Language learning as uncertainty reduction:

The role of prediction error in linguistic generalization and item-learning

Maša Vujović¹, Michael Ramscar², & Elizabeth Wonnacott³

¹ *Division of Psychology and Language Sciences, University College London, UK*

² *Department of Linguistics, University of Tübingen, Germany*

³ *Department of Education and St Johns College, University of Oxford, UK*

Correspondence concerning this article should be addressed to Elizabeth Wonnacott, Department of Education and St Johns College, University of Oxford, 15 Norham Gardens, Oxford, OX2 6PY, United Kingdom

Email: elizabeth.wonnacott@education.ac.uk

Declarations of interest: None

Abstract

Discriminative theories frame language learning as a process of reducing uncertainty about the meaning of an utterance by discriminating informative from uninformative cues via the mechanisms of prediction error and cue competition. Previous work showed that discriminative learning is affected by the order in which information is presented during language learning. Specifically, learning suffixes, where complex stems *precede* affixes, promotes better generalization than prefixing, which tends to promote better item-learning instead. We explored this in two large-scale web-based artificial language learning experiments with adult learners (total N = 434), as well as two computational simulations implementing a discriminative learning model. While we did not find an overall benefit of suffixing over prefixing in generalization, consistent with our theoretical and computational predictions, we found that participants in the prefix condition were unable to discriminate between frequent, but uninformative cues and low-frequency, informative cues. This resulted in them being more likely to show incorrect overgeneralization of that feature for low frequency test items than participants in the suffix condition. We did not find a benefit of prefixing in item learning (although there was overall better item-learning of low type-frequency items), which we discuss in terms of the methodological limitations of our empirical paradigm. Taken together, these results underline the crucial role prediction error plays in learning linguistic generalization, and have implications for how generalization interacts with item-learning.

Key words: Language Learning; Discriminative Learning; Prediction Error; Generalization; Item Learning

Introduction

Learning a productive language system has traditionally been described as learning a system of abstract symbolic rules which operate across multiple discrete linguistic items, and which is “unlearnable” unless at least some aspects of the system are innate (e.g., Chomsky, 2000). An alternative proposal suggests that productive systems emerge as a consequence of learning from the input via domain-general learning mechanisms, specifically, the ability to extract probabilistic structure from (linguistic) input.

Support for this alternative account comes from artificial language learning experiments in which learners are exposed to novel linguistic input, and are tested to see whether the input statistics affects the learning of linguistic structures. For example, in a now seminal study, Saffran et al. (1996) showed that infants and adults can identify “words” in an unsegmented stream of artificial speech by tracking the transitional probabilities between syllables. At the level of syntax, various studies reported effects of input statistics on generalization to novel instances. For example, Wonnacott et al. (2012) found that 5-year-olds were more likely to generalize verb argument constructions to novel verbs if those argument structures occurred with more verbs in the training input. Tamminen et al. (2015) similarly found that adults showed greater implicit generalization of a new morpheme to novel stems when the morpheme occurred with a larger number of distinct stems in the input, as well as when the semantics of the morpheme was more consistent across stems. Other factors have also been found to affect generalization in artificial languages, for example, skew in the input distribution (Casenhiser & Goldberg, 2005; Wonnacott et al., 2017) and the extent to which individual items occur in overlapping linguistic contexts (Reeder et al., 2013).

Further support for a statistical learning approach comes from a range of computational modelling methods, which have demonstrated that various aspects of linguistic behaviour can be

extracted from the input via statistically-based procedures (e.g., Elman, 1990; Finch & Chater, 1992; Mintz, et al. 2002; Perfors et al. 2010; Rumelhart & McClelland, 1986; Xu & Tenenbaum, 2007). Particularly influential are connectionist models, which represent human linguistic knowledge as a network of processing units interconnected to varying degree, and where learning in the network is typically error-driven, occurring via incremental changes in the strength of the connections between units that occur in response to the accuracy of their contributions to predictions about the environment. Critically, connectionist models lack in-built symbolic rules, meaning that any rule-like behaviour in the model is an emergent property of learning from input statistics (e.g., Elman, 1990; Rumelhart & McClelland, 1986). An apparent attraction of connectionist models is that to some extent their behaviour may mimic that of the networks of units (i.e., neurons) in the brain, giving them psychological plausibility, something that has been a stumbling block for other statistical learning approaches, such as those which model language learning as Bayesian inference (see Bowers & Davis, 2012; Jones & Love, 2011; Marcus & Davis, 2013). However, although connectionist work proposed gradual error-driven adjustment of the strengths of connections between processing units as the mechanistic model of human learning (e.g., Rogers & McClelland, 2004), the theoretical status of this assertion is unclear. For example, while the classic Rumelhart and McClelland (1986) model of English past tense learning uses a discriminating error-driven learning algorithm, the actual problem of learning inflectional morphology is framed in generative terms, such that the model learns to transform root forms into past tense forms based on the properties of the distribution of root forms encountered in the input. By contrast, the current work explores the role of error-driven learning in the context of a particular theoretical framework, *discriminative learning* (as introduced in Ramscar & Yarlett, 2007; Ramscar et al., 2010), which proposes that since

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

classical learning theory supports error-driven models, and since these models embody discriminative learning algorithms (Ng & Jordan, 2002), questions about human communication might be best framed in discriminative terms (such that the problem of learning inflectional morphology is framed as that of learning to discriminate the *contexts* in which root versus inflected forms should be produced/expected, such that the production of inflected forms is based on the distribution of the inflected forms encountered in context in the input). This approach has been applied to various aspects of language learning, including word learning (e.g., Ramscar et al. 2010; Ramscar, et al. 2013a), grammatical gender (Arnon & Ramscar, 2012), inflectional and derivational morphology (e.g., Ramscar, 2013; Ramscar & Yarlett, 2007; Ramscar & Dye, 2011; Ramscar, et al. 2013b), and verb argument structure (Ambridge et al., 2020), as well as processing effects in reading and naming (e.g., Baayen, et al., 2011; Hendrix et al., 2019; Milin et al., 2017). The key appeal of this framework is that it is firmly grounded in the principles of human and animal learning (next section) and yields accounts of the processes and outcomes of language learning that can be directly related to these principles. Importantly, because of this, it makes precise – and testable – predictions about when learning about linguistic inputs should and should not occur. In particular, because it explicitly treats learning as a predictive, discriminative process, it also makes predictions about the effect of the order in which events occur over the course of learning on subsequent generalization. In the current work, we test this prediction using the difference between suffixing and prefixing – that is, where affixes differ in their order with respect to the stem – as a test-case in a series of artificial language learning experiments with adult learners.

In the remainder of this introduction, we introduce the discriminative learning approach to language, and the closely related information-theoretic approach, and then consider

predictions about the effects of the order of information on learning and the evidence for these. Lastly, we turn to why the learning of suffixing versus prefixing makes an interesting test-case, and state our specific predictions.

Language Learning as Uncertainty Reduction

Learning theory defines learning as the process of reducing uncertainty about outcomes from environmental cues (Rescorla, 1988). Critically, and in contrast with Hebbian approaches, *contiguity* or simple cue-outcome co-occurrences (“Cells that fire together, wire together”, Hebb (1949, p. 15)) is neither sufficient nor necessary for learning (Kamin, 1968; Rescorla, 1967, 1968; Wagner, 1969). A key insight is that, the strength of association between a cue and an outcome depends not on contiguity but on the *predictive value* (amount of uncertainty reduced) of that cue. Given that uncertainty is finite, cues (which in classical learning theory, are thought to represent basic ‘elements’ that become associated in learning; Wagner, 2008) compete for predictive value: when a cue successfully predicts an outcome, the predictive value of that cue increases at the expense of competing cues; conversely, when the predicted outcome does not occur (i.e., when there is prediction error), the cue loses predictive value to other competing cues (Rescorla & Wagner, 1972). This process of *cue competition* is the learning mechanism which reduces uncertainty about the environment by discriminating uninformative from informative cues and, critically, it relies on the “negative evidence” provided by prediction error.

The discriminative learning framework proposed by Ramscar and colleagues adopts a discriminative as opposed to elemental approach to the representation of cues, in that representations are not assumed to be composed of atomic elements, but are themselves discriminated from an abstract representational space by the error-driven learning process itself (Ramscar, 2010; Ramscar & Port, 2015; 2016). This assumption in turn emphasizes the important discriminatory role that negative evidence (leading to negative associations between

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

cues and outcomes) plays in the learning process (Ramscar et al., 2010; Ramscar et al., 2013b).

In the context of language learning, “outcomes” are typically a system of linguistic forms, “cues” come from the world and earlier parts of an utterance, and “learning” involves discriminating informative cues from the large sets of uninformative cues available for any particular outcome.

Critically, this framing makes predictions about how the ordering of linguistic and other cues affects learning. Consider the case of learning the relationship between words and meanings.

This can logically occur in two directions: (i) predicting features in the environment from verbal labels or (ii) predicting verbal labels from features in the environment. Within the discriminative learning framework, qualitatively different learning outcomes are predicted to occur under these scenarios because only (ii) allows cue competition across the environmental features, and thus learning via negative evidence. For instance, imagine learning the meaning of the label *dog*. If the learner first sees a dog, a large number of cues (visual appearance: for example, *brown*, *furry*, *has a snout*; behaviours: for example, *barks*, *wags tail*, etc.) compete for predictive value for the label *dog*. Through this process, uninformative cues are gradually *unlearned* (i.e., they lose predictive value): whenever the learner encounters a *furry* animal that is *not* a *dog*, they are provided with negative evidence which down-weights the cue *furry*, and, via cue competition, leads to an increase in predictive value of other, more informative cues (and concomitant reductions in value of uninformative cues). In contrast, if the learner first hears the label *dog* and then predicts its meaning, there is no cue competition over the features in the environment. In the absence of cue competition, the predictive value of the label simply increases when the predicted features occur, and decreases when they do not occur, approximating the conditional probability of the features given the label.

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

An artificial language learning experiment by Ramscar et al. (2010) provided empirical support for this theory. Adult learners were taught category labels for novel items. In one condition, participants first saw a novel object and then were presented with the written category label (“That was a *wug/dep*”), and in the other, participants first heard the category label (“This is a *wug/dep*”) and then saw the novel object. Critically, some *wugs* had the same body shape as the majority of *deps* – therefore, appropriate learning of the two labels involved dissociating body shape as an uninformative cue (despite its saliency and frequency), and reinforcing a more complex informative set of cues that make something a *wug* or a *dep*. Ramscar et al. (2010) predicted that body shape will be dissociated from the category labels only in the conditions where the objects were viewed before the label was heard: only in this condition could participants have the opportunity to view a *wug*, and expect to see the label *dep* (because the exemplar had same body shape as most of the *deps*), and then, upon reading *wug* instead, to dissociate body-shape and assign greater predictive value to other, more informative cues. If the label is heard before the object is viewed, uninformative cues cannot be “unlearned” as there is no negative evidence: every time the label *wug* is encountered, it is followed by an actual exemplar of the category *wug*, that is, by a visual image containing only the features of *wugs*. The prediction was supported by the results: following exposure to the language, only participants in the object-first condition were above chance at selecting the appropriate label for novel items with the body-shape which was frequently associated with the alternative label (i.e., “minority” *wugs*), demonstrating that cue competition and prediction error are the driving mechanisms of learning.

The key notions of learning theory – and the discriminative learning framework – are closely related to information theory (Shannon, 1948), whereby communication occurs in

inherently noisy channels, and the goal of optimal communication is to reduce the uncertainty about the message, with uncertainty measured mathematically as *entropy* (section Entropy Calculations). In a re-analysis of learning theory in information-theoretic terms, Gallistel (2002) reformulates the role of predictive value in learning (Rescorla, 1967, 1968) as follows: what determines learning is not a frequent individual pairing of events, but how information is structured in a flow of events (closely related to the notion of order effects in learning). While numerous studies have applied information-theoretic concepts to various domains of language learning, processing, and use (e.g., Aylett & Turk, 2004; Boston, et al., 2011; Frank, et al., 2015; Frank, 2013; Gibson et al., 2013; Hale, 2003; Jaeger, 2010; Levy, 2008; Maurits et al., 2010; Piantadosi et al., 2011; Pluymaekers et al., 2005; Van Son & Van Santen, 2005), few have made the explicit link between entropy and predictive value. One example is the work of Dye et al. (2017), who conducted a corpus study of gendered-article+noun pairs in German. Dye et al. showed that individual nouns have higher entropy than individual items from any other part of speech, and argued that, during language processing, encountering a gendered article reduces the entropy of the noun, which in turn makes the noun easier to process (and makes noun classes more discriminable on the whole). For example, in German, hearing the masculine article *der* reduces the set of possible upcoming nouns to only those nouns that co-occur with that article¹. Eye-tracking experiments show that this entropy-reducing property of gendered articles is readily exploited during language processing by both adults (Dahan et al., 2000; Lew-Williams &

¹ In a similar analysis, Dye et al. (2018) showed that languages without gendered articles, or without any gender agreement at all, can use other means to smooth entropy over the utterance. For instance, in English, the sequence of prenominal adjectives *cute little* frequently precedes the nouns *baby*, *puppy*, or *kitten*. While these adjectives do not add much to the meaning of the noun (most babies, puppies, and kittens are *cute* and *little*), they reduce the set of probable upcoming nouns, and thus the information required to process the nouns.

Fernald, 2010) and children (Lew-Williams & Fernald, 2007), who, as soon as they hear the gendered article, direct their gaze towards an item of that gender.

Critical to the current study, there is evidence that the processing benefit provided by a pre-nominal determiner may also help learning the meaning of novel nouns. Arnon and Ramscar (2012) taught adult English speakers an artificial language in which half of the nouns were preceded by one article and the other half by another and compared learning in contexts where participants either (i) were first exposed to the nouns without their articles, and then to nouns with their articles versus (ii) were first exposed to nouns with their articles, followed by nouns without their articles. Their key interest was in the learning of article-noun pairings, which was found to be better in (ii), where the bigrams were presented before the nouns in isolation². However, most relevant to the current study is that in addition to better learning of noun-article pairings, Arnon and Ramscar (2012) report higher learning of *noun-object* mappings in that condition, which they interpret as showing that in the early exposure to the article-noun mappings in condition (ii), the articles reduced the entropy of the nouns, thus making the nouns easier to process, which in turn made it easier to learn the “meaning” of the noun (its corresponding picture).

To summarize, experimental, computational, and corpus-based evidence suggests that the order in which information is presented during learning and processing enables qualitatively

² This finding illustrates a classical learning effect, whereby learning only occurs when the cue reduces the uncertainty about the outcome (i.e., the blocking effect, Kamin, 1968; Ramscar et al., 2013a). Specifically, in (i), encountering the noun-article bigrams after having been exposed to the nouns in isolation did not add any new information about the picture; the noun itself was fully predictive of the picture and therefore the overall learning of noun-article pairings was weaker in this condition compared to (ii).

different learning. In the next section, we consider the role of ordering with respect to differences between *suffixes* and *prefixes* in inflectional morphology.

Discrimination Learning and Effects of Order on Learning: Inflectional Morphology as a Test-case

Across languages of the world, strongly suffixing languages (e.g., English, Turkish, Polish, Japanese, Tamil, Somali, Piraha) are much more common than strongly prefixing languages (e.g., Navajo, Chipewyan, Zulu) (Dryer, 2013; Greenberg, 1963). There is also evidence that children learning a suffixing language tend to learn inflectional morphology faster than children learning a prefixing language (Clark, 2007). This has led to hypotheses that there may be a cognitive “bias” which underpins this crosslinguistic tendency. One approach has been to look for evidence for a universal domain-general preference towards placing important information at the beginning of the sequence (i.e., the stem in suffixing). Hupp et al. (2009) presented English-speaking participants with novel labels and referents, such as hearing “*this is a tate*” whilst viewing a heart-shaped object, and then asked them to extend the label to a novel affix – viewing a heart-shaped and a star-shaped object and being asked “which one is a *tate-be* / *be-tate*?”. If the label was suffixed (i.e., *tate-be*?), participants were more likely to choose the heart-shaped object than if the label was prefixed (i.e., *be-tate*?), suggesting that the suffixed label was more likely to be interpreted as part of the same category of words denoting heart-shaped objects. Although the authors attributed this to a universal bias, a recent study by Martin and Culbertson (2020) found the *reverse* preference with speakers of a strongly prefixing language (Kiitharaka), suggesting that these experiments capture a bias reflecting linguistic experience.

