictors are considered, cf. Figure 2b

tinguish between individual input samples instead of generalising over the whole population (see *Addressing the methodological issues: Individual variation* below, also Figure 2a). Second, in a parallel analysis we employ the production probability instead of production frequency, to account for the order of verbs produced by speakers (more detail below, under *Addressing the methodological issues: Order of preference*, also Figure 2b).

At the final step three we test various prediction models to select the one that explains the simulated data sets best, using the two types of design from step two (see Figure 3). The following sections describe the essential components of the study: computational model, input data, experimental setup, and predictor variables.

Computational model

The model used in the current study is based on a model of human category learning, which was shown to replicate multiple experimental findings in this area (Anderson, 1991). Alishahi and Stevenson (2008) employed the same learning algorithm for simulating early learning of argument structure constructions (which is sometimes seen as a categorisation task: Goldberg et al., 2004). The model of construction learning demonstrated similarity to human data in terms of U-shaped

Table 2: An instance for the verb usage *We sold the house.*

Feature	Value
Head predicate	*sell*
Predicate semantics	EXCHANGE, TRANSFER, POSSESSION, CAUSE
Number of arguments	2
Argument 1	*we*
Argument 2	*house*
Argument 1 semantics	REFERENCE, PERSON ..., ENTITY
Argument 2 semantics	DWELLING, HOUSING ..., BUILDING
Argument 1 thematic role	COMPANY (N$_1$), PERSON (N$_1$) ..., CIVILISATION (N$_1$)
Argument 2 thematic role	RELATION (N$_1$), MATTER (N$_3$) ..., OBJECT (N$_1$)
Argument 1 case	N/A
Argument 2 case	N/A
Syntactic pattern	ARG1 VERB ARG2

learning patterns, use of syntactic bootstrapping (both in production and comprehension), phenomena of over-generalisation and recovery (Alishahi & Stevenson, 2008, 2010). Finally, the model was adapted for simulating bilingual construction learning, demonstrating effects of amount of input similar to those in human learning (Matusevych et al., 2015b).

The model relies on some theories of cognitive linguistics and construction grammar, in particular those of Goldberg (1995); Tomasello (2003); for more details, see Alishahi and Stevenson (2008). Most importantly, the input is processed iteratively, so that constructions gradually emerge from categorising individual instances item by item (similar to the theory described by Goldberg et al., 2004). At the end of the learning process, the model uses its knowledge of argument structure constructions in the elicited verb production task. While the learning model has been used before, the implementation of the test task for this model is novel. We describe these steps in more detail.

Input representations

The model is exposed to a number of instances, each of which represents a single verb usage in a specific construction. Each instance comprises several information cues characterising the respective verb usage. Table 2 shows such a usage, with the full set of features listed in the left column.

We make a simplifying assumption that the model can infer the values of all the provided features from the utterance and the respective perceptual context. This means, in particular, that the model can recognise the words in the utterance and infer their meanings and linguistic cases (where appropriate)[3], as well as to identify the role of each participant in the described event.

Each feature F_k is assigned a value within an instance I, so that I is a unique combination of specific feature values (F_k^I). Following some linguistic theories (e.g., Dowty, 1991; McRae, Ferretti, & Amyote, 1997), features expressing semantic and thematic role properties are represented as a set of elements each, and these sets were semi-automatically obtained from the existing resources (see *Input data and learning scenarios* below). Regarding the thematic roles, it has been shown that the model used in this study can learn representations of "traditional" thematic roles (e.g., AGENT, THEME) from distributed sets of properties (Alishahi & Stevenson, 2010). A distributed representation of the thematic roles in the current study provides at least two advantages over representing each role as a single symbol. First, set representations enable the model to estimate how similar lexical meanings or thematic roles are to each other. Second, computing the semantic prototypicality of a verb is rather straightforward for set representations of verb meanings (see *Predictor variables* below). As can be seen in Table 2, each verb meaning is represented as a set of semantic primitives describing this meaning: e.g., {EXCHANGE, TRANSFER, POSSESSION, CAUSE} for the verb *sell*. These elements are automatically extracted from available sources (see section *Input data and learning scenarios* below). An argument structure construction (henceforth ASC) emerges as a generalisation over individual instances, where each feature contributes to forming the generalisation. An ASC combines the feature values from all the participating instances, but it is impossible to recover individual instances from an ASC (unless it only contains a single instance). An individual instance is a set F^I of feature values F_k^I ($F_k^I \in F^I$), and an ASC S is a set F^S of feature values F_k^S ($F_k^S \in F^S$), but in an ASC each feature value ($e \in F_k^S$) may occur more than once, depending on the number of participating instances with the value $F_k = e$.

Learning mechanism

The learning is performed using an unsupervised naive Bayes clustering algorithm. As we mentioned, the model receives instances one by one, and its task is to group the incoming instances into ASCs by finding the "best" ASC (S_{best}) for each given instance I:

$$S_{best}(I) = \underset{S}{\operatorname{argmax}}\ P(S|I) \tag{1}$$

In other words, the model considers each ASC it has learned so far, seeking the most suitable category for the encountered instance. It makes little sense to talk about the probability of an ASC (prior knowledge) given an instance (new evidence), therefore, the Bayes rule is used to estimate the conditional probability in equation 1:

$$P(S|I) = \frac{P(S)P(I|S)}{P(I)} \tag{2}$$

The denominator $P(I)$ is constant for each ASC, and therefore plays no role in making the choice. The choice of ASC for the new instance is affected by the two factors in the numerator:

1. The prior probability $P(S)$, which is proportional to the frequency of the ASC in the previously encountered input (or the number of instances that the ASC contains so far, $|S|$):

$$P(S) = \frac{|S|}{N+1}, \tag{3}$$

where N is the total number of instances encountered by that moment. The learner always has an option to form a new ASC from a given instance. Although initially such a potential ASC contains no instances, its value $|S|$ is assigned to 1, to avoid 0s in the multiplicative equation 2. The determining role of frequency is grounded in usage-based linguistics: a frequent ASC is highly entrenched and is easier to retrieve from memory, so that new instances are more likely to be added to it.

2. The conditional probability $P(I|S)$, which takes into account how similar an instance I is to S. The higher the similarity between I and S, the more likely I to be added to S: this is based on studies pointing to the importance of similarity in categorisation tasks (e.g., Hahn & Ramscar, 2001; Sloutsky, 2003). The model compares each instance to each ASC by looking at the independent features listed in Table 2, such as the head predicate, argument roles, etc. For example, all being equal, two usages of the same verb are more likely to be grouped together than two usages of different verbs, yet this can be compensated by other features. Technically speaking, the overall similarity is a product of similarities for individual features:

$$P(I|S) = \prod_{k=1}^{|F^I|} P\left(F_k^I \middle| S\right) \tag{4}$$

   The probability $P\left(F_k^I \middle| S\right)$ in this equation is estimated differently depending on the feature type, see appendix.

Based on the computed values of the prior and the conditional probability, the model either places I into an existing ASC or creates a new ASC containing only one instance I. Note that when the model receives instances from two languages during a simulation, L1 and L2 instances are not explicitly marked as such. The only relevant information is implicitly present in the values of such features as head predicate, arguments, and syntactic pattern (in case it has prepositions). This ensures the model treats all instances equally, irrespective of their language.

Input data and learning scenarios

Following the original experiments, we simulate L1 English (as in N. C. Ellis et al., 2014a) and L2 English learning (as in N. C. Ellis et al., 2014b). Although the latter study was carried out with native speakers of German, Spanish, and Czech, we only use L1 German due to poor data availability. Manual annotation of argument structures proved to be rather time-consuming, therefore we used available annotated resources for English and German to automatically extract the data we needed.

We use the data sets available from Matusevych et al. (2015b); here we briefly outline how they were obtained.

1. The Penn Treebank for English (WSJ part, Marcus et al., 1994) and the TIGER corpus for German (Brants et al., 2004) were used to obtain syntactically annotated simple sentences.

2. Argument structures were extracted from these sentences, using the annotations in English PropBank (Palmer, Gildea, & Kingsbury, 2005) and the German SALSA corpus (Burchardt et al., 2006).

3. We further used only the sentences containing FrameNet-style annotations (Ruppenhofer, Ellsworth, Petruck, Johnson, & Scheffczyk, 2006), either via the PropBank–FrameNet mappings in SemLink for English (Palmer, 2009), or in the SALSA corpus for German.

4. Word semantic properties were obtained from WordNet (Miller, 1995) and VerbNet (Schuler, 2006).

5. Symbolic thematic roles were semi-automatically replaced by sets of elements through the WordNet–FrameNet mappings (Bryl, Tonelli, Giuliano, & Serafini, 2012).

The resulting German and English data sets contain 3,370 and 3,624 ASC instances, respectively, which are distributed across 301 (German) and 319 (English) verb types. The corpora mentioned above were the only large sources of English and German data for which the annotations of argument structure were available. We acknowledge that the kind of language in these corpora (mostly newspaper texts) differs from what L1 and L2 learners are normally exposed to. Moreover, the distributions of verbs and constructions in the corpora may be genre- or domain-specific and differ from English and German in general, and the data sets are limited in size: many constructions occur with only a few verb types (we look at this in more detail below, see *L1 simulations*). This prevents us from making statements about specific English verbs or constructions, yet the extracted data sets do suit our goal of studying the impact of individual input-related factors on the production of verbs in constructions.

Input to the computational model is sampled randomly from the distribution of instances in the presented data sets. This way, the exact input to the model varies between simulations, to simulate a population of learners with individual linguistic experiences. In the L1 learning setup, 100 simulated learners receive a cumulative number $N = 6{,}000$ English instances. Clearly, human adult speakers are exposed to much more input than 6,000 utterances, but given the size of our data sets, this value is large enough: an earlier study (Matusevych et al., 2015b) showed that the model achieved a stable level of ASC knowledge on the target input data set after receiving 6,000 instances. In the L2 setup, 100 learners are exposed to $N = 12{,}000$ instances: 6,000 L1 German instances, followed by 6,000 instances of "bilingual" input, in which English and German are mixed in equal proportions. This way, L2 learners only encounter $\frac{1}{2} \times 6{,}000 = 3{,}000$ English instances, to simulate non-native speakers whose L2 proficiency is lower than L1 proficiency.

Test data and elicited production

Learning was followed by the elicited production task. The model was provided with a number of test items, each of which was intended to elicit the production of verbs in a single construction. Following the original experiments, we looked at the representation of verbs within form-based constructions, without the semantic component: just as EOR's participants, the model is free in its interpretation of the arguments' thematic roles. We further refer to these units as "constructions", to distinguish them from the emergent ASC representations in the computational model. We did not limit our analysis to prepositional constructions with only two arguments (as did EOR), because this would substantially reduce the amount of the available data in our case. Instead, we used all the available constructions. In terms of ASC representations used by the model, each construction was defined as a syntactic pattern, e.g. ARG1 VERB *about* ARG2 (for a full list of patterns, see Table 4 below). To follow the design of the original experiments, we constructed the test stimuli as follows.

Following EOR's approach, two stimuli were generated for each construction: the first one had either a pronoun *he* or a pronoun *she* (randomly selected) as the first argument head, and the second one had a pronoun *it* as the first argument head. This way, each stimulus occurred once with an animate (*s/he*) and once with an inanimate pronoun (*it*). The other argument heads were masked, together with the verb. Therefore, during the testing the model was provided with a number of test ASC instances I_{test}, which only contained the values of a few features: number of arguments, syntactic pattern, the first argument (the selected pronoun) and its semantics (e.g., {REFERENCE, PERSON ..., ENTITY} for *he*). As a result, test stimuli were similar to those used in the original experiments (in this case, *he* __ *about the...*). Given a test instance, the model's task was to produce a list of verbs fitting the empty slot. Such elicited production is implemented as a generation of a set of verbs enumerated with their respective probabilities of production ($V_{produced}$). There is no upper boundary for the number of verbs produced, but verbs with low probabilities of production are excluded from the analysis. The probability of each $V_j \in V_{produced}$ given a test instance I_{test} is calculated as follows:

$$P\left(V_j \middle| I_{test}\right) = \sum_{S} P\left(V_j \middle| S\right) P(S|I_{test}) \tag{5}$$

The right side of equation 5 is a sum of the products of two probabilities, computed for each acquired ASC. $P(S|I_{test})$ is estimated as provided in appendix (equation 9), and $P\left(V_j \middle| S\right)$ is transformed and computed in exactly the same way as during the learning (see equations 2–4). In other words, to select verbs to fill in a test stimulus, the model first computes how similar the stimulus is to each ASC, and assigns the similarity weights to ASCs. Next, the model considers each verb associated with an ASC, and takes into account both the frequency of the verb in this ASC and the similarity weight of the ASC, to obtain the evidence from this ASC in favour of selecting particular verbs. Finally, such evidence values from all the existing ASCs add up, determining the final selection probability of each verb.