An alternative method for exploring the difference between suffixing and prefixing is to contrast the learning of these in an artificial language taught *in the experiment*. Critically, these

studies – in common with the current work – compare the learning of matched languages where affixes either precede (prefixing) or follow (suffixing) stems. A key assumption is that stems have features (phonological or semantic) which are associated with a smaller set of simpler affixes, and of interest is how well these features are discriminated under the two affixing conditions. St Clair et al. (2009) taught adult native speakers of English two noun classes characterized by phonological features (class A: onset and offset consonant clusters, unrounded high vowels and fricatives; class B: no consonant clusters, rounded low vowels; nasals and stops) and associated with suffixes (suffixing condition) or prefixes (prefixing condition). During training, participants heard noun+affix pairs. Participants were tested with novel pairs which were either compatible (categoryA_noun + categoryA_affix) or incompatible (categoryA_noun + categoryB_affix) with training pairs, and were asked to indicate whether pairs were “similar” or “dissimilar” to the training language. While both conditions performed above chance, participants in the suffix condition were significantly more accurate. The authors interpret this suffix benefit in terms of the effect of order in uncertainty reduction: a suffixing benefit occurs because cue competition between (here phonological) features of the lexical items can only occur when lexical items *precede* affixes (i.e. suffixing). It is this cue competition which promotes the learning of the *discriminating features* – that is, the common dimensions which group nouns with the same affixes.

In later work, Ramscar (2013) demonstrated that it is not that suffix learning is generally “better” than prefix learning; while suffixing promotes the abstraction and thus generalization of the discriminating features which group items with the same suffix, prefixing promotes better item-learning. This occurs in the same way as with preceding articles (discussed above): just as hearing “der” reduces the entropy of a following noun (Dye et al., 2017; 2018), entropy can be

reduced by a prefix: the entropy of *advantage* is higher in *He was advantaged* than in: *He was disadvantaged*. As discussed, this processing advantage may make the stem more learnable, resulting in better learning of trained stems under prefixing. Ramscar (2013) tested these benefits of suffixing and prefixing in an artificial learning experiment where adult learners were trained on prefix + noun + suffix strings denoting pictures of objects (e.g., *bahlot* – fork; *pikroo* – car). Nouns either occurred with prefixes consistently while the suffix varied (e.g., *oos-bahlot-lek*; *oos-bahlot-bol*), or occurred with suffixes consistently while the prefix varied (e.g., *mal-pikroo-sem*; *oos-pikroo-sem*). When tested on recall of trained items, participants were more accurate at recalling the affix-noun and noun-picture mappings when the nouns occurred with consistent prefixes compared to consistent suffixes in the training (further corroborating the result reported in Arnon and Ramscar (2012), where article-noun and noun-picture learning was better in the condition where article-noun pairs were taught first). However, items occurring with the same suffix were judged more similar to each other than items occurring with the same prefix because predicting suffixes from complex items (noun-picture pairs) involved cue competition, which meant that over time, learners could discriminate consistent or predictive cues (i.e., dimensions that are shared across items occurring with the same suffix) from unpredictable cues.

The studies reviewed above suggest that prefixing and suffixing facilitate qualitatively different learning, with prefixing leading to better learning at individual item-level (Ramscar, 2013), and suffixing facilitating the learning of discriminating features – semantic and/or phonological features which group items together (St. Clair et al., 2009; Ramscar, 2013). Importantly, these differences can be parsimoniously explained by the mechanisms of uncertainty reduction during learning and processing. Our focus is on unifying the two sets of predictions as an *in-principle* demonstration of uncertainty reduction as a driving mechanism in

language learning, and specifically its sensitivity to the order in which information is structured. Although this account has relevance for any linguistic structure where information is ordered (for example, differences between head-initial and head-final compounds, Marelli & Luzzatti, 2012 – we return to this in the Implications section in the General Discussion), we look at suffixing and prefixing as a useful test-case, since it provides us with both a generalizable aspect of the artificial language (affixes), *and* item-specific aspects (individual stems and associated visual items). This allows us to test the two predictions which bring together two closely theoretical approaches (learning theory and information theory) in a single paradigm, something that has not been done to date. While Ramscar (2013) was the only study to test item-learning and discriminating feature learning within the same paradigm, this study did not have a direct test of generalization. That is, there was no test with novel items containing (or not containing) a key discriminative feature, since the design of the study was such that individual objects were randomly assigned to affixes, with no consistent mapping along any dimension. Ramscar et al. (2010), on the other hand, did test generalization, but this involved category labels only and no item-level learning. In addition, the labels in that study were presented in written form, rather than auditorily. Therefore, here, and unlike previous work, we directly test both learning of individual items and generalization, and we do this with spoken stimuli, making it easier to draw parallels with spoken language, and more straightforward to incorporate both semantic and phonological cues within the design.

Current Study

The current study builds on previous research by comparing the learning of suffixing and prefixing using artificial language learning. Specifically, we test for benefits of suffixing on generalization (Ramscar et al., 2010; Ramscar, 2013) and for benefits of prefixing in the learning of individual items (Arnon & Ramscar, 2012; Dye et al., 2017, 2018; Ramscar, 2013).

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

A key hypothesis is that *generalization* occurs via discriminative learning through cue competition. We thus expect clearest benefits of suffixing where appropriate generalization involves dissociating frequent, salient, but non-predictive features, boosting the key discriminating features. This requires negative evidence, which is only provided when the noun precedes the affix. To explore learning in these circumstances, we borrowed the artificial languages from Ramscar et al. (2010) and adapted them for learning suffixes and prefixes. We used a language with CVC stems denoting novel objects and two CV affixes, which either preceded (prefix condition) or followed (suffix condition) the stems. Half of the stems occurred with one affix, and half with the other. Each stem was uniquely associated with a visual referent. We used the visual stimuli from Ramscar et al. (2010) called *fribbles*³, and manipulated the semantic features analogously to that study: there were two sub-classes of items associated with each of the two affixes, which we call “low type-frequency” (LTF) and “high type-frequency” (HTF). Critically, 75% of the items which occurred with one affix (HTF items) shared a visual feature (body shape) with the LTF items which occurred with the *opposite* affix (and which made up 25% of all items occurring with that affix). Thus we simulated a learning environment where appropriate generalization involves “unlearning” an uninformative yet frequent, salient feature (body shape). Unlike Ramscar et al. (2010), who used written label stimuli, but similar to St Clair et al. (2009), the present study used audio verbal stimuli, and these also included a phonological cue to category membership – stems of the same category contained the same vowel. This cue was included because semantic and phonological cues for grammatical generalization generally co-occur in natural languages (MacWhinney et al., 1989; Mirković et

³ Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, accessed here: https://wiki.cmbc.cmu.edu/Novel_Objects

al., 2005; Zubin & Kopcke, 1981). Unlike the visual stimuli (which we treat as “semantic” features, while acknowledging that in natural languages these semantic dimensions tend to be more abstract), phonological cues in our study were 100% consistent with the affixes, since we were concerned about floor effects in this relatively complex learning environment (i.e., in which participants are learning both individual nouns and the relationships between nouns and affixes, and in which there are multiple different cue types).

We used tests of generalization (Semantics-Phonology Generalization test; Phonology Generalization test) and item learning (Item Learning test). We derived three priori predictions from the underlying theory. First, for generalization (i.e., determining which affix was appropriate for a novel item on the basis of semantic/phonology cues) we predict that participants in the prefix condition would have more difficulty matching novel low type-frequency stems with the correct affix compared to high type-frequency stems, due to the presence of uninformative yet frequent, salient features associated with the alternative affix, whereas participants in the suffix condition should be unaffected by frequency-type (we expect to see this in the data as an interaction between affix and type-frequency in the Semantics Phonology Generalization test; Ramscar et al., 2010). Secondly, we predict overall better generalization in the suffix condition regardless of frequency-type (a main effect of affix in both generalization tests; Ramscar et al., 2010; Ramscar, 2013; St Clair et al., 2009). Thirdly, for item learning, we predict that participants in the prefix condition would show stronger performance compared to the suffix (Arnon & Ramscar, 2012; Ramscar, 2013), due to the reduction in entropy caused by hearing the prefix before hearing the stem and seeing the *fribble* during training. Finally, we included a test where participants were asked to recall the correct affix for trained stem-*fribble* pairs. Although we predict higher performance here compared with the

generalization items, no explicit prediction concerning difference between conditions was made since participants could use item-specific knowledge (i.e., learned associations between trained stem-*fribble* pairs and the affixes with which they co-occurred in the input) and/or the more general features of the stems/*fribbles* used in the generalization test. We explored the latter possibility with the same affix-by-type frequency interaction from the generalization test. Before presenting the experiment (section Experiment 1), we present computational simulations and entropy analyses using the same input sets as used with human learners which confirm the theoretical predictions laid out above.

Predictions for Human Learners: Computational Simulations and Entropy Analysis

Ramscar et al. (2010) based their predictions for label learning on a computational simulation implementing a simplified version of the Rescorla-Wagner rule⁴ (Rescorla & Wagner, 1972). The computational principles which predicted a benefit of presenting pictures before labels in that experiment should apply equally to seeing a benefit of presenting the stem and its semantics before presenting the affix in the current experiment (i.e., our predicted suffix benefit). Nevertheless, the input set in the current experiment differs from that in Ramscar et al. (2010); specifically, we use a different number of categories (four rather than six) and the “features” to be associated with the affixes come not only from the visual features of the fribbles (as in Ramscar 2010) but also the phonological features of the stems. The latter includes both idiosyncratic features of the individual labels and the vowels that are shared across stems of the

⁴ As pointed out by Stone (1987), when simplified (by the removal of an additional parameter that arbitrarily modifies the saliency of cues), the Rescorla-Wagner update rule is essentially the same as the Least Mean Square or the Widrow-Hoff rule (Widrow & Hoff, 1960), developed in the area of applied mathematics as an algorithm for solving sets of linear equations (see Sutton & Barto, 1981 for further discussion).

same category (i.e., the phonological features which provides a cue to affixing). Given these differences, it is worth confirming the prediction for a benefit of suffixing for generalization with this specific input set. In addition, in this work we also have a second hypothesis for a benefit of prefixing for item learning. Ideally, we would base our predictions for item learning within the same simulations as for generalization. However, testing this prediction involves testing the relationship between noun-labels and visual features which are presented simultaneously during training, and this is not straightforwardly possible using the architecture presented Ramscar et al. (2010) (which allows looking at relationships between temporally ordered *cues* and *outcomes* only). We present instead an information-theoretic analysis of the two conditions in order to establish the predictions for item-learning – we are able to do this since the item-learning hypothesis is rooted in information theory, which (as discussed) is closely aligned with learning theory, such that learning theory concepts can be reformulated in information-theoretic terms (see Gallistel, 2002 for a demonstration).

Learning Model Based on the Simplified Rescorla-Wagner Rule

We define a single discrete learning trial as a row vector of cues: $\varphi(S) = (\varphi_1(S), \varphi_2(S), \dots, \varphi_n(S))$, where $\varphi_i(S) = 1$ if the cue is present in the trial, and 0 otherwise. If an n -dimensional row vector of associative strengths (weights) is θ , the predicted response is given by:

$$V(S, \theta) = \theta \varphi(S)^T$$

The predicted response is compared with the outcome y ; if the outcome occurs, y is 1, otherwise it is 0. Prediction error is computed as follows (this corresponds to L2 loss):

$$L(S, \theta) = \frac{1}{2} (y - V(S, \theta))^2$$

The simplified Rescorla-Wanger (1972) learning rule for a vector of weights θ^{k+1} for a given input vector i and outcome j is given as follows:

$$\begin{aligned}\theta^{k+1} &= \theta^k - \beta_j \nabla_{\theta} L(S_i, \theta^k) \\ &= \theta^k + \beta_j (y_j - V(S_i, \theta^k)) \varphi(S_i)\end{aligned}$$

where β is learning rate. In this experiment, a default value of 0.01 was used for this hyperparameter following Ramscar et al. (2010).

The simulations were implemented as fully connected feedforward two-layer networks. Due to the lack of hidden layers, the output of the model directly corresponds to the weights of the connections between the input and the output, that is, the strength of association between individual cues and individual outcomes. As such, the model allows us to compare the strength of association between informative cues and the outcome on the one hand, and uninformative cues and that outcome on the other.

Simulation Design

Two models were each trained using input sets representing those used with human learners in the prefix and suffix conditions in Experiment 1 (source code at <https://doi.org/10.5281/zenodo.4459512>). The models are trained by presenting a set of *cues* which are used to predict a set of *outcomes*; in the prefix condition, there was a single cue (the affix) and the outcomes were phonological features of the stem-label and semantic/visual features of the referent; in the suffix condition, the affix was the outcome and the features were the cues. For simplicity, we used binary coding for all features. In particular, the discriminating phonological feature of the stem is the vowel, and this was coded as a discrete phoneme (*vowel_1*, *vowel_2*; acknowledging that this is not a realistic model of how humans perceive continuous speech). Similarly (and following Ramscar et al., 2010) the semantic features

relevant for generalization were coded as individual discrete features: non-discriminating but frequent features (e.g., *body_shape_1*), and discriminating features (e.g. *discriminating_features_1*, *discriminating_features_2*, etc.). Finally, for each item in training, we included two additional features intended to represent the fact that individual items have sets of idiosyncratic phonological and semantic features (e.g., *item1_idiosyncratic_phonology*, *item1_idiosyncratic_semantics*). A simplified representation of the input is given in Figure 1. The networks were trained on 7000 trials each.

Following Ramsar et al. (2010), generalization was assessed by activating the non-idiosyncratic features of each frequency-type (cues relevant for generalization; e.g., *shape1*, *discriminating_feature1*, *vowel1*), and summing the associative strengths of those features for each of the affixes. If the model can discriminate informative from uninformative cues, the informative cues will contribute to the sum the most, as the weights for the uninformative cues will be lower, and therefore the correct affix will be chosen. The sums of raw associative strengths were normalized using the choice axiom (Luce, 1959), whereby the probability of the correct affix i out of all affixes j for a randomly chosen exemplar (sum of weights w) is given as follows:

$$P(w_i) = \frac{w_i}{\sum_j w_j}$$

Results and Discussion

From Figure 2, which shows the associative strength between cues and a randomly chosen outcome over the course of learning, we can see that the relative associative strength of features in the suffix condition (right panel) corresponds to their relative differences in *predictive value* rather than in *frequency* of co-occurrence with the affix, which is the case in the prefix condition (left panel). Specifically, after a few hundred trials, in the suffix condition, *shape1* (a

high-frequency, but uninformative cue) loses predictive value to the HTF discriminating feature; similarly, *shape2* (a high-frequency feature which occurs with the *opposite* affix 75% of the time) loses predictive value to the benefit of the LTF-discriminating feature, which is a less frequent, but more predictive feature. In the prefix condition (left panel), on the other hand, this never happens – this condition remains unable to discriminate between uninformative features (*shape1* and *shape2*), and informative features (HTF and LTF, respectively).

How do the differences in the dynamics of weight updates between the conditions affect generalization, and particularly with LTF items, which share a frequent feature with HTF items of the *opposite* affix? To assess this, following Ramscar et al. (2010), we used the choice axiom to compute the normalized probability of choosing the correct affix for (i) a combination of HTF features (i.e., *shape1*, *discriminating_feature1*, *discriminating_vowel1*, *item1_idiosyncratic_phonology*, *item1_idiosyncratic_semantics*) and (ii) a combination of LTF features (i.e., *shape2*, *discriminating_feature2*, *discriminating_vowel1*, *item7_idiosyncratic_phonology*, *item7_idiosyncratic_semantics*). The results (Figure 3) showed that generalization in the prefix condition is affected by type frequency much more than in the suffix condition, and this is because in the prefix condition the high-frequency non-discriminating feature has the same associative strength as the LTF discriminating feature (i.e., the model is unable to discriminate based on predictive value). In the suffix condition, on the other hand, while HTF generalization is *slightly* better than LTF generalization, this difference is much smaller compared to the prefix condition, since the cue-outcome structure in this model facilitates cue competition, allowing the model to learn that the frequent non-discriminating feature has less predictive value than the low frequency discriminating feature. This finding demonstrates that cue competition is the mechanism underlying discrimination learning, and

forms the basis of our key prediction for human learners: at generalization, participants in the prefix condition will be more strongly affected by frequency-type than participants in the suffix condition, as they will not be able to discriminate between uninformative and informative cues due to a lack of cue competition in that condition.