Note that our model is not equipped with explicit language control mechanisms, which human speakers can use for inhibiting activated representations from a non-target language (Green, 1998; Kroll, Bobb, Misra, & Guo, 2008). Therefore, the model may produce L1 verbs in the L2 elicited production task, which is taken into account in our analysis of production data.

Predictor variables

The predictor variables proposed in the original experiments are the joint verb–construction frequency $F(v,c)$, the ΔP-contingency $\Delta P_A(v,c)$, and the prototypicality of verb meaning $Prt(v,c)$. These measures are used for predicting the selection of verbs within each construction. Therefore, the measures are obtained based on the input data which the input to the model is sampled from. Two different methods are used for computing the values.

Our first goal is to simulate the original experiments of EOR closely following their analysis, therefore we adopt their approach of calculating the values of $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ from the whole English data set, without accounting for the individual variation in the input. The value of joint frequency $F(v,c)$ is extracted from the input data set directly, together with additional measures such as the marginal verb frequency $F(v)$, and the marginal construction frequency $F(c)$: these were needed for computing the value of contingency $\Delta P_A(v,c)$:

$$\Delta P_A(v,c) = P(v|c) - P(v|\neg c) = \frac{F(v,c)}{F(c)} - \frac{F(v) - F(v,c)}{N - F(c)}, \tag{6}$$

where N denotes the total size of the input data, in this case 3,624 instances. In simple terms, ΔP-contingency is the probability of a verb given a construction minus the probability of the verb's occurrence in all the other constructions. ΔP can take values as high as 1 (when the verb mostly occurs with the target construction) and as low as -1 (when the verb is proportionally much more frequent in other constructions).

As for prototypicality, recall that each verb meaning in ASC instances is represented as a set of elements (e.g., {EXCHANGE, TRANSFER, POSSESSION, CAUSE}), and we consider a verb v to have a higher prototypicality in a construction c when its meaning M_v shares more elements with the meanings M_i of all the other verbs i (excluding v) occurring in c ($i \in c \setminus v$):

$$Prt(v,c) = \frac{\sum_{i \in c \setminus v} \frac{|M_i \cap M_v|}{|M_v|}}{|c \setminus v|}, \tag{7}$$

where $|c \setminus v|$ is the number of verb types participating in c, excluding v. We did not use EOR's betweenness centrality values, because they were based on a so-called path similarity between verbs in WordNet, but the hierarchy of verbs in WordNet did not reflect the true hierarchy of verb meanings in our data sets.[4] At the same time, $Prt(v,c)$, as defined here, operates on the actual sets used in ASC instances, and suits our setup. The two measures, however, are conceptually similar: more general verbs with fewer semantic components (*give*: {POSSESSION, TRANSFER, CAUSE}) tend to score higher than more specific ones (*purchase*: {BUY, GET, POSSESSION, TRANSFER, CAUSE, COST}).

Our second goal is to address the methodological issues, in particular individual variation, therefore in the respective analysis the values of the three measures are calculated for each simulated learner individually, based on the actual input sample it receives. To do this, during each simulation we record the information about the occurrence of individual verb usages in the actual input: $F(v)$, $F(c)$, and $F(v,c)$. Thus, the value of joint frequency $F(v,c)$ is directly available from the recorded information, and the values of contingency $\Delta P_A(v,c)$ and prototypicality $Prt(v,c)$ are calculated as given above (equations 6–7), but based on a particular input sample instead of the whole data. N in this case is equal to the actual amount of input: $6{,}000$ for L1 or $12{,}000$ for L2 simulations.

The goal of our final study is to identify the best set of variables predicting verb selection. In particular, when presenting the three types of contingency measures, we have mentioned that we plan to test one measure of each type. A raw frequency measure $F(v,c)$ is available directly, and a complex measure $\Delta P_A(v,c)$ is calculated according to equation 6. Therefore, we only need a measure of the second type, a conditional probability. We use $Attraction(v,c)$, henceforth $A(v,c)$, which normalises the joint verb–construction frequency by the marginal construction frequency:

$$A(v,c) = P(v|c) = \frac{F(v,c)}{F(c)} \tag{8}$$

The next section describes our simulations and the obtained results. First, we simulate the original experiment for L1 (N. C. Ellis et al., 2014a, experiment 2) and for L2 (N. C. Ellis et al., 2014b), keeping our setup and analysis as close as possible to the original experiments, to see whether our model produces results similar to those of the original experiments. Next, we address the two methodological issues by reanalysing the data obtained from the same simulated learners, to examine whether the original results still hold in the new analysis. Finally, we use a number of regression models which include different combinations of predictions, to determine which factors predict the production data best.

Simulations and results

Simulating the original experiments

In this section we employ the elicited production task described under *Test data and elicited production* above to obtain a list of produced verbs. Using this list, we look at the verbs produced within some individual constructions, run correlation tests for individual constructions, and perform a combined analysis on the whole data set as described next.

Methodological details

Each simulated learner has produced a list of verbs fitting every given construction. EOR in their experiments limited the number of produced verbs by allocating a minute for each stimulus. To adopt a similar approach, we had to filter out verbs whose probability of production was lower than a certain threshold. The value of .005 was established empirically, by testing values between .05 and .001. Using this threshold value, for each verb in a certain construction we calculate the total production frequency of this verb by all learners, henceforth $PF(v,c)$. If a verb has not been produced by any learner in a certain construction, the verb–construction pair is excluded from the analysis, to obtain data similar to EOR's. For analysing L2 production data, we exclude all L1 verbs produced by the model, because these are irrelevant for our analysis.

First we look at the verbs produced within a sample of ten individual constructions: four most frequent constructions in our data set, and six constructions present in both EOR's and our data set.

Next, to compare our model to EOR's human subjects, we look at whether each of the three factors—$F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$—correlates with $PF(v,c)$ within each construction in our data set, using Pearson correlation coefficient.[5]

Finally, we proceed with a combined regression analysis on the whole data set. Again, to make the results comparable with EOR's findings, we first consider only the six constructions present in both their and our data set. However, this is a rather small sample, therefore we run an additional regression analysis on our whole data set of 44 constructions. Before fitting the models, we standardise all the variables, to make the β coefficients directly comparable and to reduce the collinearity of predictors. We run multiple regression analyses to predict $PF(v,c)$ by the three factors: $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$. Note that the values of the mentioned variables in this simulation set are computed using the first method described in the section *Predictor variables*—that is, for the whole input data set, following the original experiments.

L1 simulations

First we look at the verbs produced by the model within ten individual constructions selected as described above: the produced lists are provided in Table 3. We can see substantial differences between the frequencies of occurrence of individual constructions in the input data. Some of them are rather frequent: e.g., A1 V A2 occurs 2,508 times with 224 verb types, and A1 V occurs 724 times with 119 verb types. In contrast, most prepositional constructions are infrequent: in particular, the six constructions from EOR's data set occur only 1 to 11 times with 1 to 6 verb types. Respectively, the number of verb types generated by the model per construction also varies between 2.4 and 84.2 in this subset of ten constructions. It is also clear from the table (see bold font) that the model sometimes produces verbs which are unattested in the target construction in the input. We discuss this in the interim discussion below.

To see whether the frequencies of verb production correlate with each of the three target factors, as in EOR's study, we run a series of correlation tests reported in Table 4. We can see that both the joint frequency $F(v,c)$ and ΔP-contingency are correlated with the production frequency $PF(v,c)$ for almost all constructions: verbs which appear more frequently in a construction or which are associated more strongly with a construction are also produced more frequently by the model. This is not always the case for the third predictor, prototypicality $Prt(v,c)$: significant correlations of this variable with production frequency are only observed for 23 out of 44 constructions. In particular, there is no such correlation for any of the six constructions present in EOR's data (marked with a star in Table 4). We address this issue below in the interim discussion. The next step, as we mentioned above, is to provide combined regression analyses of the data set.

The summary of the three models is provided in Table 5a,b. Overall, the results are similar to what EOR report: all the three variables contribute to predicting the verb production frequency. However, the difference is that $Prt(v,c)$ in our experiment appears to be a less important predictor, which is reflected in the β values (from 0.05 to 0.06 in our study, depending on the set of constructions, vs. 0.29 in the original study). We have run an additional analysis, in which we kept the verbs that appeared in a construction in the input, but were not produced in this construction by the model: $PF(v,c)$ for such verbs was assigned to 0. Besides, we have run mixed effects models (e.g., Baayen, 2008), as implemented in R (Bates, Mächler, Bolker, & Walker, 2015), for the same two sets of constructions, with a random intercept and random slopes for all the three factors over individual constructions. The results appeared to be very similar to what is reported here, therefore we leave them out for brevity.

Table 3: Ten constructions with their frequencies and produced verbs. Verbs in bold are unattested with target construction in input

	Construction				
Property	A1 V A2	A1 V	A1 V A2 A3	A1 V A2 *to* A3	A1 V *about* A2
Verb tokens in input	2,508	724	112	52	11
Verb types in input	224	119	8	12	4
Verb types produced	228	115	66	47	146
Avg. verb types produced	84.2	35.5	9.7	11.6	10.6
Verb types with	*want*: 185	*want*: 169	*give*: 143	*send*: 139	*complain*: 175
their production	*buy*: 184	*begin*: 135	*send*: 117	*give*: 137	*inquire*: 154
frequencies	*sell*: 182	*die*: 108	*pull*: 90	*elect*: 99	*brag*: 131
	announce: 170	*exist*: 104	*tell*: 58	*propose*: 87	*shout*: 96
	receive: 169	*happen*: 103	*place*: 37	***disclose***: 77	***listen***: 41
	hold: 167	*expire*: 102	*disclose*: 36	*donate*: 71	***sit***: 19
	see: 162	*rise*: 99	*drag*: 33	*pass*: 70	***groan***: 17
	start: 159	*sell*: 96	***elect***: 32	*pressure*: 51	***scoff***: 14
	post: 154	*decline*: 90	*hang*: 31	*explain*: 41	***live***: 11
	lead: 153	*drop*: 89	***pressure***: 24	*peg*: 39	***send***: 11
	...	...	...	...	...
	unnerve: 1	*exhale*: 1	***wear***: 1	***want***: 1	***withdraw***: 1

Ten constructions with their frequencies and produced verbs. Verbs in bold are unattested with target construction in input (page 2 of 2)

	Construction				
Property	A1 V *into* A2	A1 V *with* a2	A1 V *for* A2	A1 V *against* A2	A1 V *of* A2
Verb tokens in input	9	7	3	1	1
Verb types in input	6	5	2	1	1
Verb types produced	206	106	77	20	21
Avg. verb types produced	24.0	10.4	6.0	2.6	2.4
Verb types with	*buy*: 107	*join*: 174	*search*: 164	*lean*: 174	*disapprove*: 154
their production	*run*: 88	*cooperate*: 141	*scream*: 135	***groan***: 17	***scoff***: 14
frequencies	***sell***: 78	*merge*: 138	***sit***: 43	***scoff***: 16	***sit***: 14
	eat: 69	*respond*: 134	***scoff***: 20	***sit***: 13	***groan***: 11
	erupt: 68	*sit*: 118	***obtain***: 19	***gaze***: 7	***gaze***: 7
	pack: 64	***scoff***: 23	***glance***: 17	***live***: 6	***live***: 7
	turn: 63	***glance***: 21	***groan***: 17	***rely***: 5	***squint***: 7
	acquire: 62	***groan***: 19	***gaze***: 8	***listen***: 4	***rely***: 5
	hold: 51	***scream***: 18	***live***: 8	***squint***: 4	***glance***: 4
	want: 50	***gaze***: 16	***rely***: 6	***glance***: 3	***listen***: 4
	...	...	...	...	...
	thrill: 1	***write***: 1	***steal***: 1	***shout***: 1	***spout***: 1

Table 4: Summary of correlation tests between $PF(v,c)$ and each of the three factors for individual constructions in L1 replication data

	$F(v,c)$		$\Delta P_A(v,c)$		$Prt(v,c)$	
Construction	r	p	r	p	r	p
A1 V	.96	<.001	.17	.002	.05	.372
A1 V A2	.94	<.001	.13	.020	.08	.162
A1 V A2 A3	.44	<.001	.22	<.001	.11	.044
A1 V A2 *about* A3	.18	.001	.18	.001	.21	<.001
A1 V A2 *above* A3	.21	<.001	.21	<.001	.14	.011
A1 V A2 *across* A3	.33	<.001	.33	<.001	.22	<.001
A1 V A2 *among* A3	.19	.001	.19	.001	.03	.622
A1 V A2 *as* A3	.43	<.001	.42	<.001	.13	.020
A1 V A2 *at* A3	.43	<.001	.28	<.001	.17	.003
A1 V A2 *by* A3	.28	<.001	.28	<.001	.13	.023
A1 V A2 *for* A3	.43	<.001	.43	<.001	.12	.029
A1 V A2 *from* A3	.36	<.001	.31	<.001	.18	.001
A1 V A2 *in* A3	.35	<.001	.34	<.001	.21	<.001
A1 V A2 *into* A3	.30	<.001	.30	<.001	.09	.125
A1 V A2 *of* A3	.22	<.001	.21	<.001	.12	.034
A1 V A2 *on* A3	.46	<.001	.33	<.001	.15	.006
A1 V A2 *over* A3	.42	<.001	.42	<.001	.15	.008
A1 V A2 *through* A3	.27	<.001	.27	<.001	.20	<.001
A1 V A2 *to* A3	.61	<.001	.39	<.001	.19	.001
A1 V A2 *under* A3	.16	.003	.16	.003	.21	<.001
A1 V A2 *until* A3	.32	<.001	.32	<.001	.14	.011
A1 V A2 *with* A3	.49	<.001	.46	<.001	.10	.062
A1 V *about* A2*	.27	<.001	.22	<.001	.02	.663
A1 V *against* A2*	.36	<.001	.36	<.001	−.01	.875
A1 V *at* A2	.31	<.001	.29	<.001	.02	.692
A1 V *below* A2	.13	.020	.13	.020	.13	.021
A1 V *by* A2	.29	<.001	.23	<.001	.02	.779
A1 V *for* A2*	.65	<.001	.63	<.001	−.12	.294
A1 V *from* A2	.13	.022	.13	.022	−.09	.095
A1 V *from* A2 A3	.56	<.001	.56	<.001	.10	.086
A1 V *in* A2	.25	<.001	.17	.002	−.04	.468
A1 V *into* A2*	.21	<.001	.17	.002	−.07	.220
A1 V *of* A2*	.35	<.001	.35	<.001	.08	.133
A1 V *on* A2	.40	<.001	.31	<.001	.12	.037
A1 V *on* A2 A3	.23	<.001	.23	<.001	.20	<.001
A1 V *to* A2	.15	.009	.13	.020	.01	.828
A1 V *to* A2 A3	.23	<.001	.23	<.001	.16	.003
A1 V *to* A2 *about* A3	.48	<.001	.48	<.001	.09	.101
A1 V *to* A2 *of* A3	.49	<.001	.49	<.001	.09	.094

A1 V *up* A2	.09	.107	.09	.107	.19	.001
A1 V *upon* A2	.23	<.001	.23	<.001	.06	.255
A1 V *with* A2*	.36	<.001	.32	<.001	−.06	.285
A1 V *with* A2 *in* A3	.26	<.001	.26	<.001	.07	.216
A1 V *with* A2 *on* A3	.44	<.001	.44	<.001	.17	.002

* Constructions present in EOR's data.

L2 simulations

For the sake of space we omit the lists of verbs produced in the L2 simulations, as well as the correlational results per construction. There were some differences between the actual sets of verbs produced in L1 and L2 simulations, but these would not be immediately obvious from verb lists or correlation tables. Although comparing L1 to L2 simulations was not our goal in this study, to further demonstrate that our model performed as expected on the simulated task, we quantified the differences between the verbs produced in L1 and L2 simulations, to compare these differences to what Römer, O'Donnell, and Ellis (2014) report. We adopted an approach similar to theirs and ran a mixed effects regression analysis predicting the frequencies of verbs produced in L2 simulations from those in L1 simulations, with the random slope over individual constructions. The model fit was reasonable (marginal $R^2 = .57$, conditional $R^2 = .65$[6]), and the β-coefficient reflecting the correlation between the produced verb frequencies in L1 and L2 simulations was equal to 0.71, which is rather close to the average value of 0.75 reported by Römer et al. (2014) for native English vs. native German speakers.

Next, we proceed with reporting on the combined regression analysis of the L2 simulation data set. Table 5c,d summarises the regression results for the simulated L2 production data. Overall, the results are similar to those for L1, and to those of EOR. Note that the values of the three target variables, following EOR's study, were computed for English constructions only. For the same reason, although the model produced some German verbs in the test task, these verbs were excluded from our analysis. However, the input to the model consisted of both English and German constructions, many of which are shared by the two languages. Since our model treated L1 German and L2 English instances in exactly the same way, it could be fairer to compute the values of $F(v,c)$ and $\Delta P_A(v,c)$, and $Prt(v,c)$ for the whole data set, assuming that each construction may be associated with both English and German verbs. This is why we ran an additional analysis, in which all the produced German verbs were kept during the analysis, and the values of the three variables were computed for the whole bilingual data set. Again, the results were very similar to the ones reported above.

Interim discussion

To summarise, the model performs as expected on the target task: verbs which appear in a construction in the input tend to populate the top of the respective list of produced verbs for this construction. Since there are six constructions present both in this study and in EOR's study, we would ideally compare the verbs produced by the model and by human participants. Yet, in our input data set these constructions occur with only 1 to 6 verb types, and the model tends to produce these verbs first. In contrast, naturalistic language input to human participants is more varied: each

Table 5: Summary of the multiple regression models fitted to the L1 replication data

a. L1 simulations: constructions present in EOR's data set

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG[a]	VIF
$F(v,c)$	0.69	0.03	$< .001$	.59	2.75
$\Delta P_A(v,c)$	0.25	0.03	$< .001$	.40	2.74
$Prt(v,c)$	0.05	0.02	.008	.01	1.02
Multiple R^2 = .83, adjusted R^2 = .82					

b. L1 simulations: all constructions

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG	VIF
$F(v,c)$	0.57	0.01	$< .001$	.73	1.13
$\Delta P_A(v,c)$	0.25	0.01	$< .001$	.25	1.14
$Prt(v,c)$	0.06	0.01	$< .001$	.02	1.02
Multiple R^2 = .50, adjusted R^2 = .50					

c. L2 simulations: constructions present in EOR's data set

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG	VIF
$F(v,c)$	0.70	0.02	$< .001$	.57	2.73
$\Delta P_A(v,c)$	0.29	0.02	$< .001$	.41	2.73
$Prt(v,c)$	0.05	0.01	.002	.02	1.02
Multiple R^2 = .90, adjusted R^2 = .90					

d. L2 simulations: all constructions

$PF \sim F + \Delta P + Prt$

Variable	β	SE	p	LMG	VIF
$F(v,c)$	0.59	0.01	$< .001$	.75	1.12
$\Delta P_A(v,c)$	0.24	0.01	$< .001$	.23	1.14
$Prt(v,c)$	0.06	0.01	$< .001$	.02	1.03
Multiple R^2 = .51, adjusted R^2 = .51					

[a] This measure is used in EOR's studies: it computes the importance of each predictor relative to the other predictors by analysing how the regression coefficients change when various combinations of predictors are excluded from the model. The measure was proposed by Lindeman, Merenda, and Gold (1980) and implemented in R by Grömping (2006).

construction occurs with a greater variety of verb types, and EOR's participants are not as limited in their verb choice as the model is. Besides, the distribution in the input per construction differs across the two studies: human participants are mostly exposed to colloquial language, while our input data set is based on business newspaper texts from the Penn Treebank (WSJ part). This is reflected in verb selection: human participants tend to produce colloquial verbs (e.g., *go*, *be*, *dance with ...*), while the model often prefers specialised verbs (*join*, *cooperate*, *merge with ...*), although in both cases verbs produced first tend to be the most frequent ones in the respective input data set.

Given the low number of verb types in some prepositional constructions, the model generalises and produces verbs unattested in these constructions, marked with bold in Table 3. These verbs mostly appear at the bottom of the list for each construction, with a few exceptions, such as A1 *elect* A2 A3, A1 *disclose* A2 *to* A3, and A1 *sell into* A2. Although these usages may not be the most common ones, they are not ungrammatical either, and could easily appear in a larger language sample: e.g., *they elected him president*; *he ... discloses it to others*, *rivals ... sell into that market* (examples taken from the BNC). This suggests that our model is able to find reasonable generalisations using the input. At the same time, some occasionally produced verbs are ungrammatical, such as A1 *send about* A2, A1 *listen of* A2, etc. This happens because the model's exposure to the target construction is limited in terms of participating verb types, and there may not be enough support for making correct generalisations. Besides, as we argue below in this section, verb semantic representations in the input data are not rich enough. This is why the model overgeneralises and produces such ungrammatical usages. However, as we mentioned, the ungrammatical usages tend to appear at the bottom of the list, and do not compromise the model's performance on the verb production task. Besides, the difference between the frequencies of verb production in L1 and L2 simulations is very close to the value reported by Römer et al. (2014), which further defends the performance of our model on this task. Nevertheless, the fact that we could not compare the model's performance to human data in terms of specific verbs leaves the possibility that the model does not perform exactly like humans in the target task.

As for the correlations and the combined regression analysis, the frequency of production of verbs in our simulations can be predicted by joint verb–construction frequency, ΔP-contingency, and to some extent by verb semantic prototypicality. However, prototypicality does not correlate with the production frequency in all constructions, and its contribution to predicting production frequency is smaller than in EOR's studies. We propose three possible explanations of this result.

The first explanation is that our computational model does not rely on this factor to the extent human speakers do when generating verbs in constructions. This, indeed, may be the case, because the predicate semantics is only one out of many features in our representation of verb usages (recall Table 2). In other words, our model may underestimate the importance of the verb meaning in learning argument structure constructions. Note, however, that EOR in one of their studies (N. C. Ellis et al., 2014b) also did not observe significant correlations between the production frequency and semantic prototypicality for 5 out of 17 constructions in the data obtained from L1 English as well as L1 German speakers. In our simulations prototypicality was correlated with the production frequency in 23 out of 44 constructions, and it had an independent contribution in all the regression models reported above.

The second explanation relates to the type of semantic representations that the model operates on. Human speakers are often believed to possess fine-grained semantic representations of verbs: for example, Pinker (2013) proposes such narrow semantic rules as "transfer of possession mediated by separation in time and space" (p. 129). In contrast, semantic representations in our data set are extracted from WordNet and VerbNet and are more simplistic than that (e.g., give: {POSSESSION, TRANSFER, CAUSE}). This is not critical for the simulated learning process, because the discrimination between different verbs is supported by other features in the data, such as arguments' thematic proto-roles. However, in our analysis the prototypicality values are computed based on the verb semantics only, and the impoverished semantic representations may lead to the lower impact of semantic prototypicality in our study.