Of note is that, although the raw associative strength for the HTF discriminating features (and the *vowel* discriminating feature) was higher in the prefix condition compared to the suffix condition (Figure 2), this did not result in better performance with that feature in generalization – in fact, generalization as assessed by the choice axiom was slightly better in the suffix condition (grey bars in Figure 3). This is because with this axiom, even though the weights of discriminating features may be high, as long as they are not higher than the weights of non-discriminating features (as was the case in the prefix model), this will not improve performance, and will in fact hinder it with low-frequency features. This result is also consistent with previous work which observed the same overall suffixing benefit (Ramscar, 2013; St Clair et al., 2009). Therefore, we predict overall better generalization in the suffix condition compared to the prefix condition (main effect of affix), though we note that metric which yields this result (the choice axiom), is not part of the Rescorla-Wagner learning rule itself and that, while chosen to reflect our theoretical assumptions (and following Ramscar et al., 2010), previous work in the discriminative learning framework has not systematically investigated whether this metric indeed best captures human performance compared to alternatives, making this prediction somewhat exploratory. We re-evaluate this choice of metric in the General Discussion.

Finally, since our tests with humans use phonological cues in isolation (Phonology Generalization test), we also tested the model by only activating the discriminating phonological features. Using the choice axiom, the probability of correct affix alone was 1 in the suffix

condition (reflecting the fact that this cue is fully predictive of affix choice), and 0.85 in the prefix condition.⁵ This again reflects the use of the choice axiom which means that although the absolute associative strength of the correct feature was greater in the prefix condition compared to the suffix, the difference between the strength of the *correct* discriminating feature and the feature of the *opposite* affix was greater in the suffix condition compared to the prefix condition. The result is also consistent with St Clair et al. (2009) who found a suffixing advantage with fully predictive features. Therefore, we also predict a suffixing advantage in our test of phonology generalization, but again, this hypothesis is less central to the theory, since it is specific to the choice axiom and not learning theory itself.

To summarize, the results of our simulations implementing the Rescorla-Wagner learning rule provide a mechanistic basis for our predictions regarding generalization. The key prediction is a greater type-frequency effect (HTF better than LTF) in the prefix condition compared to the suffix condition. We make two further predictions: overall better generalization in the suffix condition in the test including both semantic and phonological features, and in the test of phonological learning alone.

Entropy Calculations

We predict better item-learning in the prefix condition from previous work (Arnon & Ramscar, 2012; Ramscar, 2013), and our calculations of entropy distribution under different conditions given below.

We know that for a joint probability distribution of stems s and affixes a , $P(s, a)$, the marginal probability of a stem s is given as follows:

⁵ As the choice axiom can only be used with non-negative values, prior to normalization we scaled the weights to a 0 – 1 range.

$$P(s) = \sum_a P(s, a)$$

The conditional probability of a stem s given a is:

$$\begin{aligned} P(s|a) &= \frac{P(s, a)}{P(a)} \\ &= \frac{P(s, a)}{\sum_s P(s, a)} \end{aligned}$$

The uncertainty of a probability distribution is measured as *entropy* (expressed in bits), given as:

$$H(p) = - \sum_{i=1}^N p(x_i) \log p(x_i)$$

Therefore, the entropy of the probability distributions above can be computed using the entropy equation. Below we demonstrate this with a specific example.

In Experiment 1, participants are exposed to 16 stems and two affixes, where half of the stems occur with one affix, and the other half with the other. Formally, the joint probability distribution of stems s and affixes a for this experiment is given as:

$$P(s = i, a = j) = \begin{cases} \frac{1}{16} & \text{if } i \leq 8, j = 1 \\ 0 & \text{if } i \leq 8, j = 2 \\ 0 & \text{if } i > 8, j = 1 \\ \frac{1}{16} & \text{if } i > 8, j = 2 \end{cases}$$

Across the two conditions, we manipulate the order in which the stems and the affixes are presented to participants, and we are interested in probability of a stem s_i just before it is heard. In the suffix condition, the stems are played before the affixes. Therefore the probability of s_i just before it is heard corresponds to the marginal probability of s . In our experiment, this equals to:

$$P(s_i) = \sum_a P(s_i, a) = \frac{1}{16}$$

In the prefix condition, on the other hand, the affix is played just before the stem is played. Therefore, in this case, the probability of s_i just before the noun is heard is the conditional probability of s_i given affix j , which gives:

$$P(s_i|a_j) = \begin{cases} \frac{1}{8} & \text{if } j = 1 \\ 0 & \text{if } j = 2 \end{cases}$$

Therefore, in idealized conditions, where participants have learned the probability distribution perfectly, and where they generate expectations about the identity of the upcoming stem according to the distribution, in the suffix condition, the greatest probability of successfully predicting the stem is equal to chance, or 1 in 16, whereas in the prefix condition, this is twice as high – 1 in 8. Using the entropy equation defined above, we can calculate the entropy of the marginal probability of s , as well as the entropy of the conditional probability of s given a in the two conditions. In the suffix condition, the two entropies are the same: 4 bits. In the prefix condition, on the other hand, the entropy of the marginal probability of s is 4 bits, but the entropy of s given a is 3 bits. This shows that in the prefix condition, at the point at which the stem is heard, entropy is lower than in the suffix condition. As discussed in Language Learning as Uncertainty section, lower entropy should enable lower processing cost; and based on Arnon and Ramscar (2012) and Ramscar (2013), we hypothesise that this processing benefit will be reflected in stronger item learning (i.e., better learning of which stem occurred with which picture and with which affix) in the prefix condition than in the suffix condition.

Experiment 1

Method

Experiment Versions

For both Experiments 1 and 2, two versions of the paradigm were used: an “extended” version (Version A) and a “condensed” version (Version B), with roughly half of the participants in each version. This was necessary in order to fine-tune a relatively novel paradigm which differs from previous work in several important aspects. In each case, the key difference between the two versions of the paradigm was that we dropped two generalization tests from Version A, allowing us to have more items in the remaining test in Version B (other changes are outlined in Procedure). While statistics were computed for all tests and each version of each experiment separately as well as combined, for clarity of exposition, we only report results from those tests which are common to both paradigms and only statistics computed over the combined datasets (Version A + Version B). Our choice of inferential statistic – the Bayes factor – remains a valid measure of evidence regardless of how the data were collected (Dienes 2016), and the larger combined sample allows for the best measure of evidence. The other analyses are available online at: <https://osf.io/kd4br/>.⁶

The two Versions were designed in succession as opposed to in parallel (insights from Version A were used to inform the design on Version B). In both experiments, Version A was

⁶Analyses on data from the “B” versions alone (as per the pre-registration where we planned separate analyses of versions A and B) yielded the same set of qualitative inferences as to whether evidence was for H1/H0 or ambiguous as the analyses over combined data, with three exceptions where we were following up our main analyses with cell-by-cell comparisons to chance, and there was a cell with ambiguous evidence in the B experiment but substantial evidence for H1 in the combined dataset. For further transparency, we acknowledge that results of the additional two generalization tests which were only included in versions A (and therefore run with smaller samples) were inconsistent across the two experiments, in particular with regards to the key interaction between affix and type frequency: one test showed evidence for H1 in Experiment 1A, but evidence for the null in Experiment 2A, and the other test showed evidence for H1 in Experiment 2A, and ambiguous evidence in Experiment 1A. However, the evidence as to whether these were real differences between the experiments was ambiguous. We reported and discussed these results, and possible limitations of one of the tests, in an earlier draft of this paper submitted for review, and this discussion is now included in the online materials: <https://osf.io/kd4br/>.

ran in one batch and the data were analysed before Version B was designed and run as a slightly modified replication. The predictions based on Version A were pre-registered online (Experiment 1: <https://rpubs.com/MasaVujovic/discrimlearning1b>; Experiment 2: <https://rpubs.com/MasaVujovic/discrimlearning2b>). In both cases, participants were recruited to Version B using a pre-registered stopping rule procedure whereby data were collected in a first batch of 20 participants and then in additional batches of 10 participants, inspecting the data after each batch and stopping if there was substantial evidence for or against (i.e. $BF > 3$ or $< 1/3$) the key prediction – the interaction between affix and type frequency in Semantics and Phonology generalization tests, or if a maximum of $N = 160$ was reached (with the maximum based on resources available).⁷

Participants

For Version A, data from 84 participants were collected, with sample size based on obtaining a sample approximately three times the one in Ramsar et al. (2010). Data for Version B were collected using the pre-registered stopping rule described above; our final sample was $N = 120$. Across the two versions, 204 participants (102 per condition; prefix: $M_{age} = 35.65$, $SD_{age} = 11.13$, 45 female; suffix: $M_{age} = 35.18$, $SD_{age} = 10.76$; 55 female) were recruited through Prolific

⁷ Dienes (2016) explains that this type of stopping rule does not present the same problem for Bayes factors as it would for a frequentist analysis which relies on the interpretation of p values. In that case, where testing until a significance threshold is reached (optional stopping) is guaranteed to find a significant p -value even if H_0 is actually true, since the p -value will do a random walk and eventually dip below .05. In contrast, a Bayes factor is symmetric – in the long run, if H_0 is false then B will be driven upwards, but if H_0 is true B will be driven towards zero. Since Bayes factors remain a valid measure of evidence regardless of the data collection process employed, they therefore respect the “stopping rule principle” which states that the only evidence about a parameter should be contained in the data and not the stopping rule used to collect them.

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

Academic (<https://prolific.co/>) and participated on-line, with the experiment hosted on Gorilla Experiment Builder (Anwyl-Irvine et al., 2019). All participants were adult monolingual native speakers of English with no known language, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for participation.

Stimuli

Audio stimuli consisted of 32 nouns (16 training, 16 testing – with assignment to training/testing randomized across participant) and four affixes. Each noun was a CVC syllable, such that the vowel was /u:/ for one half of the nouns (Category A nouns) and /i:/ for the other half (Category B nouns), and the consonants were re-used across categories. Each of the two noun classes was accompanied by a different affix. One Category A affix and one Category B affix was chosen randomly on a participant-by-participant basis, from two Category A affixes (/pɛ/ and /gɛ/) and two Category B affixes (/dʌ/ and /mʌ/) (following St Clair et al. (2009), we used lax vowels, consistent with word-initial, rather than word-final vowels in English, thus favouring the prefix condition, so that any suffixing advantage reflects a robust effect rather than an artefact of the stimuli design). The audio stimuli were synthesized using the MBROLA speech synthesizer (Dutoit et al., 1996) with a male British English voice. For visual stimuli, in each category, there were six HTF items and two LTF items (that is, 75% versus 25%). An additional four HTF and four LTF items per category (16 in total) were reserved for testing. See Figure 4 for a sample training set of pictures and labels.

Procedure

Participants were told that they would “learn an alien language” and see pictures of aliens and hear “how they are referred to in the alien language”. In order to discourage explicit

learning, participants were told not to focus on learning the rules of the language, but to instead focus on the pictures and the sounds.

Training. The 15-minute training consisted of 4 blocks of 64 trials each, giving 256 trials in total and 16 exposures per item (in Version A only, following Ramscar et al. (2010), there were four additional trials, in which a “control” fribble was presented; this was a blue fribble with a single label: *po frab* (prefix condition) or *frab po* (suffix condition))⁸. The fribbles were presented in pseudo-random order, such that LTF items from one category were not followed or preceded by HTF items from the opposite category, in order to minimize explicit learning. In the prefix condition, the affix was played first (average duration: 225ms), and this was immediately followed by the noun (average duration: 350ms) and the picture of the fribble. The picture and the noun were presented at the same time. The picture remained on-screen for an additional 450ms after the noun had finished, and then the picture disappeared. In the suffix condition, the picture was presented for 450ms, after which the noun was played. The picture remained on-screen for the duration of the noun. After this, the picture disappeared and the affix was played. In both condition, each trials was followed by a blank screen displayed for 1000ms. Importantly, the duration of a single trial, and the amount of time that the picture was on-screen was identical for the two conditions, and the only difference was the order in which labels (affixes) and features (nouns and pictures) were presented (see Figure 5 for a schematic representation).

⁸ The learning of the control category was tested within the Item Learning test, in which a control category item was tested against a randomly chosen experimental category item. This category was originally included in order to detect learners who were not attending to the task. However, performance on the control category test did not predict participants’ performance on any other test, and including it in the models never improved the fit (see <https://osf.io/svpab>), and therefore the control category was not included in Version B.

Training was followed by a series of tests, each tapping a different aspect of item-learning and the learning of discriminating features.

Item Learning Test. We tested participants' knowledge of the associations between individual nouns and individual visual items using a two-alternative forced choice test (2AFC) in which participants saw two items on-screen. After 500ms, participants heard a label from training (stem + affix in the suffix condition, or affix + stem in the prefix condition) that was associated with the target item. This test differed across experiment versions. In Version A, the foil item was a different trained item from the same category, whereas in Version B, the foil item was a novel item from the same category. HTF and LTF items were tested against HTF and LTF foils, respectively. Half of the trained served as targets (chosen randomly), and the remaining half as foils, giving a total of 8 trials – 6 HTF and 2 LTF trials.

For all the tests, participants were told they would have a limited amount of time to respond. If no response was received after 3500ms, a buzzer sounded and no response was collected. A new trial would start after 1000ms.

Semantics and Phonology Tests: Generalization. Generalization on the basis of semantic and phonological features was tested by presenting participants with a previously unseen fribble, and playing two novel stem+affix bigrams. After viewing the picture for 500ms, participants heard the first bigram. As the first bigram was played, a (blank) speech bubble appeared in the bottom left corner of the screen; after 500ms, the second bigram was played, and it was accompanied by a speech bubble in the bottom right corner of the screen. Participants selected the speech bubble corresponding to the audio stimuli matching the picture by pressing the corresponding arrow key on the keyboard. The target and the foil audio bigrams were played in random order. Critically, the foil label matched the target image in terms of the vowel, but not

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

affix. For example, if the target picture was from Category A, and the target label was *foop ge* (vowel A + affix A), the foil label was *kood ma* (vowel A + affix B). In order to perform above chance in this test, participants need to have learned the association between semantic cues and affixes, and/or the association between phonological cues and affixes. (Note that it is not possible, or at least is very unnatural to test an association between semantic features and the affix in the absence of phonology, since we can't remove phonology from the noun). Each test item appeared as the target once. In Version A there were a total of 8 trials in this test, to increase power, this was increased to 16 trials in Version B.

Semantics and Phonology Test: Trained items. The set-up was identical to the generalization test above, except that (i) we used items that had occurred in training and (ii) the same noun was used in the target and in the foil label (e.g., target: *jeed ge*, foil: *jeed ma*), since using different nouns in this test would have required using some novel items as foils (i.e. due to the limited numbers of trained items) meaning participants could ignore the affix and base their responses on their memory of stem-picture pairing (which is already testing in the vocabulary test). As in the generalization test, however, the concern of drawing participants' explicit attention to the affix remained. Given that the generalization test was our key test of interest, we always presented it first in order to avoid interference from any explicit learning from the trained items test. Note that participants could respond correctly either using their memory of specific item-affix correspondences or from the more general associations between affixes and semantic/phonological cues. Each trained item appeared as the target once, giving a total of 16 trials (due to a technical error, the data from this test were not recorded in Version A).

Phonology Generalization Test. We tested the learning of phonological cues in isolation by playing two novel labels (using two speech bubbles, as described above), but without

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

displaying any visual items. Participants were told that they would hear two sounds, one which followed the rules of the language they had been learning, and another which did not, and were instructed to choose the sound that did follow the rules of the language by pressing the corresponding arrow key. The two nouns came from the same noun class but the affix differed. For example, if the target label was *feep ge* (categoryA vowel + categoryA affix), the foil label was *jeed ma* (categoryA vowel + categoryB affix). Note that this is identical to Semantics and Phonology Generalization test, but without a picture. Each test item appeared as the target once, giving a total of 8 trials in Version A and 16 trials in Version B.