Our final explanation relates to how the prototypicality measure operates on a large and dense (as in EOR's study) vs. a small and sparse data set (as in our study). EOR computed semantic prototypicality of a verb in a construction based on a rich semantic network of all verbs that appear in this construction in the BNC. BNC is a rather large source, and it is unlikely that EOR's participants, given a construction, would produce a verb which is unattested in this construction in the BNC. In contrast, some constructions in our data set appeared with only a few verb types, in which case the prototypicality values were computed based on a rather small set of these few verbs. Yet the model often produced verbs which were unattested in this construction (non-members), but were semantically similar to other verbs that did appear in the target construction (members). To give an example, a construction ARG1 VERB ARG2 *for* ARG3 appeared in our data set with only five verbs: *substitute*, *elect*, *hail*, *criticise*, and *remove*. In the production task, the model generated these five verbs rather frequently, but there were other frequent verbs, in particular *praise*, *chastise*, and *indict*. Clearly, these verbs are allowed in the target construction, partly because they are somewhat synonymic to the construction members, at least when used in the target context (*to* VERB *someone for a reason*): *chastise* and *indict* are similar to *criticise*, while *praise* is similar to *hail*. In fact, the non-members must have been included into the target set of verbs, and the semantic prototypicality of all the verbs must have been calculated on this extended set. Since we had no way to predict beforehand which verbs would be produced by the model (and thus, should be included into the set), we computed all prototypicality values on the smaller set of verbs. This was particularly the case for the six constructions shared between our data set and EOR's data set: recall that these constructions appeared in the input with only a few verb types. As a result, prototypicality values for such constructions might not be very objective, hence the rather low contribution of this variable to predicting the frequency of verb production. At the same time, the correlation between prototypicality and production frequency is also very small for some frequent constructions, such as ARG1 VERB ARG2 and ARG1 VERB (at the top of Table 4), which can not be explained by the account outlined above. We believe this has to do with the incoherence of semantic networks for such constructions, and we leave this issue for the final discussion.

The small effect of semantic prototypicality in data simulated by our model should be addressed in the future; for now it is important to keep in mind that the reported impact of semantic prototypicality in the current study may be underestimated. Apart from the described limitation, our model was able to replicate the main effects reported in the original studies, both for L1 and L2. In the next section we address the two methodological issues of the original study discussed earlier (see *Methodological issues* above).

Addressing the methodological issues: Individual variation

In this second analysis we take into account the individual variation in the linguistic input, while trying to keep the rest of the design as close as possible to the previous analysis. We use the same set of simulated learners described in the previous section to predict verb production by the three target variables. The differences from the previous analysis are described next.

Methodological details

This time we do not calculate the cumulative frequency of production of each verb in a specific construction, $PF(v,c)$, as we did earlier. Instead, for each verb produced by each simulated leaner we define a binary outcome variable, which is set to 1 if the probability of production of this verb equals at least .005 (the threshold value from the previous analysis), and to 0 otherwise. This way, we now do not combine the data from all learners into a single $PF(v,c)$ value, but instead have data from individual simulated learners, while keeping the rest of the design very close to what was reported in the previous section. Besides, we compute the values of the three target variables—$F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$—for each simulation individually, based on a specific input sample. To keep up with the previous analysis, we apply the same data transformations as described before. To account for potential individual variation between constructions and learners, we use logistic mixed effects models with the binary outcome variable described above, with $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ as fixed factors, and with constructions and learners as random factors. All the mixed effects models for both L1 and L2 simulated data were fit to the two data sets: EOR's constructions only, and the whole data set, just as in the previous section. We started from maximal random effect structure with the random intercept and three random slopes (for each predictor), however the maximal model only converged for EOR's subset of L2 simulated data, therefore we removed some random slopes.

Results

The results are provided in Table 6. We did not use the LMG relative importance measure from the previous analysis, because it could not be applied to mixed effects models. In this set of models the β-coefficients for $\Delta P_A(v,c)$ are generally small (0.02 to 0.08), with the exception of model fitted to EOR's constructions in L1 simulations ($\Delta P_A(v,c) = 0.26$). However, even in the latter case the respective SE value is rather high (0.18), suggesting high variation in the data regarding the effect of $\Delta P_A(v,c)$. Besides, there is substantial variability among the coefficients for $Prt(v,c)$: between -0.08 and 0.38. The coefficients are greater in the models fitted to all constructions (0.38 and 0.37), compared to the models fitted to EOR's constructions only (0.10 and -0.08). Note that, surprisingly, in the latter case this coefficient has a negative value, however the respective variation in the data is high again ($SE = 0.09$). Besides, the respective model (fitted to EOR's constructions in L2 simulations) is the only one which includes random slopes for $Prt(v,c)$ over individual constructions and individual learners (see Table 6c), suggesting that some of this variation may come from accounting for the individual variation in the data.

Interim discussion

The models reported above predict verb production while taking into account differences in individual linguistic experiences of simulated learners. By comparing this kind of analysis to the original one, we can investigate whether taking into account individual variation may potentially lead to different results. Although our goal was to keep the data and the analysis maximally consistent with the previous setup, there are still differences in the type of outcome variable used (numeric production frequency vs. binary outcome) and, as a result, in the type of models fitted to the data

Table 6: Summary of the mixed effects models accounting for individual language experience

a. L1 simulations: constructions present in EOR's data set

$Prod. \sim F + \Delta P + Prt + (1 + \Delta P|learner) + (1 + \Delta P|constr.)$

Variable	β	SE^{a}	$95\%CI^{a}$	VIF
$F(v,c)$	0.58	0.01	[0.56, 0.60]	1.03
$\Delta P_A(v,c)$	0.26	0.18	[−0.09, 0.61]	1.00
$Prt(v,c)$	0.10	0.02	[0.06, 0.13]	1.03

b. L1 simulations: all constructions

$Prod. \sim F + \Delta P + Prt + (1|learner) + (1|constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.89	0.00	[0.88, 0.90]	1.95
$\Delta P_A(v,c)$	0.02	0.00	[0.01, 0.02]	2.01
$Prt(v,c)$	0.38	0.01	[0.36, 0.39]	1.07

c. L2 simulations: constructions present in EOR's data set

$Prod. \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt|learner) + (1 + F + \Delta P + Prt|constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.75	0.06	[0.62, 0.87]	1.32
$\Delta P_A(v,c)$	0.08	0.06	[−0.04, 0.20]	1.41
$Prt(v,c)$	−0.08	0.09	[−0.26, 0.09]	1.09

d. L2 simulations: all constructions

$Prod. \sim F + \Delta P + Prt + (1|learner) + (1|constr.)$

Variable	β	SE	$95\%CI$	VIF
$F(v,c)$	0.89	0.01	[0.88, 0.90]	1.42
$\Delta P_A(v,c)$	0.05	0.00	[0.04, 0.06]	1.50
$Prt(v,c)$	0.37	0.01	[0.35, 0.39]	1.07

[a] Due to the large sizes of the data sets, the reported SE and CI values for all the models are approximate, based on the Wald tests (Bates et al., 2015).

(linear vs. logistic regression). This does not allow us to compare coefficients pairwise across the two types of analysis, however the general pattern of difference suggests that the effect of ΔP-contingency may not be as high as predicted earlier, as soon as individual variation is taken into account.

The results on the individual variation in terms of semantic prototypicality are somewhat inconclusive. On the one hand, the positive effect of semantic prototypicality is present in the new models fitted to the full data sets, both in L1 and L2 simulations, and in the new model fitted to EOR's constructions in L1 simulations. On the other hand, there is not enough evidence for such effect in EOR's constructions obtained from L2 simulations. This must relate to whether the respective prediction model accounts for the variation between individual learners regarding this factor: we fitted an additional model to the same data, this time without the random slope for prototypicality over individual learners, and this model did predict a positive effect of semantic prototypicality. In other words, our data suggest that semantic prototypicality may play a role for some learners, but not for others.

Addressing the methodological issues: Order of preference

In the third set of analyses we look into the order of verb production by the same simulated learners, trying again to keep the rest of the design as close as possible to the original procedure.

Methodological details

In this set of analyses we record the actual probability of production of each verb by each simulated learner in each construction and then compute the cumulative probability, $PP(v,c)$, using it as the outcome variable in regression, instead of cumulative frequency. Cumulative frequency of a verb only shows how many times it is produced overall, while cumulative probability preserves the order of verb production by adding up the actual values of verb production probability for each learner. Unlike in the previous section, we are not interested in the variation between learners' individual experiences, therefore we use the values of $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ computed for the overall data set, to keep this analysis as close as possible to the original one. Again, we use the threshold value of .005 and apply the same data transformations as before. To account for the variation between constructions, we use linear mixed effects models with $PP(v,c)$ as the outcome variable, with $F(v,c)$, $\Delta P_A(v,c)$, and $Prt(v,c)$ as fixed factors, and with the random intercept and three random slopes (for each predictor) over individual constructions. One random slope has been removed from one final model to ensure its convergence. The rest of the analysis follows the originally outlined procedure.

Results

The summaries of the prediction models are provided in Table 7. Just as in the previous set of analyses, we can see that the effect of ΔP-contingency is small (the greatest β-coefficient is 0.03), and even negative (-0.04) for one of the models. Besides, in all cases the respective 95%CI includes 0, suggesting that the contributions of ΔP-contingency are not significant in these models.

In other respects this new set of models is similar to the original analysis. The other two factors, joint frequency $F(v,c)$ and prototypicality $Prt(v,c)$, have their independent contributions, although in one case the 95%CI for prototypicality includes 0. The overall fit of the models to the data is

Table 7: Summary of the mixed effects models accounting for the order of verb preference

a. L1 simulations: constructions present in EOR's data set

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt|constr.)$

Variable	β	SE[a]	95%CI[a]	VIF
$F(v,c)$	0.56	0.08	[0.41, 0.72]	1.74
$\Delta P_A(v,c)$	−0.04	0.09	[−0.21, 0.13]	1.70
$Prt(v,c)$	0.06	0.05	[−0.03, 0.15]	1.39
$R^2_m = .28$, $R^2_c = .34$[b]				

b. L1 simulations: all constructions

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt|constr.)$

Variable	β	SE	95%CI	VIF
$F(v,c)$	0.84	0.04	[0.77, 0.93]	1.86
$\Delta P_A(v,c)$	0.03	0.02	[−0.01, 0.07]	1.96
$Prt(v,c)$	0.14	0.02	[0.10, 0.18]	1.09
$R^2_m = .47$, $R^2_c = .64$				

c. L2 simulations: constructions present in EOR's data set

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P|constr.)$

Variable	β	SE	95%CI	VIF
$F(v,c)$	0.53	0.08	[0.37, 0.68]	1.64
$\Delta P_A(v,c)$	0.02	0.08	[−0.13, 0.18]	1.63
$Prt(v,c)$	0.14	0.04	[0.06, 0.22]	1.02
$R^2_m = .32$, $R^2_c = .37$				

d. L2 simulations: all constructions

$PP \sim F + \Delta P + Prt + (1 + F + \Delta P + Prt|constr.)$

Variable	β	SE	95%CI	VIF
$F(v,c)$	0.85	0.05	[0.75, 0.95]	2.20
$\Delta P_A(v,c)$	0.03	0.02	[−0.01, 0.08]	2.25
$Prt(v,c)$	0.14	0.02	[0.10, 0.17]	1.05
$R^2_m = .47$, $R^2_c = .66$				

[a] The reported SE and CI values are estimated via parametric bootstrap with 1,000 resamples (Bates et al., 2015).

[b] R^2_m and R^2_c stand for marginal and conditional R^2 coefficients.

lower than reported in our first analysis: they explain 34 to 66% of the variance in the data (see R_c^2 values in the table), and only 28 to 47% of this is explained by the fixed factors (R_m^2 values): to compare, the overall fit of the models in the original analysis varies between 50 and 90%. [7]

Interim discussion

The models reported above predict the cumulative probability of verb production by the simulated learners. Unlike the originally reported models (see *Simulating the original experiments*), this type of analysis accounts for the order of verb preference by our simulated L1 and L2 learners. Most importantly, none of the four models suggest that $\Delta P_A(v,c)$ is an independent predictor, when the order of verb production is taken into account. Recall that both joint frequency and ΔP-contingency are measures of the contextual frequency: this may explain why we do not observe the independent effects of both measures at the same time. Indeed, the approximate correlation coefficient between βs for joint frequency and ΔP-contingency (this coefficient is not included into the tables) appears to be rather large, between -0.50 and -0.74. In other words, the higher the β for frequency, the lower the β for ΔP-contingency, and vice versa.

The poorer fits of the models support our idea that there is space for refining the original prediction model used so far in the analyses: another set of variables may explain the data better without predicting so much random variation between constructions. We will investigate this issue in the next section.