Language Awareness Questionnaire (LAQ). Participants were presented with a questionnaire about explicit awareness of the relationship between the fribbles and the affixes (semantics and affix association), and between the nouns and the affixes (phonology and affix association) (see Appendix A). The questionnaire was presented on-screen one question at a time, to prevent participants from altering their responses after learning that there were rules and patterns in the language, as revealed in one of the questions.

Statistical Approach

Our primary method of inference was Bayes Factors (BF). These are used to compute the strength of evidence for H1 over H0 or vice versa. Quantifying evidence of H0 is not possible using frequentist p-values, since a $p > 0.05$ does not provide evidence for the null (indeed in some cases there may be more evidence for H1), despite this common misinterpretation in the literature (Dienes 2008; 2014).

We used the procedure outlined in Dienes (2014) to compute the evidence for H1 over H0 given the data. This procedure requires a model of the data and a model of H1.

The data are modelled using a mean (i.e., a mean difference between two conditions or a comparison to baseline) and a standard deviation. For this we used values from the corresponding logistic mixed effect models described below: the betas were used as mean differences, and SEs of the relevant beta coefficients were used as the SD. Note that using the log-transformed values from logistic mixed-effect models allows us to meet normality assumptions.

To model the H1, we used a half-normal distribution with a mean of 0 and an SD of a rough estimate – x – of the mean difference. Our choice of x depended on the contrast in question. Wherever possible, we used independent data – from a different experiment or from a different condition from the same experiment. For example, if there was evidence for an effect in one experiment, we used that value as the H1 for the same or similar effect in another experiment; or if there was evidence for an effect in one condition, we used that value as the H1 estimate for the other condition. Where using same or similar effects was not possible, we determined a plausible maximum effect and set x to half of this value, since the maximum for the half normal is equal to 2SD. Appendix B provides details of the methods used to compute x from the plausible maximum, and Appendix C shows a table indicating which method was used to compute x for effect.⁹ In-text, we report the Bayes factor, the H0 estimate (always 0) and x (the H1 estimate) as $B_{(H0, H1)}$. Finally, all Bayes factor analyses were one-tailed.

⁹ As noted in Experiment Versions section (Experiment 1), Version B of both this experiment and Experiment 2 were pre-registered. In those two pre-registrations, estimates of effect size used to evaluate x were laid out, generally based on the values obtained from Version A (i.e., an independent data set). We cannot use those values here, as the data from Version A are included in this combined dataset and it is circular to base a measure of effect size on a measure of that same effect from a subset of the same data (see also footnote 7).

Since the choice of value to inform H1 is subject to debate, we also calculated a “robustness region” for each Bayes factor, which shows the range of estimates of H1 (i.e. the value used as so the SD of the theory in the models of the half normal) for which our data would support the same conclusion, noted as RR $[x_1, x_2]$ with x_1 being the smallest SD and x_2 the largest SD leading to the same conclusion. These Robustness Regions should be interpreted bearing in mind larger values of x bias the computation to find evidence for the null, whereas smaller values bias in favour of H1. To find the robustness regions, we tested values of x which are reasonable given the scale, specifically: from 0 in log odds to 4.595 in log odds space (corresponding to odds/odds ratio of 1.041, that is, if one group has minimal performance and is at chance, or we are comparing to chance, and the other has almost perfect performance at 99% accuracy) examining values in increments of 0.01. For those hypotheses where independent data were not available, and where there was no principled way of determining a plausible maximum, we computed Bayes Factors for every plausible value of x and reported the ranges of estimates of H1 for which the data would provide support for the H1, the null, and for which the evidence would be ambiguous.

We interpret BFs as follows: values $<.33$ suggest substantial evidence for H0, whereas values > 3 suggest substantial evidence for H1; values between 0.33 and 3 suggest inconclusive/ambiguous evidence (Jeffreys 1961; Dienes 2008). Given their greater familiarity to the reader, frequentist statistics were computed alongside Bayes Factors using logistic mixed effect models (Baayen et al., 2008) with the lme4 package (Bates et al., 2014) in R (R Core Development Team, 2013). This method is appropriate, as the dependent variable was a correct/incorrect response (binary accuracy data). In all models, all experimentally manipulated variables and all interactions between those variables were included as fixed factors, regardless

of whether they contributed significantly to the model. All fixed factors were centred to reduce collinearity between main effects and interactions, and so that the intercept reflected the grand mean. In each model, participant was included as a random effect, and by-participant slopes for any within-participant effects were used, as recommended by Barr et al. (2013). All reported models converged with Bound Optimization by Quadratic Approximation (BOBYQA) optimization (Powell, 2009). Although we report p -values given their greater familiarity, we do not interpret them, and only refer to the Bayes Factors for statistical inference. We only report and interpret inferential statistics for the fixed effects relevant to our specific hypotheses, along with comparisons to chance for individual cells in the experimental design. The latter are used to shed light on whether there is evidence for “no learning” in some circumstances.

Finally, although statistical analyses were computed in log odds space, our figures plot proportions for ease of visualization.

Results

The data¹⁰ for this experiment are on OSF (<https://osf.io/37ywe>). All the analyses were carried out in R (R Team & others, 2013), and the R script is on OSF (<https://osf.io/svpab>). The results from each test in this experiment are plotted alongside the results from Experiment 2 in Figures 6-9.

Item Learning Test

The data are shown in Figure 6. Recall from the Experiment Versions section that in Version A, this test tested participants’ learning of noun-picture pairs: participants saw two trained pictures from the same category and heard a trained noun-affix label. The task was to

¹⁰ The total data set consisted of 4032 trials (with 56 trials per participants lost due to a technical error in Experiment 1A where no data were collected in the Trained Items test). In each cell, the total proportion of trials lost due to participants not responding within the 3500 ms time window ranged between 2-3%.

select the picture which matched the label. To answer correctly, participants needed to have learned which nouns occurred with which fribble (they cannot use the affix as it is correct for both target and foil). We predicted better item learning in the prefix condition. However, on inspection of Figure 6 (left panel), learning looks to be at chance in all cells. We therefore tested both for greater performance in the prefix than suffix condition and for above chance performance in each cell. In the absence of prior data to inform our estimate of H1 (and since low overall performance means we cannot use values from within the current data set) we computed Bayes Factors for every plausible value of H1. This revealed that there are no plausible estimates of H1 for which the current data would provide evidence for learning in any cell (prefix HTF: $\beta = 0.06$, $SE = 0.14$, $z = 0.41$, $p = 0.679$, $B_{(0 [0.581: \infty])} < .33$); prefix LTF: $\beta = -0.11$, $SE = 0.25$, $z = -0.46$, $p = 0.648$, $B_{(0 [0.481: \infty])} < .33$; suffix HTF: $\beta = 0.02$, $SE = 0.14$, $z = 0.14$, $p = 0.886$, $B_{(0 [0.461: \infty])} < .33$; suffix LTF: $\beta = -0.24$, $SE = 0.24$, $z = -0.96$, $p = 0.335$), $B_{(0 [0.331: \infty])} < .33$), meaning that there is a floor effect.

Version B used an easier item-learning test where participants again matched the label to one of two pictures, but the foil was an unseen fribble rather than another trained fribble (see Stimuli). To answer correctly, participants needed to distinguish between seen and unseen items of the same type using their idiosyncratic features, and here again we predicted better learning in the prefix condition. While learning in some cells – specifically, in LTF cells – was above chance (prefix LTF: $\beta = 0.84$, $SE = 0.16$, $z = 5.46$, $p < .001$, $B_{(0,0.84)} = 646149.23$, $RR[0.03, >4.5]$; suffix LTF: $\beta = 1.1$, $SE = 0.16$, $z = 6.75$, $p < .001$, $B_{(0,0.84)} = 1340315744.04$, $RR[0.03, >4.5]$; prefix HTF: $\beta = 0.08$, $SE = 0.08$, $z = 0.94$, $p = 0.347$, $B_{(0,0.84)} = 0.25$, $RR[0.63, \infty]$; suffix HTF: $\beta = 0.06$, $SE = 0.08$, $z = 0.68$, $p = 0.494$, $B_{(0,0.84)} = 0.19$, $RR[0.47, \infty]$), there was substantial evidence for the null as to the hypothesized prefixing advantage (prefix: $M = 56.18\%$,

$SD = 14.26\%$, suffix: $M = 56.76\%$, $SD = 14.82\%$; $\beta = -0.05$, $SE = 0.11$, $z = -0.44$, $p = 0.663$, $B_{(0,0.29)} = 0.26$, $RR[0.23, \infty]$).

Although not part of our original predictions, since only LTF items were above chance, we tested the hypothesis that overall, LTF items were better identified than HTF items: this was the case ($\beta = 0.9$, $SE = 0.12$, $z = 7.34$, $p < .001$, $B_{(0,0.29)} = 6568219972.23$, $RR[0.02, >4.5]$), indicating that participants were better at identifying idiosyncratic features of trained fribbles

Semantics and Phonology Test: Generalization

The data are shown in Figure 7. Our key prediction was regarding an affix-by-type-frequency interaction, specifically, that participants in the prefix condition will show a greater type-frequency effect than participants in the suffix condition, with weaker generalization over low frequency features, due to their being unable to “unlearn” features which are high-frequency but low predictive-value, in favour of low-frequency but high predictive-value features. There was evidence for this interaction ($\beta = 0.58$, $SE = 0.18$, $z = 3.17$, $p = 0.001$, $B_{(0, 0.42)} = 55.05$, $RR[0.07, >4.5]$). Looking at the conditions separately, there was a simple effect of type-frequency in the prefix condition ($\beta = 0.69$, $SE = 0.13$, $z = 5.16$, $p < .001$, $B_{(0, 0.1)} = 248.57$, $RR[0.04, >4.5]$), but this was ambiguous in the suffix condition, ($\beta = 0.1$, $SE = 0.14$, $z = 0.76$, $p = 0.45$, $B_{(0, 0.69)} = 0.4$, $RR[0, 0.83]$).

We also predicted that, across HTF and LTF features, participants in the suffix condition will perform better due to better overall learning of the discriminating semantic and phonological features for each affix. However, there was substantial evidence for the null for this prediction (suffix: $M = 61.1\%$, $SD = 19.206\%$; prefix: $M = 56.61\%$, $SD = 16.8\%$; $\beta = -0.12$, $SE = 0.12$, $z = -0.95$, $p = 0.344$, $B_{(0, 0.45)} = 0.14$, $RR[0.18, \infty]$).

Semantics and Phonology Test: Trained items

Recall that, due to a technical error, data were collected for this test in Version B only (120 participants). The data are shown in Figure 8.

This test tested whether participants could associate a trained stem-fribble pair with the correct affix.¹¹ Our predictions were less clear for this test, since although participants could choose the correct affix based on item-specific learning, they could also use the same cues as in the generalization task, in which case we would predict the same affix-by-type-frequency interaction as in the previous generalization task. However, the evidence for the interaction was ambiguous ($\beta = 0.4$, $SE = 0.27$, $z = 1.48$, $p = 0.138$, $B_{(0, 0.58)} = 1.88$, $RR[0, 4.39]$). To elucidate this finding, we conducted further analyses (not included in the pre-registration). First, we looked for a three-way affix-by-type-frequency-by-test interaction comparing the two-way interaction for novel and trained items, which was ambiguous ($\beta = -0.14$, $SE = 0.35$, $z = -0.39$, $p = 0.695$, $B_{(0, 0.308)} = 0.54$, $RR[0, 0.73]$). Second, given that it was unclear whether this test would tap generalization or item-learning, we looked at correlations between participants' performance between this task and the Item Learning and Semantic and Phonology Generalization tasks.¹² This showed evidence for a positive correlation for both the item learning test ($r(118) = 0.21$, $SE = 0.092$, $p = .02$, $B_{(0, 0.203)} = 7.33$, $RR[0.06, 0.83]$) and for the semantics and phonology generalization test ($r(118) = 0.57$, $SE = 0.092$, $p < .001$, $B_{(0, 0.343)} = 4592572054$, $RR[0.02, >4.5]$), suggesting that learning the associations between affixes and trained items was related to learning both the features relevant for generalization and the idiosyncratic features of trained

Phonology Generalization Test

¹¹ Although not part of the predictions related to our hypotheses, performance is expected to be better in this test (trained items) than in the previous generalization test, and our analyses found this to be the case (both here and in Experiment 2; see online analyses <https://osf.io/svpab>).

¹² We thank an anonymous reviewer for suggesting these analyses.

The data are shown in Figure 9. In this test, participants were asked to determine the correct affix for novel stems using phonology only (no picture was shown). We predicted that the suffix condition would be better than the prefix condition at learning the association between vowels and affixes, for example, that all nouns which contain the *-oo* sound occur with affix *ma*. The evidence was ambiguous (suffix: $M = 54.79\%$, $SD = 13.59\%$; prefix: $M = 53.53\%$, $SD = 14.57\%$; $\beta = 0.1$, $SE = 0.09$, $z = 1.11$, $p = 0.267$, $B_{(0, 0.2)} = 1.14$, $RR[0, 0.83]$). There was evidence for above-chance performance in both conditions (suffix: $\beta = 0.25$, $SE = 0.06$, $z = 4.01$, $p < .001$, $B_{(0, 0.15)} = 752.81$, $RR[0.02, >4.5]$; prefix: $\beta = 0.15$, $SE = 0.06$, $z = 2.46$, $p = 0.014$, $B_{(0, 0.25)} = 8.34^{13}$, $RR[0.04, 0.8]$).

Language Awareness Questionnaire

Sixty-four percent of participants reported no explicit awareness of the patterns in the language (see Appendix A on how the questionnaire responses were coded). No participants showed any awareness of the role of the discriminating features required for generalization with the LTF items, despite showing above-chance generalization of these in the suffix condition. Some participants did appear aware of the features relevant to generalization of HTF items (i.e. the fribbles shape), as well as the relationship with vowels in the nouns. To assess the effect of explicit awareness on learning, for each type of generalization, we removed the participants who reported explicit awareness of the pattern critical for that generalization and repeated the analyses. We found that the effects of interest persisted after removing participants whose reports suggested explicit awareness (analyses available at: <https://osf.io/kd4br/>).

Discussion

¹³ For Experiment 2B alone, evidence was ambiguous ($B=2.883$)

Experiment 1 tested the effects of linear order in the context of suffixing versus prefixing on generalization and item learning. We discuss each in turn.

Starting with generalization, the key positive finding is that the type-frequency effect was greater in the prefix condition compared to the suffix condition (affix-by-type-frequency interaction). This result confirms the key theoretical prediction of this work, based on the computational model as well on the results of Ramscar et al. (2010). The prediction comes from the differences in the dynamics of cue structure over the course of learning in the two conditions. Specifically, in the suffix condition, prior to hearing the suffix, the learners view the picture and hear the stem, meaning that numerous semantic (visual) and phonological (auditory) features can compete for predictive value for the outcome (the affix). These cues compete for predictive value, and ultimately, via prediction error, informative cues gain higher predictive value than uninformative cues, resulting in accurate generalization. In the prefix condition, on the other hand, the hypothesis is that no comparable cue competition occurs, given that the learners are only exposed to the semantic and phonological features of the stem and the picture *after* the affix was heard. As a result, in the prefix condition, learning reflects the frequencies of cue-outcome co-occurrence (associative learning), but there is no discrimination of cues based on predictive value. Therefore, when generalizing to novel items, learners in the prefix condition are only successful when the predictive value of a cue is positively correlated with the frequency of its co-occurrence with the outcome. In our learning context, this is the case for HTF features only, whereas with LTF generalization, some features co-occur more frequently with the *opposite* outcome, and as such are not *predictive*. Therefore, we predicted that participants in the prefix condition will show a strong advantage in HTF generalization compared to LTF generalization, whereas this will not be the case in the suffix condition. There was strong evidence for this in the

data, with strong evidence for an effect of type frequency in the prefix condition, and ambiguous evidence in the suffix condition (meaning that although we have evidence that the type-frequency effect is smaller in the suffix condition, we cannot say whether it is absent – and recall that a small type-frequency effect *was* found in the model). This finding provides evidence for the discriminative learning framework, and is directly analogous to the finding in Ramscar et al. (2010) where participants showed better learning of features discriminating word meanings when features were viewed before the label was played, particularly for LTF items.