Refining the prediction model

Our next goal is to test whether there is a better set of predictors explaining the production data. Based on our theoretical overview, we have three issues to address. First, theoretical accounts suggest that the marginal verb frequency may play an independent role in verb selection, therefore we believe that including marginal frequency into the prediction model would improve its fit to the data. Second, the presence of two contextual frequency (association) measures in the model may not be well justified, and eliminating one of them might not necessarily damage the model. Finally, there are multiple measures of contextual frequency, three of which we plan to test: joint frequency, ΔP (as in the previous analyses), and Attraction.

Methodological details

We start by fitting a number of mixed effects models of the type described in the second analysis (logistic models taking into account individual differences) and in the third analysis (linear models taking into account order of preference). To ensure that the models generalise well over different constructions, we use the full set of constructions for fitting each model, and not EOR's subset.

The structure of fixed factors in the models is defined as described below.

			I		**II**		**III**
(m1a)	$Production$	~			$joint\,freq. \times \Delta P\text{-}assoc.$	×	$protot\text{-}ty$
(m2a)	$Production$	~			$joint\,freq. \times attraction$	×	$protot\text{-}ty$
(m3a)	$Production$	~			$attraction \times \Delta P\text{-}assoc.$	×	$protot\text{-}ty$
(m4a)	$Production$	~			$joint\,freq.$	×	$protot\text{-}ty$
(m5a)	$Production$	~			$attraction$	×	$protot\text{-}ty$
(m6a)	$Production$	~			$\Delta P\text{-}assoc.$	×	$protot\text{-}ty$
(m1b)	$Production$	~	$verb\,freq.$	×	$joint\,freq. \times \Delta P\text{-}assoc.$	×	$protot\text{-}ty$
(m2b)	$Production$	~	$verb\,freq.$	×	$joint\,freq. \times attraction$	×	$protot\text{-}ty$
(m3b)	$Production$	~	$verb\,freq.$	×	$attraction \times \Delta P\text{-}assoc.$	×	$protot\text{-}ty$
(m4b)	$Production$	~	$verb\,freq.$	×	$joint\,freq.$	×	$protot\text{-}ty$
(m5b)	$Production$	~	$verb\,freq.$	×	$attraction$	×	$protot\text{-}ty$
(m6b)	$Production$	~	$verb\,freq.$	×	$\Delta P\text{-}assoc.$	×	$protot\text{-}ty$

In all the equations above, component I represents the marginal verb frequency, component II comprises contextual frequency measures, and component III is the semantic prototypicality. We start with the original model tested in the previous sections, m1a. Models m2a–m3a resemble m1a, but they test alternative pairs of the three contextual frequency measures. Models m4a–m6a, in contrast, eliminate one of the contextual frequency measures, keeping only one. Finally, the other six models (m1b–m6b) mirror models m1a–m6a, respectively, but add the marginal frequency measure to their counterparts. Note that the models are multiplicative due to the log-transformation of all the variables: $log(y) = log(a) + log(b) + log(c) \Rightarrow y = abc$. Studying and interpreting interactions between variables in such models is not straightforward, and for simplicity we do not include any interaction terms in the prediction models.

We compare the fit of all the 12 models using their corrected Akaike information criterion (AICc), as implemented in R (Bolker & R Development Core Team, 2016). This is a common method to compare models in a multimodel inference paradigm (Burnham & Anderson, 2002).[8]

Results: model comparison

The ranked list of the models with their respective AICc values is provided in Table 8, which is also visualised in Figure 4.

First we have to note that models m2a–m3a and m2b–m3b in some cases yielded multicollinearity problems. This was caused by the presence of two contextual frequency measures in these models, which sometimes were highly correlated even after applying the data transformations. The models which show this problem, even if ranked rather high, may not be very informative in terms of their coefficients.

Further, we notice that the order of the models in the four lists is not identical, although there are clear similarities. The original model m1a is far from being the best one in any list. A pairwise comparison of the models demonstrates that m1b–m6b, which include the marginal verb frequency $F(v)$, always fit the data better than their respective counterparts without $F(v)$: m1a–m6a. In other words, adding $F(v)$ to any model improves its fit. If we further look only at the ranks of the "better" models m1b–m6b, we can see that the models with two contextual frequency measures (m1b–m3b) generally outperform the models with only one such measure (m4b–m6b). The only exception from this pattern is the single-measure model m4b, which is ranked third in each list, always higher than m3b. In all the four lists, the best model is m2b, therefore we look at this model in more detail in the following section.

Table 8: Model rankings

	L1 data				L2 data			
	Individual differences		Order of preference		Individual differences		Order of preference	
Rank	Model	$\Delta AICc$	Model	$\Delta AICc$	Model	$\Delta AICc$	Model	$\Delta AICc$
1	m2b	0	m2b*	0	m2b	0	m2b*	0
2	m1b	2,601	m1b	20	m1b	1,893	m1b	21
3	m4b	5,842	m4b	39	m4b	5,236	m4b	41
4	m2a	9,867	m3b*	57	m2a	5,810	m3b*	66
5	m1a	13,892	m5b	131	m1a	8,282	m5b	135
6	m4a	16,539	m6b	614	m4a	11,475	m6b	579
7	m3b*	23,403	m2a*	846	m3b*	19,627	m2a*	749
8	m5b	34,996	m1a	855	m5b	29,330	m1a	760
9	m3a*	36,100	m4a	858	m3a*	30,887	m4a	762
10	m5a	55,234	m3a*	921	m5a	43,980	m3a*	833
11	m6b	64,844	m5a	1,023	m6b	53,950	m5a	932
12	m6a	93,828	m6a	1,496	m6a	72,918	m6a	1,363

* Models which showed multicollinearity problems ($VIF > 3$ for some predictors).

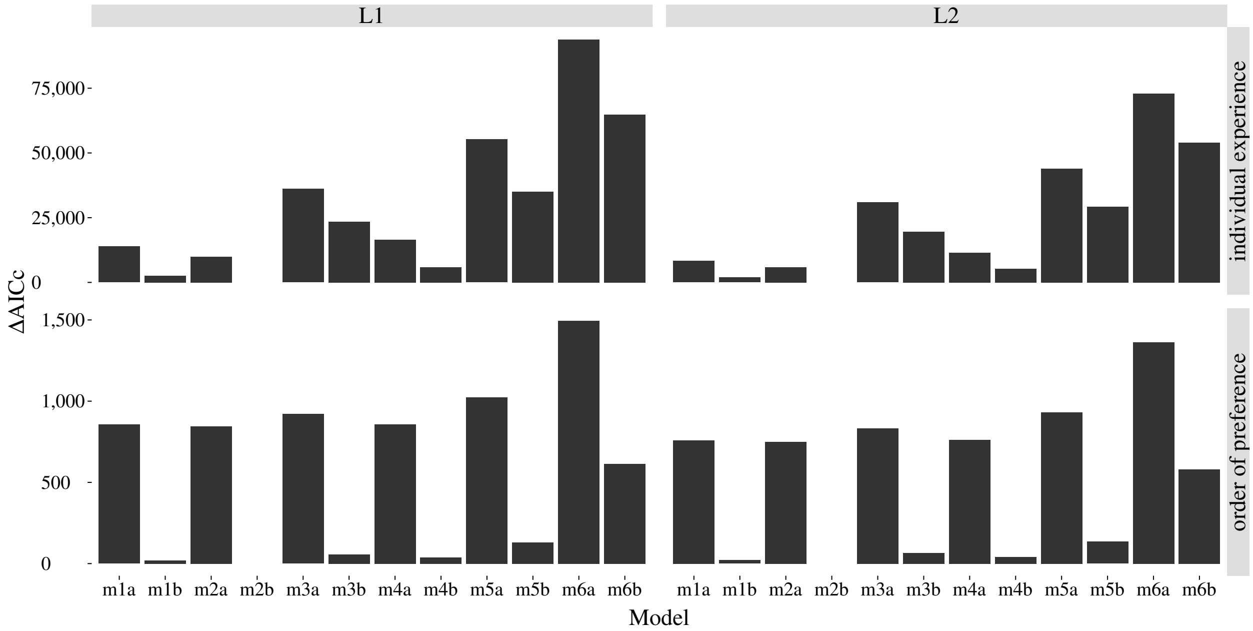


Figure 4: Model rankings visualised. $\Delta AICc$ for a model M in each subplot shows the difference between the $AICc$ of the best model in that subplot and the $AICc$ of the model M. $\Delta AICc$ of the best model in each subplot is 0, and higher $AICc$ values correspond to worse model fits.

Predictive power of each factor

To look at the impact of individual predictors in the refined model, in Table 9 we provide the summary of the model m2b ranked highest in each list. To account for the random variance, we refit m2b to each data set, this time including random slopes for each predictor (linear models accounting for order of preference), or random intercepts (logistic models accounting for individual variation).

Looking at the summary in Table 9, we first observe that the four fixed factors in the linear models explain 52% of the variance for L1 data, and 53% for L2 data (see R^2_m coefficients). This is higher compared to the original prediction models for the same data sets (47%, see section *Addressing the methodological issues: Order of preference*).

Next, we can see that all the models yield collinearity problems: the variance inflation factor for $A(v,c)$ and $F(v,c)$ varies between 5.14 and 6.79. This suggests high collinearity between the two predictors. This is supported by the high correlation between βs for $A(v,c)$ and $F(v,c)$ in all models, varying between -0.89 and -0.91. Considering that the random slopes for $A(v,c)$ and $F(v,c)$ could not be included into the logistic models, the random variation in these models may be underestimated. In sum, even though $A(v,c)$ and $F(v,c)$ demonstrate their independent effects in the two logistic models, the respective β-coefficients may not be very informative.

The coefficients for $Prt(v,c)$ in linear models are also rather small, 0.10 and 0.08. Most importantly, the effect of $F(v)$ is high in all the models.

Interim discussion

The comparison of prediction models supports our proposal that the marginal verb frequency plays an independent role in predicting verb production in our simulated data. The parallel use of two contextual frequency measures appears to improve the model fit overall, contrary to our expectations. Yet, including two contextual frequency measures leads to collinearity issues: there is often a trade-off between the overall fit of the model to the data and the informativeness of its β coefficients. The use of a single measure is supported by our analysis of individual predictors, which suggests that the contextual frequency can be considered as a single component: joint frequency and Attraction capture the same type of syntagmatic relation between verbs and constructions. In other words, it is the combined effect of contextual frequency which is important, but not the individual effect sizes of joint frequency and Attraction. If one needs to chose a single contextual frequency measure between joint frequency, Attraction, and ΔP-contingency, our analysis suggests that joint frequency is the best measure: recall the high ranks of model m4b.

Considering contextual frequency as a single component, its individual impact in all the models is the highest, compared to the other predictors. The impact of prototypicality appears to be rather small in some refined models, but so it is in the original models as well: again, recall that our computational model may underestimate the importance of this factor.

General discussion

In this study we examined whether the selection of verbs within constructions could be explained by the distributional and semantic properties of these verbs and constructions, to see which factors may be responsible for establishing links between verbs and constructions in speakers' minds.

Table 9: Summary of the best models of m2b type

a. L1 simulations: model accounting for individual differences

$Prod. \sim F(v) + F(v,c) + A + Prt + (1|learner) + (1|constr.)$

Variable	β	SE^{a}	$95\%CI^{\text{a}}$	VIF
$F(v)$	0.64	0.01	[0.63, 0.66]	1.03
$F(v,c)$	0.65	0.01	[0.64, 0.67]	6.48
$A(v,c)$	0.11	0.00	[0.10, 0.11]	6.54
$Prt(v,c)$	0.33	0.01	[0.31, 0.35]	1.07

b. L1 simulations: model accounting for order of preference

$PP \sim F(v) + F(v,c) + A + Prt + (1 + F(v) + F(v,c) + A + Prt|constr.)$

Variable	β	SE^{b}	$95\%CI^{\text{b}}$	VIF
$F(v)$	0.29	0.04	[0.21, 0.37]	1.36
$F(v,c)$	0.73	0.10	[0.53, 0.92]	5.14
$A(v,c)$	0.07	0.05	[−0.02, 0.17]	5.78
$Prt(v,c)$	0.10	0.02	[0.06, 0.14]	1.14
$R^2_m = .52$, $R^2_c = .72$				

c. L2 simulations: model accounting for individual differences

$Prod. \sim F(v) + F(v,c) + A + Prt + (1|learner) + (1|constr.)$

Variable	β	SE^{a}	$95\%CI^{\text{a}}$	VIF
$F(v)$	0.63	0.01	[0.62, 0.66]	1.06
$F(v,c)$	0.60	0.01	[0.58, 0.62]	6.29
$A(v,c)$	0.16	0.01	[0.15, 0.17]	6.31
$Prt(v,c)$	0.33	0.01	[0.32, 0.35]	1.07

d. L2 simulations: model accounting for order of preference

$PP \sim F(v) + F(v,c) + A + Prt + (1 + F(v) + F(v,c) + A + Prt|constr.)$

Variable	β	SE^{b}	$95\%CI^{\text{b}}$	VIF
$F(v)$	0.30	0.04	[0.22, 0.38]	1.45
$F(v,c)$	0.77	0.11	[0.55, 1.00]	6.18
$A(v,c)$	0.06	0.05	[−0.05, 0.16]	6.79
$Prt(v,c)$	0.08	0.02	[0.05, 0.12]	1.13
$R^2_m = .53$, $R^2_c = .75$				

[a] Values are based on the Wald tests.