An additional prediction from this test was that we would observe *overall* better generalization in the suffix condition compared to the prefix (for HTF and LTF features alike). This is predicted by the computation modelling in this work, two previous studies comparing generalization for prefixes and suffixes (Ramscar, 2013; St Clair et al., 2009), and by an equivalent main effect of order in the word learning experiment by Ramscar et al. (2010). However, in the current experiment, no evidence for this overall benefit of suffixing was found, with evidence for the null. Recall from the modelling section that this prediction was derived from the choice axiom used to assess model performance, which itself is not part of learning theory – in the General Discussion, we shall re-visit this and consider some theoretical reasons why modelling human performance with this metric may not always be appropriate.

Turning to the Phonology Generalization test, we predicted that the suffix condition should outperform the prefix condition, based on the computational model and previous work which manipulated the learning of phonology (St Clair et al., 2009). However, although the suffixing condition showed numerically higher performance than the prefix condition, and we only had evidence of above-chance performance in the suffix condition, the evidence for above-chance learning in the prefix condition and for a difference between conditions was ambiguous.

This means that we cannot draw any conclusions for this hypothesis. The relatively low performance in this test contrasts with the strong learning of “phonological” features in the computational model since these are fully consistent (-*ee* always occurs with one affix, -*oo* with the other). We return to this point in the General Discussion. It is worth noting that performance in this test was also low compared with the test where semantic features were available alongside the phonology, indicating the important role of semantic features in learning.

Finally, results of the informal analysis of the LAQ suggest that the generalization was not mainly driven by explicit awareness of the patterns in the language. However, we acknowledge that the questionnaire may not be a reliable measure of explicit awareness – it is possible that participants who performed above chance were explicitly aware of the correct patterns, but were unable (due to the objects being novel) and/or unwilling to verbalize them. Nevertheless, other studies *have* found that generalization is contingent on explicit awareness, as captured by a similar questionnaire (Brown et al., 2018) and we see the questionnaire as a useful supplement for understanding participants’ performance. We return to this in the General Discussion.

Turning to the Item Learning test, we predicted better item learning in the prefix condition compared to the suffix condition. This hypothesis stems from an entropy-based analysis of how uncertainty is distributed over the utterance during learning. In the prefix condition, the entropy is smoothed over the prefix-noun pair, such that after the prefix is known, only a half of all possible nouns can occur, thus reducing the uncertainty of the noun. Following previous work (Arnon & Ramscar, 2012; Ramscar, 2013), we hypothesised that this makes it easier to process the noun and therefore to associate that word with the corresponding picture. In the suffix condition, on the other hand, this does not happen, as the entropy of the utterance is

not distributed between the noun-suffix pair. Instead, all uncertainty is carried by the noun: at the point before the noun is heard, any noun is equally likely, but after the noun is heard, only one affix is possible and therefore there is no uncertainty. We predicted that this would be reflected in poorer learning of noun-picture pairs (“item” learning) in the suffix condition. However, in Version A, a floor effect meant that we were not able to draw conclusions as to this prediction, and thus in Version B, a less demanding item learning test was used: rather than having participants choose between two trained items, participants were presented with one trained and one untrained item. Although this did not test participants’ ability to differentiate trained items from each other, it did require them to recognize trained items, which they could *not* do if they were only tracking the features which are relevant for generalization. Our prediction still holds for this test: if the processing load was distributed more evenly across the utterance in the prefix condition, and if this resulted in stronger associations between the noun and the picture, participants in this condition should find it easier to select the correct picture *given* the noun in this condition. Again, we expected this to be better in the prefix condition, where processing load was smoothed over the utterance.

While the new version of the test did show item-learning, contrary to our prediction, there was no prefixing benefit, with substantial evidence for the null – we return to this in the General Discussion. However, in both affix conditions, LTF items were learned better than HTF items, with cell-by-cell comparisons to chance showing substantial evidence for the null for HTF items in both conditions. This difference cannot be understood in terms of the frequency of *individual* fribbles, since HTF and LTF fribbles were each seen the same number of times in training. Although we did not originally predict this difference, it is sensible in terms of the information structure of the *fribble* sets: participants are only able to recognize *fribbles* when they come from

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

the category with fewer members (LTF) since discriminating between individual items from the same category is perceptually harder, and particularly so when the set size is larger (meaning greater uncertainty).

In addition to the Item Learning test, we collected data in a second test using trained items where, rather than testing the learning of stem-picture pairs, we tested the learning of trained item-affix pairs (i.e., stem-picture to affix) in the same set-up as the Semantics and Phonology Generalization test, but with fribbles which had occurred during training. Although we did not have a clear prediction, if participants also rely to some extent on the features used in generalization, we would expect the same affix-by-type-frequency interaction we observed with novel items. The evidence for an interaction in the data was ambiguous, as was the evidence for a three-way by-test interaction (novel versus trained items). This means that we cannot say whether or not an interaction was present, or whether it was smaller than that observed in generalization. Given this, we cannot draw conclusions as to whether participants in this test rely more on general or item specific cues, though we note that additional analyses found evidence that performance in this test was related to that in both the Item test and the Semantics Phonology generalization tests. We also note that this test was always administered *after* the generalization test, and it is possible that this ordering also affected performance.

To summarize, Experiment 1 provided evidence for the critical role of prediction error in linguistic generalization. In particular, participants in the prefix condition had difficulty with LTF discriminating features, resulting in as strong effect of type-frequency, whereas this effect was smaller in the suffix condition, where greater cue competition provided greater error signal in learning, allowing participants to learn the relevant discriminating but low frequency cues. With respect to item-learning, while it is encouraging that the test used in Version B elicited

item-learning, this was only true for LTF-items, and, contrary to prediction, there was no benefit of prefixing over suffixing. One possibility is that stronger item-learning is required to observe an effect of affix. In Experiment 2, we address this possibility by reducing the vocabulary size by 50%. Rather than training participants on six HTF and two LTF items per category, in Experiment 2, participants will be trained on three HTF and one LTF item per category (thus preserving the 3:1 ratio of items). Our hypothesis regarding generalization is unchanged, since the category structure of stems and their relationships with affixes, and thus the cue structure, remains unchanged. This was also confirmed by re-running the computational model with a smaller number of idiosyncratic features reflecting the smaller number of items (16 as opposed to 32, representing semantic and phonological idiosyncratic features for 8 rather than 16 nouns). We do not report the full results of this simulation, since they are highly similar to those in Experiment 1. Critically, we found the same key effects in generalization: (1) overall suffixing benefit and (2) a greater effect of type-frequency in the prefix condition compared to the suffix condition¹⁴. Note also that our findings regarding entropy reduction (on which we base our predictions for item-learning) also hold for this smaller set of items, where uncertainty is reduced from a set of eight possible items in the suffix condition, to four in the prefix condition (as hearing the prefix eliminates the four items in the other category).

¹⁴ As noted in the introduction, other studies show that that decreasing the number of lexical items can reduce generalization (e.g., Wonnacott et al., 2012; Tamminen et al., 2015). This can be explained under discriminative learning, since increasing lexically-specific features generally increases cue competition and thus learning of the invariant cues which are key to the generalization. However, importantly, weaker generalization is *not* predicted in Experiment 2 compared with Experiment 1 – and is not found in the computational simulation – since what is key in the current design is the *relative* strength of the association between the predictive features and the affixes, and how this interacts with affix order, and this is identical across the two experiments. Thus, for both input sets, what matters for generalization is the competition between the discriminating and non-discriminating features, which is the same across the two input sets.

Finally, Experiment 2 also provides as conceptual replication for the findings regarding generalization. As in Experiment 1, two versions of this experiment were run, but in both versions we used the item-test from Version B of Experiment 1 (because the item learning test from Version A showed a floor effect).

Experiment 2

Method

Experiment Versions

Same as Experiment 1.

Participants

Two hundred and thirty participants took part in the experiment. For Version A, sample size was determined via a power simulation based on the interaction we observed in Experiment 1 Version A (see <https://osf.io/258w3/>). Data were inspected only once, after the full sample of 110 had been collected. For Version B, we used the same pre-registered optional stopping procedure as for Experiment 1 Version B. Although as in Experiment 1, there was evidence for the key interaction with 30 participants, since we had additional resources (bearing in mind that Bayes Factors remain a valid measure of evidence regardless of data collection procedure; see footnote 4), we collected an additional 60 participants (in one batch), making the final sample $N = 120$.

Stimuli

These were drawn from the set as in Experiment 1, however, there were fewer training audio and visual items. In the current experiment, there were 8 training items in total: four per category, three HTF, and one LTF item per category (Figure 10).

Procedure

Identical to Experiment 1, except that in training, there were 32 exposures per item (same total number of training trials as in Experiment 1), and that in both versions of the experiment, we used the Item Learning test from Experiment 1 Version B.

Statistical Approach

Same as Experiment 1.

Results

The data¹⁵ for this experiment and the R analysis script are on OSF (<https://osf.io/37ywe> and <https://osf.io/svpab>, respectively).

Item Learning Test

We predicted better learning of trained stem-fribble pairs in the prefix condition than in the suffix condition, but the evidence for this was ambiguous ($M = 57.78\%$, $SD = 21.6\%$; suffix: $M = 55.51\%$, $SD = 18.49\%$; $\beta = 0.13$, $SE = 0.14$, $z = 0.99$, $p = 0.322$, $B_{(0, 0.6)} = 0.59$, $RR[0, 1.09]$) (see Figure 5).

As in Experiment 1, LTF items were learned better than HTF items (HTF: $M = 58.94\%$, $SD = 23.71\%$; LTF: $M = 86.25\%$, $SD = 23.71\%$; $\beta = 1.3$, $SE = 0.23$, $z = 5.57$, $p < .001$, $B_{(0, 0.6)} = 516951.89$, $RR[0.05, >4.5]$). However, unlike in Experiment 1, learning was above chance in every cell (prefix HTF: $\beta = 0.27$, $SE = 0.09$, $z = 3.06$, $p = 0.002$, $B_{(0, 1.82)} = 10.44$, $RR[0.03, >4.5]$; prefix LTFT: $\beta = 1.82$, $SE = 0.3$, $z = 6$, $p < .001$, $B_{(0, 1.82)} = 13065062.04$ ¹⁶, $RR[0.06, >4.5]$; suffix HTF: $\beta = 0.27$, $SE = 0.09$, $z = 2.97$, $p = 0.003$, $B_{(0, 1.82)} = 8.09$, $RR[0.04, >4.5]$; suffix LTFT: $\beta = 1.3$, $SE = 0.26$, $z = 5.09$, $p < .001$, $B_{(0, 1.82)} = 92737.02$, $RR[0.06, >4.5]$) and

¹⁵ The total data set consisted of 6160 trials. In each cell, the total proportion of trials lost due to participants not responding within the 3000ms time window ranged between 2-4%.

¹⁶ For Experiment 2B alone, evidence was ambiguous ($B=2.743$)

learning was overall better in this experiment compared to Experiment 1 (Version B) ($\beta = 0.22$, $SE = 0.08$, $z = 2.61$, $p = 0.009$, $B_{(0, 0.41)} = 10.67$, $RR[0.05, 1.93]$).

Semantics and Phonology Test: Generalization

The data are shown in Figure 6. As predicted – and as in Experiment 1 – there was evidence for an interaction indicating that type-frequency had a stronger effect on generalization in the prefix than in the suffix condition ($\beta = 0.5$, $SE = 0.21$, $z = 2.35$, $p = 0.019$, $B_{(0, 0.58)} = 7.66$, $RR[0.12, 2.08]$). While (as in Experiment 1) in the prefix condition, generalization was better with HTF compared to LTF features ($\beta = 0.56$, $SE = 0.17$, $z = 3.34$, $p = 0.001$, $B_{(0, 0.69)} = 91.45$, $RR[0.06, >4.5]$), in the suffix condition, there was evidence for no difference between HTF- and LTF-features ($\beta = 0.07$, $SE = 0.16$, $z = 0.46$, $p = 0.649$, $B_{(0, 0.69)} = 0.33$, $RR[0.68, \infty]$), whereas this was ambiguous in Experiment 1. The evidence for the difference in the key interaction between the two experiments was ambiguous ($\beta = 0.1$, $SE = 0.28$, $z = 0.36$, $p = 0.721$, $B_{(0, 1)} = 0.36$, $RR[0, 1.09]$), although the Bayes factor suggested more evidence for the null.

We also predicted overall better generalization in the suffix condition compared to the prefix condition (a main effect of affix) but as in Experiment 1, the evidence for this was ambiguous (prefix: $M = 63.6\%$, $SD = 17.06\%$; suffix: $M = 58.8\%$, $SD = 15.92\%$; $\beta = 0.11$, $SE = 0.11$, $z = 1.04$, $p = 0.297$, $B_{(0, 0.45)} = 0.66$, $RR[0, 0.94]$).

Semantics and Phonology Test: Trained Items

The data are shown in Figure 7. If participants use the same features as in the generalization test, performance should be affected by the same factors as generalization, meaning we can expect an interaction indicating a larger effect of type-frequency in the prefix condition. However, as in Experiment 1, there *was* evidence for this interaction in the data (β

= 0.18, $SE = 0.26$, $z = 0.67$, $p = 0.5$, $B_{(0,0.58)} = 0.72$, $RR[0, 1.42]$), and the three-way affix-by-type-frequency-by-test interaction comparing the two-way interaction for novel and trained items was ambiguous ($\beta = 0.03$, $SE = 0.03$, $z = 1.03$, $p = 0.303$, $B_{(0, 0.14)} = 0.66$, $RR[0, 0.29]$). Finally, as in Experiment 1, we computed correlations between participants' performance between this task and the Item Learning and Semantic and Phonology Generalization tasks. This showed evidence for a positive correlation in both cases (Item Learning test: ($r(228) = 0.2$, $SE = 0.092$, $p = .002$, $B_{(0, 0.213)} = 41.421$, $RR[0.06, 0.83]$); Semantics and Phonology Generalization test: ($r(228) = 0.33$, $SE = 0.092$, $p < .001$, $B_{(0, 0.648)} = 111730$, $RR[0.02, >4.5]$), suggesting that learning the associations between affixes and trained items was related to both their ability to learn the cues relevant for generalization and the idiosyncratic features of trained items.

Phonology Generalization Test

We predicted stronger performance in the suffix condition due to greater learning of the discriminating phonological features which reliably predict affix occurrence. However, as for Experiment 1, the evidence for this was ambiguous ($\beta = 0.02$, $SE = 0.08$, $z = 0.25$, $p = 0.804$, $B_{(0,0.2)} = 0.46$, $RR[0,0.28]$). There was evidence for above-chance performance in both conditions (suffix: $M = 53.77\%$, $SD = 15.39\%$; $\beta = 0.15$, $SE = 0.06$, $z = 2.66$, $p = 0.008$, $B_{(0, 0.25)} = 13.01$, $RR[0.04, 1.3]$; prefix: $M = 52.76\%$, $SD = 11\%$; $\beta = 0.13$, $SE = 0.06$, $z = 2.32$, $p = 0.02$, $B_{(0,0.25)} = 5.79^{17}$, $RR[0.04, 0.55]$).

Language Awareness Questionnaire

Overall, 55% of participants reported no explicit awareness of the patterns in the language (see Appendix A). As in Experiment 1, the evidence for all effects of interest in the two

¹⁷ For Experiment 2B alone, evidence was ambiguous ($B=1.295$)

generalization tests persisted after removing participants who reported explicit awareness of the patterns critical for those effects.

Discussion

Experiment 2 was identical to Experiment 1, except that participants were exposed to only half of the training set: eight items in total, three HTF and one LTF item per affix. We predicted that training participants on fewer items (but with matched total exposure) would lead to better item learning, which might in turn unmask a prefixing advantage. We also predicted that the differences between conditions for generalization seen in Experiment 1 should hold for this experiment, despite the smaller training set, since basic cue structure of the languages remains unchanged (as confirmed by computational modelling). Below we briefly summarize the key results and comparisons to Experiment 1; we discuss the common findings of the two experiments and their relation to theory more extensively in the General Discussion.