[b] Values are estimated via parametric bootstrap with 1,000 resamples.

We started from adopting the proposal by EOR that the frequency of production of a verb in a construction can be predicted by the joint verb–construction frequency, the contingency of verb–construction mapping, and the prototypicality of the verb meaning. In what follows, we first briefly recapitulate how our simulations are similar and dissimilar to the human data. Since semantic prototypicality is the main issue in this respect, we discuss it next. The discussion is continued with a comparison of the results across three types of analysis provided above. Next we explain how the prediction model can be improved by avoiding multiple measures of contextual frequency, and by including marginal frequency instead. Additionally, we discuss how the use of form-based representations of constructions may have affected the findings, and address other theoretical challenges. Finally, we briefly talk about the computational model used in this study, and provide a short conclusion.

Simulations vs. human data

We used a computational model of construction learning to simulate the verb production experiments from EOR's studies. The analysis of verbs produced in the computational simulations demonstrated the model's reasonable performance on the target task: given a construction, the model mostly produced verbs that had been attested in this construction in the input. There were some exceptions, which suggest that the model was able to perform sensible generalisations over individual verb usages. At the same time, the type of the input data used in this study made it impossible to directly compare the verbs produced by the model to those produced by human participants, suggesting that we can not claim that the model exactly replicated human linguistic behaviour in the target task.

Our initial correlational and regression analyses showed main effects similar to those in the original experiments of EOR. In particular, we observed independent contributions of all the three predictors to explaining the frequency of verb production. Additionally, a preliminary comparison of the verb lists produced by the model in L1 vs. L2 simulations demonstrated that the degree of difference between the two lists was similar to that reported by Römer et al. (2014) for native German vs. native English speakers. However, a qualitative comparison between the simulated L1 and L2 verb lists is still needed. The main difference between the results obtained in our simulations and those reported by EOR related to the effect of semantic prototypicality, which appeared to be lower in our simulated data. We discuss this issue next.

Meaning prototypicality, data sparsity, and semantic coherence

We proposed three possible explanations for the low impact of semantic prototypicality: (1) the role of verb semantics is underestimated in the learning algorithm used by our model; (2) verb semantic representations in our data sets are impoverished compared to those in human speakers; (3) our semantic prototypicality measure performs poorly on infrequent constructions due to the data sparsity. Regarding the last explanation, we also found that the correlations between semantic prototypicality and verb production frequency were also low within some frequent constructions in our data set, for which dense information on verb use was available: ARG1 VERB and ARG1 VERB ARG2. We suggest this has to do with the degree of semantic coherence of a construction. Following the setup of the original studies, we have used highly abstract constructions defined by their shallow form, which may not be semantically coherent. In particular, if we look at the verbs produced within the most frequent construction ARG1 VERB ARG2, these comprise several semantic groups: verbs of mental state (e.g., *want*), verbs of transfer (e.g., *buy*, *sell*), verbs of communication (e.g., *announce*), and many others. Given this variety, the construction is unlikely to have a single semantic core surrounded by multiple peripheral verbs. Instead, there are multiple semantic centres, and a single measure of semantic prototypicality may not capture such organisation well, in particular when some semantic verb classes within a construction are much richer than others. This might be why we do not observe an effect of prototypicality in such constructions. In contrast, the effect is larger in constructions whose semantics is more coherent, because they actually have a single "prototypical" core. To give an example, the ARG1 VERB ARG2 ARG3 construction in our data (which comprises ditransitive verb usages, but also allows for adverbial arguments) is represented by eight verbs: *drag*, *give*, *hang*, *lead*, *place*, *pull*, *send*, and *tell*. Most of these are physical action verbs, the only exception being *tell*, hence high semantic coherence and a high effect of semantic prototypicality.

To compare, Theakston et al. (2004) in their study of early verb use did not find enough support that semantic prototypicality of a verb could predict the age when this verb first appeared in the child's speech, and the constructions they used—SVO, VO, and the intransitive—were highly abstract, and thus unlikely to be semantically coherent. This may also explain why the prototypicality effect was observed in the studies of EOR: they only focused on various constructions with locative semantics in their analyses, which may be more semantically coherent.

The question whether the effect of prototypicality is related to the degree of semantic coherence of a construction requires further investigation. As a counter-argument to this claim, Ambridge, Bidgood, Pine, Rowland, and Freudenthal (2015) find the effect of semantics in the passive, a semantically general construction. Note, however, that the interpretation of semantics in their study (and in other related studies: e.g., Ambridge et al., 2012, 2014) differs from semantic prototypicality as defined in this study. The reasoning behind this study (following EOR) is that more prototypical verbs are produced more frequently (because of how the activation spreads within a semantic network). This is why semantic verb features used in our study must capture the essential properties of the respective events. In contrast to this, the idea in the series of studies mentioned above is that particular nuances of verb meanings help in acquiring restrictions on the verb use. Therefore, these studies focus on very specific fine-grained features of a verb meaning, which do not necessarily provide much information about the general semantics of the event, but do help in discriminating between different verbs or verb classes. This account is largely based on Pinker's (2013) theory, in which "it's not what possibly or typically goes on in an event that matters; it's what the verb's semantic representation is choosy about in that event that matters" (p. 127). For this reason, the effect of semantics in this study and in EOR's study is not immediately comparable to the findings of Ambridge and colleagues. Building more comprehensive verb meaning representations based on both general event features and fine-grained discriminatory features could open new prospects in this area: such representations could be used for training both our computational model and the model of Ambridge and Blything (2015).

Comparing the results across three types of analysis

We further carried out two additional analyses, to account for the potential between-learner variation in the linguistic input, and for the order of verb production by each (simulated) learner. These additional analyses of our simulated data suggest that the type of analysis may affect the main findings, in particular in terms of the observed effect of ΔP-contingency, which we address below. This is consistent across the two additional analyses, suggesting that both individual variation and order of learners' preference is important, which is in line with studies suggesting that individual differences play a role in language learning (e.g., R. Ellis, 2004), and that speakers do not arrive at the same mental grammar (e.g., Dąbrowska, 2012; Misyak & Christiansen, 2012). To verify the predictions made by our model in this respect, we would need to compare the results to human empirical data on individual variation and order of preference, which are missing yet.

Multiple measures of contextual frequency

Contingency may sometimes fail to demonstrate its independent effect because of the other variable included into the prediction model: joint verb–construction frequency. Both variables capture how well a verb and a construction go together (i.e., contextual frequency). If the hypothesised cognitive effect of the verb–construction association is loaded on both variables, one of them may show no independent impact. This issue was addressed by testing a number of alternative prediction models. One of our questions was whether models with one or with two contextual frequency measures would predict the data better. Our findings in this respect were somewhat inconclusive. On the one hand, prediction models which included two such measures were in general ranked higher than models which included only one measure. On the other hand, the independent effects of both joint frequency and contingency were not always present within the same prediction model. In fact, it was the combined impact of the two measures that was consistent across prediction models, but not the independent effect of each contextual frequency measure. This is why we suggest that it is a single effect of the contextual frequency that is cognitively plausible, while each measure (i.e., joint verb–construction frequency, Attraction, or ΔP-contingency) provides a particular quantitative representation of this effect. The correlation between the measures may be lower or higher in a specific data set, and this is why sometimes, but not always, it is justified to include two contextual frequency measures into a prediction model.

The relation between association strength and joint verb–construction frequency may also resemble the relation between the effects of entrenchment and preemption on learning argument structure restrictions, described by Ambridge, Bidgood, Twomey, et al. (2015). Both the entrenchment and preemption hypotheses predict that the distribution of verbs over argument structure constructions affects the learning of the related usage restrictions, because of the verb's occurrence in either competing constructions (preemption hypothesis), or in all constructions (entrenchment hypothesis). In fact, independent contributions of these two factors within the same prediction model have been sometimes found (e.g., Ambridge, 2013; Blything et al., 2014). Yet, Ambridge, Bidgood, Twomey, et al. (2015) suggest that entrenchment and preemption are not independent mechanisms, but only effects that may or may not be observed, depending on the exact set of constructions in a study. Similarly, the effects of both association strength and joint frequency in our study capture the same mechanism of competition between verbs in the speaker's mind.

Whenever a single measure of contextual frequency must be considered in the analysis, our study supports joint verb–construction frequency as the best measure, although more research is needed in this respect. In particular, a more advanced factor analysis (e.g., of the type employed by Maki & Buchanan, 2008) may clarify the relationship between different measures of contextual frequency.

Marginal verb frequency

The results in terms of marginal (overall) verb frequency are more straightforward. We found a consistent effect of the marginal verb frequency, in line with some data in language acquisition research (Blything et al., 2014; Theakston et al., 2004). Besides, this effect was independent from that of joint verb–construction frequency, in accordance with the proposed distinction between cotextual and cotext-free entrenchment (Schmid, 2010; Schmid & Küchenhoff, 2013). Based on this result, the effect of marginal verb frequency is worth investigating in human production data. In particular, this is theoretically supported by some existing memory research (Hockley & Cristi, 1996; Madan et al., 2010), where item memory (reflected in our case in marginal verb frequency) is believed to be independent of associative memory (in our case: contextual frequency measures).

At the same time, the marginal frequency of a verb may relate to the diversity of syntactic environments in which this verb is used. Although some frequent verbs may be used in only a few types of constructions, in general a verb's frequency is likely to be higher when the verb is used in a great variety of construction types. In this capacity, the observed effect of the marginal verb frequency in our study may be similar to what Naigles and Hoff-Ginsberg (1998) report in their child language study: verbs which appear in diverse syntactic frames are used more frequently.

Speaking about the effect of marginal verb frequency compared to that of contextual frequency, our data suggests that contextual frequency has a higher impact on verb selection than marginal frequency. This is a rather reasonable conclusion: when cued by a construction, speakers are more likely to produce frequent verbs related to the cue, rather than verbs which are frequent overall. However, if there are two verbs fitting the construction equally well, the one which is more frequent overall will be preferred. This is consistent with the fact that constructions attract only some verbs and reject other verbs (e.g., Goldberg, 1995; Stefanowitsch & Gries, 2003).

Alternative construction representations

In this study, constructions were defined solely by their shallow form. This is a common approach in corpus linguistics, because it is easy to automatically look for syntactic forms in a corpus. An efficient search for constructional meanings, on the other hand, would only be possible in a corpus that is semantically annotated, which is most often not the case. At the same time, constructions are commonly defined as pairings of form and meaning (e.g., Croft, 2001; Goldberg, 1995; Langacker, 1987). Assuming a priori that a shallow pattern has a meaning does not guarantee that this meaning is unified and coherent, and that the hypothesised construction is cognitively real. Defining a construction by explicitly describing both its form and its meaning may be a better practice.

The described problem is particularly evident in the current study, as well as in EOR's studies. Form-based patterns do not predefine the argument roles, and therefore, could be interpreted by participants in multiple ways. This sometimes resulted in the production of verbs with different argument structures within the same pattern: e.g., *come* and *throw* in *he/she/it* __ *across the ...*; or *eat* and *write* in *he/she/it* __ *as the ...*; with some usages even looking ungrammatical: *he/she/it knows as the ...*, *he/she/it climbs of the ...*, etc. (data from English native speakers in N. C. Ellis et al., 2014a). Similarly, in our study multiple semantic interpretations were possible, for example, for ARG1 VERB ARG2 ARG3. Besides, the problem in both studies is reinforced by the use of both animate (*s/he*) and inanimate (*it*) pronouns as the subject of each test stimuli: it may be argued that the animate pronouns represent an AGENT, while the inanimate pronoun is more likely to be a FORCE, hence two different constructions.

This leads us to the issue of the level of granularity of constructional patterns. It has been suggested that observed frequency effects may depend on the level of granularity of a construction under consideration (Lieven, 2010). The issue has also been touched on by Theakston et al. (2004), who show that different researchers employ different constructions in similar studies: for example, Ninio's (1999a) vo and svo constructions are combined within the same transitive construction by Goldberg (1998). In other words, the results may be also conditional on the chosen level of granularity of constructions. Together, these issues call for a similar analysis of different constructional representations. An earlier study with the same computational model (Matusevych, Alishahi, & Backus, 2015a) suggests that the observed effects of input-related factors on verb selection depend, indeed, on the type of constructional representations. Yet, the issue requires further investigation.

Further theoretical challenges

This study additionally touches on some theoretical questions that need to be addressed in the future. One of them is the relation between naturalistic and experimental verb production data. In this study, just as in EOR, the production of verbs was elicited by constructional stimuli. This is different from related studies of verb production by children (e.g., Theakston et al., 2004; Naigles and Hoff-Ginsberg, 1998; Ninio, 1999a, 1999b), which work with naturalistic samples of child language. It is unclear whether such "field" data are directly comparable to the experimental data from elicited production experiments: for example, in the natural data some verbs within a construction may be used more often simply because of the higher referential frequency of the actions, states, etc. they refer to.

This leads us to the problem of defining the true nature of such phenomena as a unit's frequency, semantic prototypicality, and entrenchment. In this study we have simplistically assumed that a unit's frequency reflects its entrenchment, and that the frequency is independent of prototypicality, but these relations are not so trivial (Geeraerts, Grondelaers, & Bakema, 1994; Schmid, in press). To mention only some complications, when a unit is perceptually salient in speech (e.g., a word which is very unusual in a given genre or context), it may contribute more to memory consolidation (and entrenchment) than when it is less salient. Besides, it has been argued that the frequency (e.g., the referential frequency) does play a role in determining prototypicality (see an overview in Gilquin, 2006). Highly controlled studies of these phenomena could clarify the theory, and computational modelling can be helpful in this respect.

Computational model of construction learning

The final issue to address is the computational model employed in this study. On the one hand, simulation results always depend to a certain extent on the chosen model. To give an example from this study, semantics in our model is only one out of many features that guide construction learning, and the role of semantics may be underestimated compared to human learners. If that is indeed the case, then the differences in the size of effects reported in this study and in EOR's study may be attributed to the model's inability to replicate the exact linguistic behaviour of human speakers.

On the other hand, when the model, as in our case, produced results similar to some existing experimental findings, this supports the plausibility of the model. The similarity of our results based on L1 and L2 simulations to those of EOR supports the assumption that incidental learning takes place in both L1 and L2 learning. Besides, the fact that the model is able to produce verbs relevant for a given construction, suggests that the emergent constructional representations in the model may approximate well what humans learn. Unfortunately, the type of the input data used in the present study does not allow us to compare the production data to the original study in terms of specific verbs and constructions, and this issue should be addressed in the future to better evaluate the potential of this computational model. One fruitful direction may be to investigate the role of frequency vs. verb semantics in the process of learning verb–construction associations (as in Ambridge & Blything, 2015), as opposed to looking at the static knowledge of such associations in simulated speakers.

Conclusion

In this article we presented a computational simulation of the verb production experiments of N. C. Ellis et al. (2014a, 2014b) using a usage-based, probabilistic model of argument structure construction learning. Our experiments showed that the model's performance in the verb production task could by predicted by the same variables as the performance of human participants in EOR's experiments. Our follow-up analyses addressed some methodological limitations of these experimental studies, and suggest a refined version of the verb production model proposed by EOR. In particular, the frequency of production of verbs within argument structure constructions in our simulated data could be predicted by joint verb–construction frequency, contingency of verb–construction mapping, and prototypicality of verb meaning, although the effect of prototypicality was lower than in the human data. We then carried out two additional analyses on the same simulated data sets, to account for individual variation between speakers and for order of their verb preference. The results suggest that the type of analysis may affect the main findings. In particular, the effects of both joint verb–construction frequency and contingency measure within the same prediction model are not always observed. Finally, we compared a number of prediction models with different variables. The best prediction model included overall verb frequency in the input data, semantic prototypicality, and two contextual frequency measures: joint verb–construction frequency and Attraction. However, the high correlation between the contextual frequency measures suggests that their effects are combined rather than independent. We believe this refined prediction model should be tested on experimental data with human subjects.

Notes

[1] Note that in EOR's setup virtually no distinction is made between first (L1) and second language (L2) speakers. This is in line with the theories of incidental (statistical) language learning, and with the proposal in cognitive linguistics that much of the L2 learning relies on the same cognitive mechanisms used in L1 learning (N. C. Ellis & Larsen-Freeman, 2006; Ervin-Tripp, 1974; MacWhinney, 2012).

[2] We follow the existing literature in assuming that the entrenchment of a unit is a mere product of its frequency, although the impact of each individual use may, in fact, be strongly modulated by pragmatics (Schmid, in press).

[3] Note that we do not assign case-marking to personal pronouns (e.g., *me* = *I*-ACC), but use the actual forms used in the corpus instead. Given the exceptionally high token frequencies of these forms, it is sometimes argued that forms such as *I* and *me* co-exist in the speaker's lexicon, without *me* being derived from *I* (e.g., Diessel, 2007; Hudson, 1995).

[4] As an alternative, we tried to calculate the similarity between verb meanings using the actual sets of semantic elements used in our data sets, build a resulting network based on these similarity values for each construction, and then calculate betweenness centrality on this network. Recall, however, that many constructions in our data sets occurred

with only a few verb types: computing betweenness centrality on such a small network yielded an abundant number of 0s, which was damaging for our analysis.

[5] As in the original study, we add 0.01 to all the predictors as well as to the outcome variable. We additionally increment $\Delta P_A(v,c)$ by 1, to avoid having negative values in the data. The last step is necessary, because we log-transform all the variables as in EOR's studies. The log-transformation is justified by the fact that practice (which in our case is reflected in production frequency) is believed to be a power function of experience (Newell & Rosenbloom, 1981), and therefore a power transformation can linearise the relationship between $PF(v,c)$ and at least one of the predictors, namely $F(v,c)$.

[6] These coefficients indicate the amount of variance explained by the fixed factors and by the full model, respectively (Johnson, 2014), and are computed using an existing R implementation (Bartoń, 2016).

[7] For a fairer comparison of model fits across the two types of analysis, we also looked at the mixed effects models mentioned in the section *Simulating the original experiments*, and their fits were still higher than reported here.

[8] It has been argued (Greven & Kneib, 2010) that using AICc to compare models with different structures of random factors leads to a bias in favour of a more complex random factor structure. For this reason, to ensure the model comparison is fair, in linear models we only use random intercepts over individual constructions. In logistic models (accounting for individual differences) we would ideally use random intercepts over individual learners and constructions, but some of the models with random intercepts did not converge, therefore we used simple logistic regression without random effects.

References

Alishahi, A., & Pyykkönen, P. (2011). The onset of syntactic bootstrapping in word learning: Evidence from a computational study. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

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

Alishahi, A., & Stevenson, S. (2010). A computational model of learning semantic roles from child-directed language. *Language and Cognitive Processes*, *25*, 50–93. doi: 10.1080/01690960902840279

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

Ambridge, B., Bidgood, A., Pine, J. M., Rowland, C. F., & Freudenthal, D. (2015). Is passive syntax semantically constrained? Evidence from adult grammaticality judgment and comprehension studies. *Cognitive Science*. Advance online publication. doi: 10.1111/cogs.12277

Ambridge, B., Bidgood, A., Twomey, K. 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*. doi: 10.1371/journal.pone.0123723

Ambridge, B., & Blything, R. P. (2015). A connectionist model of the retreat from verb argument structure overgeneralization. *Journal of Child Language*. Advance online publication. doi: 10.1017/S0305000915000586

Ambridge, B., & Brandt, S. (2013). Lisa filled water into the cup: The roles of entrenchment, pre-emption and verb semantics in German speakers' L2 acquisition of English locatives. *Zeitschrift für Anglistik und Amerikanistik*, *61*, 245–263. doi: 10.1515/zaa-2013-0304

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–273. doi: 10.1017/s030500091400049x

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

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

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295. doi: 10.1016/b978-1-4832-1446-7.50016-9

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429. doi: 10.1037/0033-295x.98.3.409

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511801686

Bartoń, K. (2016). *Package 'MuMIn': Multi-Model Inference.* Retrieved from https://cran.r-project.org/web/packages/MuMIn/MuMIn.pdf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. doi: 10.18637/jss.v067.i01

Blumenthal-Dramé, A. (2012). *Entrenchment in Usage-based Theories: What Corpus Data Do and Do Not Reveal about the Mind.* Berlin: Walter De Gruyter. doi: 10.1515/9783110294002

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

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, *89*, 1–47. doi: 10.1037/0033-295x.89.1.1

Bolker, B., & R Development Core Team. (2016). *Package 'bbmle': Tools for general maximum likelihood estimation.* Retrieved from https://cran.r-project.org/web/packages/bbmle/bbmle.pdf

Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., ... Uszkoreit, H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, *2*, 597–620. doi: 10.1007/s11168-004-7431-3

Bryl, V., Tonelli, S., Giuliano, C., & Serafini, L. (2012). A novel Framenet-based resource for the semantic web. In S. Ossowski & P. Lecca (Eds.), *Proceedings of the 27th Annual ACM Symposium on Applied Computing* (pp. 360–365). New York, NY: Association for Computing Machinery. doi: 10.1145/2245276.2245346

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., & Pinkal, M. (2006). The SALSA corpus: A German corpus resource for lexical semantics. In N. Calzolari et al. (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)* (pp. 969–974). European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). New York, NY: Springer Science & Business Media. doi: 10.1007/b97636

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*, 711–733. doi: 10.1353/lan.2006.0186

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511750526

Bybee, J., & Thompson, S. (1997). Three frequency effects in syntax. In M. L. Juge & J. L. Moxley (Eds.), *Proceedings of the Twenty-Third Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Pragmatics and Grammatical Structure* (pp. 378–388). Berkeley, CA: Berkeley Linguistic Society. doi: 10.3765/bls.v23i1.1293

Clahsen, H. (2007). Psycholinguistic perspectives on grammatical representations. In S. Featherstone & W. Sternefeld (Eds.), *Roots: Linguistics in Search of Its Evidential Base* (pp. 97–132). Berlin: Mouton de Gryuter.