As expected, reducing the number of trained items led to overall better item learning, with evidence for substantially better learning in this condition. We hoped that this stronger learning would allow us to find evidence for our original prediction of better item learning in the prefix condition than the suffix condition, but the evidence for this was ambiguous.

The pattern of evidence for generalization was remarkably similar compared to Experiment 1. With respect to our key prediction, there was substantial evidence for a stronger effect of type-frequency on generalization in the prefix condition compared to the suffix condition (an affix-by-frequency interaction), where, as in Experiment 1, participants in the prefix condition struggled with generalizing from low-frequency predictive cues. This result is directly predicted by the theory, given that the ordering of stem and affix means that participants in the prefix condition do not benefit from cue competition, and thus cannot dissociate the high frequency non-predictive cues in favour of the low frequency cues which predict the affixes (for

LTF items). Notably, this effect persisted with a significantly smaller learning set, suggesting that the same learning mechanisms were at play in Experiments 1 and 2, regardless of the different set sizes.

Turning to our predictions regarding *overall* better generalization in the suffix condition, the evidence for this was ambiguous both in the Semantics and Phonology test (whereas in Experiment 1 there was evidence for the null) and in the Phonology generalization test (ambiguous in Experiment 1, too). Note that for Phonology generalization, performance was relatively poor, suggesting a possible floor effect.

As in Experiment 1, we included the somewhat exploratory Semantics and Phonology Trained Items test, where participants mapped affixes to trained nouns. As in Experiment 1, the affix-by-type-frequency interaction, which was seen in generalization, was ambiguous in this test, and there was ambiguous evidence for a three way test-by-affix-by-type-frequency interaction. Therefore, we cannot say whether this interaction is present in this test, or whether it is smaller than with generalization items.

Finally, the findings from the LAQ revealed that, as in Experiment 1, most participants did not report explicit awareness of any patterns in the language. Critically, evidence of generalization for both HTF and LTF semantic and phonological features persisted even when we excluded participants who explicitly reported noticing these features, suggesting that learning was largely implicit.

General Discussion

The current study used artificial language learning to explore predictions of a discriminative learning approach to language, whereby language learning is a predictive process of discriminating between informative and uninformative cues from both the linguistic and non-

linguistic environment. Critically, this discrimination cannot occur without cue competition and negative evidence from prediction error. Under this view, “generalization” is the extension of appropriate predictive features to novel linguistic contexts and the theory predicts that this may be stronger or weaker depending on the ordering of linguistic elements.

We tested this idea by exposing participants to two artificial languages in which novel objects (“fribbles”) were associated with novel noun-stems as well as with novel affixes which either preceded (prefixing condition) or followed (suffixing condition) the noun. The assignment of affixes was such that the use of a particular affix was predicted by the features of the stem – both semantic (visual properties of the referent fribble) and phonological (vowel in the stem). We predicted greater learning of the features associated with the affixes – and thus more accurate generalization of the affixes to novel stems – in the suffixing condition than in the prefix condition, since only the stem-affix order allows for cue competition over the features of the stem. Specifically, we predicted this effect to be strongest for items where appropriate generalization required the learner to reject frequent, but uninformative cues in favour of cues which were low frequency but predictive (the low-type-frequency or LTF nouns in our design) leading to a type-frequency effect (Ramscar et al., 2010). We also made predictions about the learning of the idiosyncratic properties of individual items (item learning) in our paradigm. Specifically, we predicted better item-learning in the prefix condition, based on an information-theoretic approach which suggested that when an affix precedes a lexical item it smooths the entropy of that item (Dye et al., 2017, 2018). This is known to aid processing and is predicted to lead to better learning of prefixed items (based on Arnon & Ramscar, 2012; Ramscar, 2013).

These predictions were tested in two versions of the same artificial language which differed only in vocabulary size. The key results were remarkably similar across the two

experiments. First, there was evidence for poorer learning of low frequency discriminating features in the prefix condition compared to the suffix condition, evidenced by an affix-by-type-frequency interaction, although generalization was not overall better in the suffix condition compared to the prefix condition. Second, for item learning we did not find the predicted benefit of prefixing compared with suffixing. We discuss the findings for generalization and item learning in turn, then turn to consider implications and finally limitations and future directions.

Generalization

The key positive finding in the current study concerned the ability to determine affixes for novel items (generalization) and specifically how this was affected by the frequency of the cues relevant for generalization (our type-frequency manipulation). Across two experiments, the type-frequency manipulation had a greater effect in the prefix than the suffix condition, evidenced by an affix-by-type-frequency interaction. We explain this result in terms of competition between a high frequency salient feature (body shape) and the lower frequency features which are *predictive* of affix use (i.e., for LTF type nouns). This interacts with how the prefix/suffix ordering influences cue structure: appropriate generalization requires cue competition across the features of the items, and this can only occur in the condition where the order is such that these *features* are cues to the affix, leading to relatively poor performance with LTF nouns in the prefix condition.

This finding directly reflects the result of Ramscar et al. (2010), though in the domain of affix learning, and more generally adds to an existing body of evidence showing that linguistic generalization is a process of discriminating predictive from non-predictive cues in the input through a competitive, error-driven learning mechanism.

Interestingly, we did not see *overall* better generalization in the suffix condition than in the prefix condition. This would have occurred if this condition showed better generalization of

both HTF and LTF semantic features, and of the phonological features. In fact, there was no evidence for this in any generalization test in either experiment (Experiment 1 showed evidence for the null in the Semantics and Phonology test) and generalization of HTF items was on average higher in the *prefix* condition (Figure 7) (although since not part of our original predictions, we did not test for this statistically). How can this be understood in terms of the discriminative learning theory?

To answer this, we return to the computational modelling, where we predicted an overall benefit of suffixing since the suffix-model had a greater probability of “choosing” the correct affix for both HTF and LTF novel items. However, observing the figure plotting raw weights over time (Figure 3), we note that, while the negative weights between the discriminating features as the *opposite* affix are greater in the suffix condition, the positive weights between the discriminating features and the *correct* affix are in fact greater in the prefix condition. These positive weights are particularly high for the most frequent cues which are also fully predictive of affix occurrence for HTF items (body shape and colour of the fribbles, as well as the vowel), and overall 0.75 predictive of the affix. This stronger positive association in the prefix condition is not captured by the metric which we used to assess the model (following Ramscar et al., 2010), that is, the choice axiom, which fundamentally captures the fact that the probability of choosing an option from one set is related to the probability of choosing the same option from a different set. This means that the probability of choosing one affix for a particular set of features (i.e., the sum of associative weights between the features and the affix) is related to the probability of choosing the other affix (i.e., the sum of associated weights to the other affix). Viewed in this way, the metric rewards greater discrimination between the two options, that is, stronger “unlearning” of the opposite affix, and may predict a suffixing advantage in learning

situations where human learners do not in fact rely on unlearning non-discriminating features, or can “get by” without it. An important next step in the future work should be to explore alternative metrics which may be better suited for assessing learning from positive evidence. It is also important to note that the choice axiom (or any alternative metric for that matter) is not inherent to the Rescorla-Wagner model – Rescorla and Wagner (1972) analysed raw associative strengths for different cues and asymptote, and the discriminative learning framework for language is vague about how different raw weights may translate to how humans generate responses. Future work should investigate how tests such as the 2AFC task relate to the model, and in our own ongoing work we are exploring alternative tasks which may be more directly related to associative strength in the model (e.g., contingency judgment ratings, see also Kopp & Wolff, 2000). It may also be useful to revisit the previous relevant work (Ramscar et al., 2010; Ramscar, 2013; St Clair et al., 2009) and consider how the different tasks they employ might relate to different evaluation metrics for the model, and how this relates to the overall better generalization in the suffix condition that this work reported.

Before leaving the comparison between humans and the model in generalization, we consider a surprising discrepancy in the Phonology Generalization test. In addition to not seeing the predicted suffixing benefit, overall human performance was very weak, whereas the model showed very strong learning due to the fact that the “phonological” feature was fully predictive of each affix. This indicates that, for human learners, the phonological features are less salient than the semantic features, despite their greater consistency. This may help to explain the different performance with phonological features reported in studies which only involved such features (e.g., St Clair et al., 2009), where participants heard novel words and affixes without associated referents. In our study, the input contained redundant features: learners could

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

generalize correctly based on semantic features alone or phonological features alone. It is possible that semantic features were more salient than phonological features in our study, and were learned first, which may have in turn blocked the learning of phonology. Indeed, Culbertson et al. (2017) found that, when semantic and phonological cues to noun class membership were aligned in an artificial language, adult learners relied on semantic cues (but when phonological cues were taught before semantic cues, learning was better for the former, indicating the same blocking effect). Interestingly, when the same paradigm was used with 6-7-year-olds, children relied on phonological cues over semantics (Culbertson et al., 2019). It would therefore be interesting to see whether children would also show better learning of phonological than semantic cues in our paradigm. However, note that in our paradigm, it is not possible to draw strong conclusions about the learning of phonology versus semantics since the manipulation of these cues is different (we do not have a comparable type-frequency manipulation with phonological cues). Nevertheless, it is notable that phonological learning in this experiment is relatively weak. This highlights that our theoretical approach, and the model, is based on a naïve learner, and therefore does not capture biases that human learners may have towards different types of cues. Further exploration of the role played by these different information sources is an important area for future research.

A final question we considered was whether mechanisms of generalization in this paradigm are implicit. Using a post-experiment questionnaire, we found that most participants were not able to report any of the features relevant to generalization, and, critically, generalization remained even when those participants who *did* notice the relevant features were removed. While we acknowledge that this method is imperfect (for example, it is possible that participants may have explicit knowledge of a feature that they cannot clearly articulate; also,

where participants do have explicit knowledge, this could have emerged via implicit learning processes), other artificial language work has found questionnaires useful in revealing that learning is largely explicit. For example, Brown et al. (2018) taught children and adults artificial languages somewhat similar to those in the current study, and generalization with novel nouns, although above chance at group-level, was entirely driven by a small group of participants who had explicitly noticed the semantic conditioning. Here, we found no evidence that consciously identifying the semantic generalizations benefitted performance on the generalization test. Interestingly, this is despite our 2AFC generalization task requiring participants to explicitly decide about the match/mismatch of an affix and a set of features. Other experiments have found that such explicit tasks may tap a rather different type of generalization to that measured with implicit tasks, and critically that these more explicit tasks may not capture differences in generalization as a function of input statistics (e.g., Tamminen et al., 2015). In contrast, our 2AFC task does show differences due to our manipulations of input structure (and does so in the context of a single training session, although Tamminen et al., 2015 only reported such generalization after a two-day period of consolidation). We take this as a further indication that our test taps into knowledge acquired via implicit learning mechanisms, and suggest that such learning has been encouraged by the speeded procedure we used to present stimuli in the training stage of the experiment.

In sum, the results of the current experiment are consistent with an account in which generalization is underpinned by implicit learning mechanisms in line with discriminative learning theory, and broadly as implemented in the computational models presented in this study. Our discussion has highlighted the difficulties in assessing learning in the models and linking

this to human performance, and the fact that non-naïve human learners may bring biases which are not captured in this computational implementation.

Item Learning

The key test here was the Item Learning test, in which we tested participants' recall of trained frubbles. However, we did *not* find evidence for a prefixing advantage. Recall that we predicted this not only on theoretical grounds – based on the idea that prefixes aid learning and processing by reducing the entropy of upcoming nouns (Dye et al. 2017, 2018) – but also based on experiment findings (Arnon & Ramscar, 2012; Ramscar, 2013). Arnon and Ramscar (2012) exposed participants to either articles (akin to prefixes) + nouns *before* hearing nouns in isolation or the reverse, and they report better learning of noun-object pairings in the latter case (as well as article-noun pairs which was in fact the key focus of that study). The authors suggested the result is due to participants' *first* experiencing the nouns with the preceding articles, which allowed them to benefit from the hypothesized aid to processing and learning. Similarly, and more directly applicable to the current study, Ramscar (2013) exposed participants to prefix+noun+suffix strings, such that some items occurred with the same prefix while the suffix varied unpredictably in prefix condition (and the other way round for the suffix condition), and found better learning of prefix+noun and noun+picture pairs for those items which consistently occurred with the same prefix, providing corroborating evidence for an entropy-based explanation. However, both studies differed from current work in important ways discussed below.

The current study potentially provides a more direct test of whether items are better learned when they are proceeded by articles/affix, by comparing the learning of a language where they are, with one where they are not (where the affix instead follows the stem). The fact that we do *not* find this effect might therefore indicate that the hypothesis that prefixing aids

learning is wrong. However, several aspects of the current study must be considered before accepting this conclusion.

We need to consider whether the nature of our item test is suitable for capturing the effect reported in previous work (Arnon & Ramscar, 2012; Ramscar, 2013). In Arnon and Ramscar (2012), item-learning was tested by showing participants a trained picture and playing two labels, one in which the noun matched the picture, and one in which the noun was incorrect. In Experiment 1a, we used a similar test: we played a single noun+affix bigram and showed participants two pictures from training, one that matched the bigram and one that did not, but which belonged to the same category. However, Experiment 1a showed a floor effect, which is why in Experiment 1b (and subsequent experiments) we opted for a different item learning test, in which participants chose between a trained and a novel item from the same category. This test improved participants' performance, but rather than directly testing the noun/picture mappings between trained items, the test now tested participants' recognition of seen versus unseen items. However, this test is also more similar to the one used in Ramscar (2013), where participants were presented a trained picture and two labels – a correct label and a foil, and were asked whether the label was *old* or *new*. Importantly, Ramscar (2013) found a prefixing advantage whereas we did not. Therefore, it is possible that the within-category distinctions in our input set (the learning of which was necessary to complete the test) were insufficiently salient for participants to pick up on, given that all fribbles within the same category shared a cluster of visual features. It thus seems that the input structure we used was particularly conducive for grouping items together based on shared features – that is, for learning between-category distinctions, as opposed to within-category distinctions. Therefore, an input set in which individual items are different from each other enough to facilitate within-category learning may

be more appropriate for observing strong item-learning (this would, however, need to be balanced with between-group differences, which would also need to be salient enough to facilitate the learning of discriminating features). Future work will thus investigate vocabulary learning in an input set where individual items are more distinguishable, with more distinctive within-category variation. In addition, we may boost learning by removing phonological cues and using maximally discriminable artificial labels – for example: *tombat*, *deecha*, *paylig*, as opposed to: *foop*, *moob*, *kood*, and/or use visual items with more distinctive within-category variation.

Although not part of our initial predictions, one notable finding in the item test held across the two experiments – where we saw learning, it was better for LTF items than for HTF items. Importantly, this is despite the fact that *individual* items occurred equally frequently in the input (the frequency manipulation is at the type rather than token level). We suggest that this is broadly consistent with our information theoretic approach and is due to the fact that learning individual LTF-items involved lower entropy (less uncertainty) than with HTF-items, a consequence of there being fewer LTF-items. During training, in Experiment 1, once participants discriminated which category and which type of item they were presented with, for HTF-type, there were six possible items to distinguish between, whereas for LTF-type, there were only two possible items to distinguish. This may have made it harder to memorize the individual HTF items, and at test, participants were unable to distinguish them from novel items. In Experiment 2, participants were able to memorize individual items from groups of three (HTF), however, this was still poorer than LTF items, where there was only one item for each of the two types.

In addition to our item test, we included a second test of trained items but which tested knowledge of the mappings between trained stems and affixes. This test was similar to the

Semantics Phonology generalization test, but with trained rather than novel stems. Our predictions for this test were less clear, given that participants could complete it based on the association between discriminating features and affixes, in which case we predicted a similar pattern to the generalization test, or they might instead base performance on their knowledge of individual associations between particular nouns and affixes. In fact, in both experiments, unlike with novel items, we did *not* see evidence for greater type-frequency effect in the prefix condition compared to the suffix condition. This indicates that participants may assign affixes differently with trained than novel items. However, the evidence was ambiguous, both as to whether the type frequency effect was actually present/absent, and as to whether there was really a difference between trained and untrained nouns in this respect. One possibility is that participants use a combination of idiosyncratic noun-specific and general features in this test and this weakens the signature interaction seen for generalization. This is supported by the fact that individual performance in this was found to be related test both to performance in the Item Learning test and performance in the Generalization test. However, it was also the case that this test came after the generalization test and this may also have influenced performance. Given the need to avoid overinterpreting ambiguous findings, we do not speculate further.

Implications

In the current work, we based our predictions on *discriminative learning*, a theoretical framework for human communication, introduced by Ramscar and colleagues (Ramscar & Yarlett 2007; Ramscar, 2010; Ramscar et al., 2010; Ramscar et al., 2013b; Ramscar & Port, 2016), based on a computational analysis of concepts derived from classical learning theory. In its simplest form, this framework can be formalized in a computational model implementing a simplified version of the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972), which was initially proposed to account for a range of results from the field of animal learning (e.g., Kamin,

1968; Rescorla, 1967, 1968; Wagner, 1969). This approach predicts that the linear ordering of information interacts with cue structure, so that different orders can lead to different learning of discriminative cues, and we saw evidence for this prediction across two data sets, corroborating the findings in Ramscar et al (2010). We now consider how this fits with other theoretical perspectives in the literature.

First, we consider whether this same result would be predicted by other computational models. Most obviously, it is clear that other “connectionist” models, as discussed in the introduction (e.g., Elman, 1990; Rumelhart & McClelland, 1986) could in principle also capture these effects, since these models in fact implement the same error-minimization algorithm described by the simplified Rescorla-Wagner learning rule. What would be critical for capturing the difference between affix conditions is to ensure that any implementation captures the key theoretical claim that learning proceeds in a linear fashion, so that incoming linguistic features are continuously used to discriminate subsequent forms. Note that a connectionist model does not automatically necessitate a discriminative theoretical approach. As discussed in the introduction, the classic Rumelhart and McClelland (1986) model of English past tense is essentially a generative model, despite its utilization of a discriminative learning mechanism. It is also notable that, unlike the basic simple two-layer error-driven learning model employed here, most connectionist models involve at least one, and often multiple, hidden layers (e.g., Elman, 1990; Engelmann et al., 2019; Mirković et al., 2005; Rumelhart & McClelland, 1986; Seidenberg & McClelland, 1989). What the current work demonstrates is that this additional architecture may often be unnecessary when it comes to capturing core learning mechanisms at work. In a recent study, Ambridge et al. (2020) compared different approaches to modelling the learning of semantic cues to verb-argument structure using natural language corpus data. They

compared a discriminative implementation of the simplified Rescorla-Wagner rule without hidden layers (comparable to that in the current study), to ones with different numbers of additional hidden layers, and found that these did not improve the performance of a model. Therefore, while the discriminative learning observed in our study could in principle be captured by more complex connectionist architectures, we suggest that there is much to be said for the theoretical clarity that the simpler two-layer model reported in this study brings. Similarly, it may be possible to construct a model which captures the effects shown in the experiment, but where those effects are implemented at a higher level of abstraction, for example in a Bayesian model (Perfors & Wonnacott, 2011). However, for the goals of the present study, this would be less useful than the approach taken here, which provides a mechanistic account of human performance which is firmly grounded in established principles of human and animal learning. Note that this is not to say that more complex architectures may not be appropriate or required for answering questions about learning that are more fine-grained than the abstract, high-level approach to learning taken here (for example, while our simulations treat time as a discrete sequence of events, the temporal dynamics of even simple learning tasks like the ones studied here are far more subtle and complex than this, Apfelbaum & McMurray, 2017), rather, that where a simple architecture *can* capture the effect of interest, it should be preferred (see Baayen & Hendrix, 2017).

More broadly, we suggest that the current theoretical framework may provide an account of other findings in the statistical learning literature. Critically, this framework provides a precise account of what underpins “generalization” – in discriminative learning terms, this is the strength of association between informative cues and the forms which they predict; the mechanism of generalization is the strengthening of these associations at the expense of dissociating

uninformative cues, such that “generalization” in a discriminative model is simply the process of eliminating alternative outcomes. By the same token, what is sometimes viewed as item-based learning is simply a function of the degree to which less informative, idiosyncratic cues are associated with the outcome. For example, Wonnacott et al. (2012) found that when five year olds were taught a novel verb argument construction (a novel word order with a novel meaning), they had difficulty extending the construction to a novel verb if they had only encountered it in the context of a single verb in their input set. From a discriminative learning perspective, this arises because the meaning of the construction remains strongly associated with the verb in the input set, at the expense of learning the association with the novel word order. This makes it harder to extend that construction to new verbs. It is beyond the scope of the current paper to re-evaluate 20 years of experimental findings in statistical language learning, although this may be an interesting future project.

Our work also has potential implications for understanding crosslinguistic patterns. As noted in the introduction, strongly suffixing languages are more prevalent across the world’s languages, and it has been suggested that this may arise because suffixing systems are easier to learn. Our work adds to the experimental evidence demonstrating a difference in the learning of suffixing and prefixing, and thus may be seen to provide support for this explanation of crosslinguistic difference. However, we note that in some circumstances the prefix condition was just as good as (if not better than) the suffix condition. Note that this is a strength, in that it speaks against interpretations of our results as an effect of linguistic experience – it is not simply that our participants were speakers of a suffixing language and thus showed a suffixing advantage, a potential criticism that could be made of some earlier work, such as St Clair et al. (2009). On the other hand, it is clear that while in certain circumstances the cue ordering under

suffix learning is helpful for generalizing discriminative cues, it is over-simplistic to characterize this as a straightforward benefit of suffixing over prefixing. Other researchers have criticised the general assumption that cross-linguistic differences are underpinned by learner biases (Evans & Levinson, 2009; Dunn et al., 2011 for criticisms of this approach), and there may be an alternative account of suffixing preferences based on historical language change (Martin & Culbertson, 2020). Thus, though our work demonstrates how basic learning mechanisms may lead to differences in the learning of alternative linguistic forms, we hesitate to link our findings directly to the prevalence of suffixing languages.

On the other hand, although we framed our learning paradigm in terms of affix learning, we view our conclusions as potentially important for *any* aspect of language learning in which order is relevant to the way linguistic information is structured. In the domain of noun-compounds, Marelli and Luzzatti (2012) explored the processing of head-initial and head-final compounds in Italian native speakers, and found faster processing of head-final compounds (e.g., *astronave* – *spaceship*, literally: *star+ship*) compared to head-initial (e.g., *pescespada* – *swordfish*, literally: *fish+sword*). While the authors explain this in terms of the head-final structure being the default structure (and when this default assumption fails, in the case of head-initial compounds, the head and modifier roles need to be re-assigned on the fly, causing processing delays), the results may be more parsimoniously interpreted in terms of cue competition and uncertainty reduction¹⁸. Specifically, modifiers provide relational information: for example, speakers of English know that *snowball* refers to more than a ball that is somehow related to snow, but a ball that is MADE OF snow (a *teacup* is a cup USED FOR tea, etc.). The different possible relations are supported by the modifier – in this case, *snow* (e.g., a ball MADE

¹⁸ We thank Marco Marelli for this insightful suggestion.

OF snow, a ball USED FOR snow, a ball CAUSED BY snow, etc.) compete during processing (Pham & Baayen, 2013; Schmidtke et al., 2016). Critically, in head-final compounds, since competitive information is provided *first*, this allows for cue-competition to reduce uncertainty about the meaning of the compound, in contrast with head-initial compounds where the first constituent does not reduce uncertainty, leading to longer processing times.

Limitations and Future Directions

As with all experiments, there are limitations to our design. First, the artificial languages we used are not designed to directly reflect any specific natural language morphology, and the suffixing and prefixing in this study is a simplified representation. Our goal was to include the aspect of affixing whereby “stems” have features (phonological and/or semantic) which are predictive of the occurrence of more simple “affixes”. Our experiment provides an in-principle demonstration of how order affects learning in these circumstances. However, in so simplifying, we ignore details which may be highly relevant in the learning of particular inflectional systems. For example, “affixes” may have varying phonological features and (often) have associated functions. We expect that these factors could be important, depending on how they affect cue structure. We also note that our use of terms like “morpheme”, “stem”, and “affix” is for explanatory convenience, and should not be taken as support for the claim that these constructs exist as psychological “units” of representation (see e.g., Blevins, 2016; Ramscar & Port, 2016; Port & Leary, 2005; Samuel, 2020 for reviews of some of the problems with the latter). As discussed in the introduction, our aim is to provide an *in-principle* demonstration of the critical point that, in order for learning to occur, features associated with a linguistic utterance must unfold in time in a way that provides enough cue competition to produce uncertainty reduction via prediction error.

One possible concern with our paradigm is that the timing of stimuli was somewhat unnatural, as pictures were only present on-screen while stems were played, with a blank screen showing during the affix. However, this was necessary so that in the prefix condition the semantic features available of the stem were not present until the stem was actually encountered (to prevent these features from acting as “cues” for the prefix), and the timing in the suffix condition was consequently designed to be as closely matched as possible to the prefix condition. Our goal here was to mimic the situation in natural language processing where the referent of the noun is not known until that noun is heard – seeing the picture is akin to the process of accessing the noun’s semantic representation in memory. A more natural analogue of natural language learning could involve a cross-situational learning paradigm (Yu & Smith, 2007), in which multiple potential referents are visible in any given trial. In this situation, we would expect participants would gradually come to associate nouns with their referents using cross-trial co-occurrence statistics (e.g., Medina et al., 2011; Ramscar et al., 2013a; Roembke & McMurray, 2016; Smith et al., 2011; Smith & Yu, 2008; Vouloumanos & Werker, 2009; Yu & Smith, 2007). With this method, the referent could be visibly present throughout the trial, but the participant would only be able to identify it when the noun was played. We would expect that, eventually, the same pattern of learning as seen in the current experiment would emerge, although this would require more exposure, possibly over multiple sessions. This is an interesting direction for future work. In the meantime, our results provide in-principle evidence that cue competition over features of the noun can lead to different patterns of learning in suffixing compared with prefixing, where cue competition over those features is not possible.

One observation concerning the current results is that, although we have clear evidence at the group level for at least some of our predicted effects, there is wide variance of performance

in each of the test tasks: our experience is that this is common for artificial language paradigms, although the still relatively common practice of using bar charts (in contrast to our own plots which capture individual differences) can obfuscate this. One reason is that individuals vary greatly in their ability to attend to and process the limited input in these experiments, which is an interesting topic beyond the scope of the current paper. Another reason is that we generally used tests not well designed in terms of reliability and within-participant consistency, due to having few trials per test-type. While in principle we could design longer tests (which we did with some tests here in the move to from Version A to Version B), this has to be balanced against the possibility that participants learn from the test exposures themselves, including incorrect associations with the “foil” affix. We do acknowledge that our results are much noisier than in the similar experiment in Ramscar et al. (2010). Since the paradigms are not identical (we use audio stimuli, and items-labels as well as category-labels but have fewer categories), it is not possible to compare the results directly. However, one possibility is that the move to online testing may have introduced noise (in Ramscar et al., 2010, participants were tested on the same PC, sitting the same distance from the screen, with the same headphones, and were informed they were being video recorded, such that they may have been more likely to attend to the stimuli in the lab context, etc.). In ongoing work, we are exploring measures which can mitigate against noise in online experiments, such as including attention checks within training.

Despite the extent of individual differences, we emphasize that we *do* find substantial evidence, at the group level, for the key interaction between affix and type-frequency in both experiments. We also found evidence for the null for an overall benefit of suffixing in the Semantics and Phonology Generalization test at least in in Experiment 1. In fact, the only key hypotheses where we are left with ambiguous evidence were the benefits of prefixing for item-

learning (which was ambiguous in Experiment 2, the study where we thought we had best chance of seeing the effect), and the benefit of suffixing in the Phonology Generalization test (ambiguous in both experiments). While we could in principle increase our participant sample until we have evidence for H1/H0 for these hypotheses (see footnote 7), we have discussed above why in retrospect we think the current paradigm is not well-designed to provide answers here (i.e., poor discriminability between individual items in the same category in the Item Learning test; floor effects in Phonology test). Importantly, basing our inferences on Bayes Factors (rather than p values) allows us to differentiate the situation where the data is insensitive from the situation where we actually have evidence for the null.

Another limitation of this work is that our test of learning is an “offline” method (2AFC post-exposure tests). Although the generalization results suggest that we *did* capture implicit statistical learning with this test task (unlike in Tamminen et al., 2015), we acknowledge that the method is crude since it does not provide us with information about learning as it unfolded in time. Thus our evidence that learning proceeds via cue competition and prediction error is indirect. A goal for future work is to adapt the existing paradigm to use online measures. One possibility is to use pupillometry, which has been used as a measure of surprise and cognitive load (e.g., Ben-Nun, 1986; Borghini & Hazan, 2018; Hyona et al., 1995; Just & Carpenter, 1993). Our prediction would be that, if our hypothesis is correct and prediction error is the driving mechanism for learning in the suffix condition, we should expect greater pupil dilation on LTF trials, which generate greater error signal. Developing more implicit measures may also allow us to adapt the current paradigm for use with child learners. This is important since ultimately our interest is in first language acquisition, and given evidence that adult and child

language learning may be different (Brown et al., 2018; Culbertson & Newport, 2015; Culbertson et al., 2019; Hudson Kam & Newport, 2005).

Conclusion

This study tested the discriminative learning framework (as introduced in Ramscar & Yarlett, 2007; Ramscar et al., 2010), which frames language learning in terms of classical learning theory, where learning is a process of uncertainty reduction by discriminating informative from uninformative cues. The critical learning mechanisms are cue competition and prediction error, and these are facilitated when more complex cues precede simpler cues in learning, such as stems preceding affixes (suffixing), compared with when simpler cues come first, such as with prefixing. Prefixes, on the other hand, may facilitate processing by reducing the uncertainty of the upcoming noun-stem. From this, we predicted better generalization in the suffix condition than in the prefix condition (and demonstrated this computationally), and better item-learning in the prefix condition. We tested this in two artificial language learning experiments, and as predicted, found that when generalization required discriminating between high-frequency uninformative cues and low-frequency informative cues, this was better in the suffix condition (affix-by-type-frequency interaction). With respect to item-learning, contrary to prediction, we did not find evidence that this was better in the prefix condition, and we identify some aspects of our paradigm which may have, in retrospect, been unsuitable for testing this hypothesis, and propose potential alternatives.

Acknowledgements

While this manuscript was prepared, the first author was supported by an Economic and Social Research Council PhD Studentship, and the final author was supported by an Economic

and Social Research Council grant. We thank two anonymous reviewers and Marco Marelli for helpful revisions, as well as Ben Davidson and Matthew Jaquiere for feedback.

References

- Ambridge, B., Tatsumi, T., Doherty, L., Maitreyee, R., Bannard, C., Samanta, S., ... & Efrati, A. (2020). The crosslinguistic acquisition of sentence structure: Computational modelling and grammaticality judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and K'iche'. *Cognition*, 202, 104310.
<https://doi.org/10.1016/j.cognition.2020.104310>
- Anwyl-Irvine, A. L., Massonnie, J., Flitton, A., Kirkham, N., Evershed, J. K. (2019). Gorilla in our midst: An online behavioural experiment builder. *Behavior Research Methods*. 1-20.
<https://doi.org/10.3758/s13428-019-01237-x>
- Apfelbaum, K. S., & McMurray, B. (2017). Learning During Processing: Word Learning Doesn't Wait for Word Recognition to Finish. *Cognitive Science*, 41(S4), 706-747.
<https://doi.org/10.1111/cogs.12401>
- Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition*, 122(3), 292-305.
<https://doi.org/10.1016/j.cognition.2011.10.009>
- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1), 31-56.
<https://doi.org/10.1177/00238309040470010201>
- Baayen, R. H., Davidson, D. J., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412. <https://doi.org/10.1016/j.jml.2007.12.005>

- Baayen, R. H., & Hendrix, P. (2017). Two-layer networks, non-linear separation, and human learning. *From Semantics to Dialectometry. Festschrift in honor of John Nerbonne. Tributes*, 32, 13-22.
- Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naïve discriminative learning. *Psychological Review*, 118(3), 438.
<https://psycnet.apa.org/doi/10.1037/a0023851>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Machler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1-48.
<https://doi.org/10.18637/jss.v067.i01>
- Ben-Nun, Y. (1986). The use of pupillometry in the study of on-line verbal processing: Evidence for depths of processing. *Brain and Language*, 28(1), 1-11.
[https://doi.org/10.1016/0093-934X\(86\)90086-6](https://doi.org/10.1016/0093-934X(86)90086-6)
- Blevins, J. P. (2016). *Word and Paradigm Morphology*. Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780199668984.013.22>
- Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: a pupillometry study. *Frontiers in Neuroscience*, 12, 152.
<https://doi.org/10.3389/fnins.2018.00152>

Boston, M. F., Hale, J. T., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3), 301-349.