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198299554.001.0001

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, *25*, 108–127. doi: 10.1016/j.newideapsych.2007.02.002

Divjak, D. (2008). On (in)frequency and (un)acceptability. In B. Lewandowska-Tomaszczyk (Ed.), *Corpus Linguistics, Computer Tools and Applications – State of the Art* (pp. 213–233). Frankfurt: Peter Lang. doi: 10.1515/9783110274059

Divjak, D., & Caldwell-Harris, C. L. (2015). Frequency and entrenchment. In E. Dąbrowska & D. Divjak (Eds.), *Handbook of Cognitive Linguistics* (pp. 53–75). Berlin: Walter de Gruyter. doi: 10.1515/9783110292022

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, *2*, 219–253. doi: 10.1075/lab.2.3.01dab

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*, 547–619. doi: 10.1353/lan.1991.0021

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*, 143–188. doi: 10.1017/s0272263102002024

Ellis, N. C. (2012). What can we count in language, and what counts in language acquisition, cognition, and use. In S. T. Gries & D. Divjak (Eds.), *Frequency Effects in Language Learning and Processing* (pp. 7–34). Berlin: Walter de Gruyter. doi: 10.1515/9783110274059

Ellis, N. C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 187–221. doi: 10.1075/arcl.7.08ell

Ellis, N. C., & Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics—Introduction to the special issue. *Applied Linguistics*, *27*, 558–589. doi: 10.1093/applin/aml028

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014a). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics*, *25*, 55–98. doi: 10.1515/cog-2013-0031

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2014b). Second language verb-argument constructions are sensitive to form, function, frequency, contingency, and prototypicality. *Linguistic Approaches to Bilingualism*, *4*, 405–431. doi: 10.1075/lab.4.4.01ell

Ellis, R. (2004). Individual differences in second language learning. In A. Davies & C. Elder (Eds.), *The Handbook of Applied Linguistics* (pp. 525–551). Malden, MA: Blackwell Publishing. doi: 10.1002/9780470757000.ch21

Ervin-Tripp, S. M. (1974). Is second language like the first. *TESOL Quarterly*, *8*, 111–127. doi:

10.2307/3585535

Evert, S. (2005). *The Statistics of Word Cooccurrences: Word Pairs and Collocations.* (Unpublished doctoral dissertation). Retrieved from http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf

Geeraerts, D., Grondelaers, S., & Bakema, P. (1994). *The Structure of Lexical Variation: Meaning, Naming, and Context*. Berlin: Mouton de Gruyter. doi: 10.1515/9783110873061

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67. doi: 10.1037/0033-295x.91.1.1

Gilquin, G. (2006). The place of prototypicality in corpus linguistics: Causation in the hot seat. In S. T. Gries & A. Stefanowitsch (Eds.), *Corpora in Cognitive Linguistics: Corpus-Based Approaches to Syntax and Lexis* (pp. 159–192). Berlin: Mouton de Gruyter. doi: 10.1515/9783110197709.159

Gilquin, G. (2010). Language production: A window to the mind? In H. Götzsche (Ed.), *Memory, Mind and Language* (pp. 89–102). Newcastle upon Tyne: Cambridge Scholars Publishing.

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

Goldberg, A. E. (1998). Patterns of experience in patterns of language. In M. Tomasello (Ed.), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure* (pp. 203–219). Mahwah, NJ: Lawrence Erlbaum.

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199268511.001.0001

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*, 289–316. doi: 10.1515/cogl.2004.011

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2005). The role of prediction in construction-learning. *Journal of Child Language*, *32*, 407–426. doi: 10.1017/s0305000904006798

Gor, K., & Long, M. H. (2009). Input and second language processing. In W. Ritchie & T. Bhatia (Eds.), *The New Handbook of Second Language Acquisition* (pp. 445–472). Bingley: Emerald.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*, 67–81. doi: 10.1017/s1366728998000133

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, *97*, 773–789. doi: 10.1093/biomet/asq042

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics*, *18*, 137–166. doi: 10.1075/ijcl.18.1.09gri

Gries, S. T. (2015). More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics*, *26*, 505–536. doi: 10.1515/cog-2014-0092

Gries, S. T., & Ellis, N. C. (2015). Statistical measures for usage-based linguistics. *Language Learning*, *65*, 228–255. doi: 10.1111/lang.12119

Gries, S. T., Hampe, B., & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*, 635–676. doi: 10.1515/cogl.2005.16.4.635

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software*, *17*, 1–27. doi: 10.18637/jss.v017.i01

Grosjean, F. (2010). *Bilingual: Life and Reality*. Cambridge, MA: Harvard University Press. doi:

10.4159/9780674056459

Haegeman, L. (1994). *Introduction to Government and Binding Theory* (2nd ed.). Oxford: Blackwell.

Hahn, U., & Ramscar, M. J. A. (2001). Conclusion: Mere similarity? In M. J. A. Ramscar & U. Hahn (Eds.), *Similarity and Categorization* (pp. 257–272). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198506287.003.0013

Hockley, W. E., & Cristi, C. (1996). Tests of the separate retrieval of item and associative information using a frequency-judgment task. *Memory & Cognition*, *24*, 796–811. doi: 10.3758/bf03201103

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, *73*, 418–433. doi: 10.1111/1467-8624.00415

Hudson, R. (1995). Does English really have case? *Journal of Linguistics*, *31*, 375–392. doi: 10.1017/s0022226700015644

Johnson, P. C. (2014). Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution*, *5*, 944–946. doi: 10.1111/2041-210x.12225

Kelly, M. H., Bock, J. K., & Keil, F. C. (1986). Prototypicality in a linguistic context: Effects on sentence structure. *Journal of Memory and Language*, *25*, 59–74. doi: 10.1016/0749-596x(86)90021-5

Kemmer, S., & Barlow, M. (2000). Introduction: A usage-based conception of language. In S. Kemmer & M. Barlow (Eds.), *Usage-Based Models of Language* (pp. 7–28). Stanford, CA: CSLI Publications.

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, *128*, 416–430. doi: 10.1016/j.actpsy.2008.02.001

Küchenhoff, H., & Schmid, H.-J. (2015). Reply to "More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff" by Stefan Th. Gries. *Cognitive Linguistics*, *26*, 537–547. doi: 10.1515/cog-2015-0053

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical Prerequisites* (Vol. 1). Stanford, CA: Stanford University Press.

Lieven, E. V. M. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, *120*, 2546–2556. doi: 0.1016/j.lingua.2010.06.005

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Glenview, IL: Scott Foresman.

MacWhinney, B. (2012). The logic of the unified model. In S. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 211–227). London: Routledge. doi: 10.4324/9780203808184.ch13

Madan, C. R., Glaholt, M. G., & Caplan, J. B. (2010). The influence of item properties on association-memory. *Journal of Memory and Language*, *63*, 46–63. doi: 10.1016/j.jml.2010.03.001

Maki, W. S., & Buchanan, E. (2008). Latent structure in measures of associative, semantic, and thematic knowledge. *Psychonomic Bulletin & Review*, *15*, 598–603. doi: 10.3758/pbr.15.3.598

Marcus, M., Kim, G., Marcinkiewicz, M. A., Macintyre, R., Bies, A., Ferguson, M., … Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In C. J. Weinstein (Ed.), *Proceedings of the 1994 ARPA Human Language Technology Workshop* (pp.

114–119). San Francisco, CA: Morgan Kaufmann. doi: 10.3115/1075812.1075835

Matusevych, Y., Alishahi, A., & Backus, A. (2015a). Distributional determinants of learning argument structure constructions in first and second language. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1547–1552). Austin, TX: Cognitive Science Society.

Matusevych, Y., Alishahi, A., & Backus, A. M. (2015b). The impact of first and second language exposure on learning second language constructions. *Bilingualism: Language and Cognition*. Advance online publication. doi: 10.1017/S1366728915000607

McRae, K., Ferretti, T. R., & Amyote, L. (1997). Thematic roles as verb-specific concepts. *Language and Cognitive Processes*, *12*, 137–176. doi: 10.1080/016909697386835

Mervis, C. B., Catlin, J., & Rosch, E. (1976). Relationships among goodness-of-example, category norms, and word frequency. *Bulletin of the Psychonomic Society*, *7*, 283–284. doi: 10.3758/bf03337190

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41. doi: 10.1145/219717.219748

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*, 302–331. doi: 10.1111/j.1467-9922.2010.00626.x

Naigles, L. R., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*, 95–120. doi: 10.1017/s0305000997003358

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. Anderson (Ed.), *Cognitive Skills and Their Acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum.

Ninio, A. (1999a). Model learning in syntactic development: Intransitive verbs. *International Journal of Bilingualism*, *3*, 111–130. doi: 10.1177/13670069990030020301

Ninio, A. (1999b). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language*, *26*, 619–653. doi: 10.1017/s0305000999003931

Onishi, K. H., Murphy, G. L., & Bock, K. (2008). Prototypicality in sentence production. *Cognitive Psychology*, *56*, 103–141. doi: 10.1016/j.cogpsych.2007.04.001

Palmer, M. (2009). SemLink: Linking PropBank, VerbNet and FrameNet. In A. Rumshisky & N. Calzolari (Eds.), *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon* (pp. 9–15). Stroudsburg, PA: Association for Computational Linguistics.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*, 71–106. doi: 10.1162/0891201053630264

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, *44*, 137–158. doi: 10.1007/s10579-009-9101-4

Pinker, S. (2013). *Learnability and Cognition: The Acquisition of Argument Structure* (New ed.). Cambridge, MA: MIT Press.

Plant, C., Webster, J., & Whitworth, A. (2011). Category norm data and relationships with lexical frequency and typicality within verb semantic categories. *Behavior Research Methods*, *43*, 424–440. doi: 10.3758/s13428-010-0051-y

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2014). Second language learner knowledge of verb–argument constructions: Effects of language transfer and typology. *The Modern Language*

Journal, *98*, 952–975. doi: 10.1111/modl.12149

Römer, U., O'Donnell, M. B., & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions. In N. Groom, M. Charles, & S. John (Eds.), *Corpora, Grammar and Discourse: In Honour of Susan Hunston* (pp. 43–71). Amsterdam: John Benjamins. doi: 10.1075/scl.73

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605. doi: 10.1016/0010-0285(75)90024-9

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., & Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice.* Retrieved from https://framenet2.icsi.berkeley.edu/docs/r1.5/book.pdf

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system. In D. Glynn & K. Fischer (Eds.), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches* (pp. 101–133). Berlin: Walter de Gruyter. doi: 10.1515/9783110226423.101

Schmid, H.-J. (in press). Introduction: A framework for understanding linguistic entrenchment and its psychological foundations in memory and automatization. In H.-J. Schmid (Ed.), *Entrenchment, Memory and Automaticity: The Psychology of Linguistic Knowledge and Language Learning.*

Schmid, H.-J., & Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics*, *24*. doi: 10.1515/cog-2013-0018

Schuler, K. K. (2006). *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon.* (Unpublished doctoral dissertation). Retrieved from http://verbs.colorado.edu/~kipper/Papers/dissertation.pdf

Shaoul, C., Baayen, R. H., & Westbury, C. F. (2014). N-gram probability effects in a cloze task. *The Mental Lexicon*, *9*, 437–472. doi: 10.1075/ml.9.3.04sha

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*, 246–251. doi: 10.1016/s1364-6613(03)00109-8

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*, 209–243. doi: 10.1037/0033-295x.89.1.1

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2004). Semantic generality, input frequency and the acquisition of syntax. *Journal of Child Language*, *31*, 61–99. doi: 10.1017/s0305000903005956

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, *4*, 253–290. doi: 10.1515/cllt.2008.011

Appendix. Estimating conditional probabilities

Conditional probabilities are estimated differently, depending on the type of the feature. A smoothed maximum likelihood estimator is used for features with values represented as a single symbol, such as head predicate, number of arguments, lexical arguments, cases and syntactic pattern:

$$P\left(F_k^I \middle| S\right) = \frac{\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right| + \lambda}{\left|F_k^S\right| + \lambda |F_k|} \tag{9}$$

where $\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right|$ shows how many times F_k^I occurs in F_k^S, and the smoothing parameter λ determines the default probability of F_k^I in S when $\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right| = 0$. The lower bound of λ (when a new ASC is created for each encountered instance) can be computed based on how many values each feature F_k in the data set can take. More specifically, $\lambda_{min} = \prod_k \frac{1}{F_k}$. For the data sets in the present study (when they are used jointly), $\lambda_{min} = 10^{-17}$. We chose a moderate value 10^{-9}.

Equation 9 can not be used for features with set values, because there is rarely a full overlap between any two sets of properties (e.g., semantic properties). In other words, $\left|\left\{F_k^I \middle| F_k^I \in F_k^S\right\}\right|$ is almost always 0. Alishahi and Pyykkönen (2011) propose the following way to compute the probability for such features, which we employ in this study:

$$P\left(F_k^I \middle| S\right) = \left(\prod_{e \in F_k^I} P(e|S) \times \prod_{e \in F_k \setminus F_k^I} P(\neg e|S)\right)^{\frac{1}{|F_k|}} \tag{10}$$

where F_k denotes the set of all values of this feature in the data, and $F_k \setminus F_k^I$ subtracts from this set all elements occurring in F_k^I. The probabilities $P(e|S)$ and $P(\neg e|S)$ can be computed using equation 9, replacing F_k^I with an individual element e.