<https://doi.org/10.1080/01690965.2010.492228>

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience.

Psychological Bulletin, 138(3), 389. <https://doi.org/10.1037/a0026450>

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2018). Brown, Smith, Samara, & Wonnacott (pre-print). Semantic cues in language learning: An artificial language study with adult and child learners. PsyArXiv. <https://doi.org/10.31234/osf.io/7hq2c>

Casenhiser, D., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science*, 8(6), 500-508.

<https://doi.org/10.1111/j.1467-7687.2005.00441.x>

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge University Press.

Clark, E. (2007). Morphology in language acquisition. In A. Spencer and A. M. Zwicky (Eds). *The Handbook of Morphology*. Oxford: Blackwell. 374-389.

<https://doi.org/10.1002/9781405166348.ch19>

Culbertson, J., Gagliardi, A., & Smith, K. (2017). Competition between phonological and semantic cues in noun class learning. *Journal of Memory and Language*, 92, 343-358.

<https://doi.org/10.1016/j.jml.2016.08.001>

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: evidence for a phonological bias. *Language*, 95, 268-293. <https://doi.org/10.1353/lan.2019.0031>

- Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, 139, 71-82.
<https://doi.org/10.1016/j.cognition.2015.02.007>
- Dahan, D., Swingle, D., Tanenhaus, M. K., & Magnuson, J. S. (2000). Linguistic gender and spoken-word recognition in French. *Journal of memory and Language*, 42(4), 465-480.
<https://doi.org/10.1006/jmla.1999.2688>
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, 5, 781. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78–89. <https://doi.org/10.1016/j.jmp.2015.10.003>
- Dryer, M. S. (2013). Prefixing vs. suffixing in inflectional morphology. In M. S. Dryer & M. Haspelmath (Eds.) *The World Atlas of Linguistic Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <http://wals.info/chapter/26>.
- Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345), 79–82.
<https://doi.org/10.1038/nature09923>
- Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non-commercial purposes. In *Proceedings of the Fourth International Conference on Spoken Language*, 3, 1393-1396. <https://doi.org/10.1109/ICSLP.1996.607874>

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms.

In F. Kiefer, J. P. Blevins, & H. Batros (Eds.) *Perspectives on Morphological Structure:*

Data and Analyses (pp. 212-239). Brill. https://doi.org/10.1163/9789004342934_011

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2018). Alternative solutions to a language design

problem: The role of adjectives and gender marking in efficient communication. *Topics*

in Cognitive Science, 10(1): 209-224. <https://doi.org/10.1111/tops.12316>

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179-211.

[https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E)

Engelmann, F., Granlund, S., Kolak, J., Szreder, M., Ambridge, B., Pine, J., ... & Lieven, E.

(2019). How the input shapes the acquisition of verb morphology: Elicited production

and computational modelling in two highly inflected languages. *Cognitive*

Psychology, 110, 30-69. <https://doi.org/10.1016/j.cogpsych.2019.02.001>

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and

its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5), 429– 448.

<https://doi.org/10.1017/s0140525x0999094x>

Finch, S., & Chater, N. (1992). Unsupervised methods for finding linguistic categories. In I.

Aleksander, & J. Taylor (Eds.), *Artificial Neural Networks, Proceedings of the 1992*

International Conference on Artificial Neural Networks (ICANN-92) (pp. 1365-1368).

North-Holland. <https://doi.org/10.1016/B978-0-444-89488-5.50115-9>

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence

comprehension. *Topics in Cognitive Science*, 5(3), 475-494.

<https://doi.org/10.1111/tops.12025>

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1-11.

<https://doi.org/10.1016/j.bandl.2014.10.006>

Gallistel, C. R. (2002). Frequency, contingency and the information processing theory of conditioning. In P. Sedlmeier & T. Betsch (Eds.), *ETC. Frequency processing and Cognition* (pp. 153-171). Oxford University Press.

<https://doi.org/10.1093/acprof:oso/9780198508632.003.0010>

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7), 1079

1088. <https://doi.org/10.1177/0956797612463705>

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Language*, 2, 73-113.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, 32(2), 101-123. <https://doi.org/10.1023/A:1022492123056>

Hebb, D. O. (1949). *The organization of behavior: a neuropsychological theory*. Wiley.

Hendrix, P., Ramscar, M., & Baayen, H. (2019). NDRA: A single route model of response times in the reading aloud task based on discriminative learning. *PloS one*, 14(7), e0218802.

<https://doi.org/10.1371/journal.pone.0218802>

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2), 151-195. <https://doi.org/10.1080/15475441.2005.9684215>

- Hupp, J. M., Sloutsky, V. M., & Culicover, P. W. (2009). Evidence for a domain-general mechanism underlying the suffixation preference in language. *Language and Cognitive Processes*, 24, 876–909. <https://doi.org/10.1080/01690960902719267>
- Hyönä, J., Tommola, J., & Alaja, A. M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, 48(3), 598–612. <https://doi.org/10.1080/14640749508401407>
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62. <https://doi.org/10.1016/j.cogpsych.2010.02.002>
- Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.
- Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4): 169–88. <https://doi.org/10.1017/s0140525x10003134>
- Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 47(2), 310. <https://dx.doi.org/10.1016%2Fj.cogpsych.2010.02.002>
- Kamin, L. J. (1968). Attention-like processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation*. Miami University Press.
- Kopp, B., & Wolff, M. (2000). Brain mechanisms of selective learning: Event-related potentials provide evidence for error-driven learning in humans. *Biological Psychology*, 51(2–3), 223–246. [https://doi.org/10.1016/S0301-0511\(99\)00039-3](https://doi.org/10.1016/S0301-0511(99)00039-3)

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177.

<https://doi.org/10.1016/j.cognition.2007.05.006>

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, 18(3), 193-198.

<https://doi.org/10.1111%2Fj.1467-9280.2007.01871.x>

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language*, 63(4), 447

464. <https://doi.org/10.1016/j.jml.2010.07.003>

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Courier Corporation.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and language*, 28(3), 255–277.

[https://doi.org/10.1016/0749-596X\(89\)90033-8](https://doi.org/10.1016/0749-596X(89)90033-8)

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24(12), 2351–2360.

<https://doi.org/10.1177/0956797613495418>

Marelli, M., & Luzzatti, C. (2012). Frequency effects in the processing of Italian nominal compounds: Modulation of headedness and semantic transparency. *Journal of Memory*

and Language, 66(4), 644-664. <https://doi.org/10.1016/j.jml.2012.01.003>

Martin, A., & Culbertson, J. (2020). Revisiting the Suffixing Preference: Native-Language Affixation Patterns Influence Perception of Sequences. *Psychological Science*,

0956797620931108. <https://doi.org/10.1177%2F0956797620931108>

Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? A uniform information density account. In J. Lafferty, C. Williams, J. Shawe

- Taylor, R. Zemel & A. Culotta (Eds.) *Advances in Neural Information Processing Systems* (pp. 1585-1593). Curran Associates, Inc.
- Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22), 9014-9019. <https://doi.org/10.1073/pnas.1105040108>
- Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS one*, 12(2), e0171935.
<https://doi.org/10.1371/journal.pone.0171935>
- Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive science*, 26(4), 393-424.
https://doi.org/10.1207/s15516709cog2604_1
- Mirković, J., MacDonald, M. C., & Seidenberg, M. S. (2005). Where does gender come from? evidence from a complex inflectional system. *Language and Cognitive Processes*, 20(12), 139–167. <https://dx.doi.org/10.1080/02701690960444000205>
- Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In S. Becker, S. Thrun & K. Obermayer (Eds). *Advances in Neural Information Processing Systems* (pp. 841-848).
- Perfors, A., Tenenbaum, J., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, 37, 607-642.
<https://doi.org/10.1017/S0305000910000012>
- Perfors, A. & Wonnacott, E. (2011) Bayesian modeling of sources of constraint in language acquisition. In: Arnon, I. and Clarke, E. V., (Eds.) *Experience, Variation and Generalization : Learning a First Language* (pp. 277-294). John Benjamins.

Pham, H., & Baayen, H. R. (2013). Semantic relations and compound transparency: A regression study in CARIN theory. *Psihologija*, 46(4), 455-478.

<https://doi.org/10.2298/PSI1304455P>

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526-3529.

<https://doi.org/10.1073/pnas.1012551108>

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, 62(2-4), 146-159.

<https://doi.org/10.1159/000090095>

Port, R. F., & Leary, A. P. (2005). Against formal phonology. *Language*, 81(4), 927-964.

<https://doi.org/10.1353/LAN.2005.0195>

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06*, University of Cambridge, Cambridge, 26-46.

R Core Development Team. (2013). R: A language and environment for statistical computing.

Ramscar, M. (2010). Computing machinery and understanding. *Cognitive Science*, 34(6), 966.

<https://doi.org/10.1111/j.1551-6709.2010.01120.x>

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, 46(4), 377-396. <https://doi.org/10.2298/PSI1304377R>

Ramscar, M., Dye, M., & Klein, J. (2013a). Children value informativity over logic in word learning. *Psychological Science*, 24(6), 1017-1023.

<https://doi.org/10.1177/0956797612460691>

- Ramscar, M., Dye, M., & McCauley, S. M. (2013b). Error and expectation in language learning: The curious absence of mice in adult speech. *Language*, 89(4), 760-793.
<https://doi.org/10.1353/lan.2013.0068>
- Ramscar, M., & Dye, M. (2011). Learning language from the input: Why innate constraints can't explain noun compounding. *Cognitive Psychology*, 62(1), 1-40.
<https://doi.org/10.1016/j.cogpsych.2010.10.001>
- Ramscar, M., & Port, R. (2015). Categorization (without categories). In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 75–99). De Gruyter.
<https://doi.org/10.1515/9783110626476-005>
- Ramscar, M., & Port, R. F. (2016). How spoken languages work in the absence of an inventory of discrete units. *Language Sciences*, 53: 58-74.
<https://doi.org/10.1016/j.langsci.2015.08.002>
- Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, 31(6), 927-960. <https://doi.org/10.1080/03640210701703576>
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label order and their implications for symbolic learning. *Cognitive Science*, 34(6), 909–957.
<https://doi.org/10.1111/j.1551-6709.2009.01092.x>
- Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, 66(1), 30–54.
<https://dx.doi.org/10.1016%2Fj.cogpsych.2012.09.001>

- Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, 74(1), 71. <https://doi.org/10.1037/h0024109>
- Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, 66(1), 1. <https://psycnet.apa.org/doi/10.1037/h0025984>
- Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American Psychologist*, 43(3), 151. <https://psycnet.apa.org/doi/10.1037/0003-066X.43.3.151>
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current Research and Theory*, 2, 64-99.
- Roembke, T. C., & McMurray, B. (2016). Observational word learning: Beyond propose-but verify and associative bean counting. *Journal of Memory and Language*, 87, 105-127. <https://doi.org/10.1016/j.jml.2015.09.005>
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.
- Rumelhart, D. E., & McClelland, J. L. (1986). Parallel distributed processing: explorations in the microstructure of cognition. Volume 1: Foundations.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926-1928. <https://doi.org/10.1126/science.274.5294.1926>
- Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, 111, 104070. <https://doi.org/10.1016/j.jml.2019.104070>

- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96(4), 523.
<https://doi.org/10.1037/0033-295x.96.4.523>
- Schmidtke, D., Kuperman, V., Gagné, C. L., & Spalding, T. L. (2016). Competition between conceptual relations affects compound recognition: the role of entropy. *Psychonomic Bulletin & Review*, 23(2), 556-570. <https://doi.org/10.3758/s13423-015-0926-0>
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379-423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Smith, K., Smith, A. D., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, 35(3), 480-498.
<https://doi.org/10.1111/j.1551-6709.2010.01158.x>
- Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558-1568. <https://doi.org/10.1016/j.cognition.2007.06.010>
- St. Clair, M., Monahan, P., & Ramscar, M. (2009) Relationships between language structure and language learning: the suffixing preference and grammatical categorization. *Cognitive Science*, 33(7), 1317–1329. <https://doi.org/10.1111/j.1551-6709.2009.01065.x>
- Stone, G. O. (1987). An analysis of the delta rule and the learning of statistical associations. In J. A. Feldman, P. Hayes, & D. E. Rumelhart (Eds.), *Parallel distributed processing* (Vol. 1, pp. 444–459). MIT Press.
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, 88(2), 135.
<https://doi.org/10.1037/0033-295X.88.2.135>

- Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, 79, 1-39.
<https://doi.org/10.1016/j.cogpsych.2015.03.003>
- Van Son, R. J., & Van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, 47(1-2), 100-123. <https://doi.org/10.1016/j.specom.2005.06.005>
- Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, 45(6), 1611. <https://doi.org/10.1037/a0016134>
- Wagner, A. R. (1969). Stimulus selection and a "modified continuity theory". *The Psychology of Learning and Motivation*, 3, 1-41. [https://doi.org/10.1016/S0079-7421\(08\)60396-5](https://doi.org/10.1016/S0079-7421(08)60396-5)
- Wagner, A. R. (2008). Evolution of an elemental theory of Pavlovian conditioning. *Learning & Behavior*, 36(3), 253-265. <https://doi.org/10.3758/LB.36.3.253>
- Widrow, G., & Hoff, M. E. (1960). Adaptive switching circuits. In Institute of Radio Engineers, Western Electronic Show and Convention, part 4 (pp. 96-104).
- Wonnacott, E. (2011) Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, 65, 1-14.
<https://doi.org/10.1016/j.jml.2011.03.001>
- Wonnacott, E., Boyd, J.K, Thomson, J.J., & Goldberg, A.E. (2012). Verb argument structure. Input effects on the acquisition of a novel phrasal construction in 5-year olds. *Journal of Memory and Language*, 66, 458-478.
<https://doi.org/10.1016/j.jml.2011.11.004>
- Wonnacott, E., Brown, H., & Nation, K. (2017). Skewing the evidence: The effect of input structure on child and adult learning of lexically based patterns in an artificial

language. *Journal of Memory and Language*, 95, 36-48.

<https://doi.org/10.1016/j.jml.2017.01.005>

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, 10(3), 288-297.

<https://doi.org/10.1111/j.1467-7687.2007.00590.x>

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414-420.

<https://doi.org/10.1111/j.1467-9280.2007.01915.x>

Zubin, D. A., & Köpcke, K. (1981). Gender: A less than arbitrary grammatical category. In *Papers from the Seventeenth Regional Meeting of the Chicago Linguistic Society*, 3. 437.

Figure 1.

Feature Vector Representations of the HTF and LTF Exemplars for Each Affix.

		HTF	LTF	HTF	LTF

Non discriminating features	Shape1	1	0	0	1
	Shape2	0	1	1	0
Discriminating features	Discrim1	1	0	0	0
	Discrim2	0	1	0	0
	Discrim3	0	0	1	0
	Discrim4	0	0	0	1
	Vowel1	1	1	0	0
	Vowel2	0	0	1	1
Affixes	Affix1	1	1	0	0
	Affix2	0	0	1	1

Note. The input matrix also consisted of idiosyncratic features which are not presented for simplicity (two features per item: total of 32). In the suffix simulation, the discriminating and non-discriminating features were cues and the affixes were outcomes, whereas in the prefix simulation it was the other way round: the affixes were cues and the features were outcomes.

Figure 2.

Associative Strength Between Affixes and Features Over Time.

Note. Associative strength between affix1 and key features of interest: *shape1* (HTF non-discriminating feature), *shape2* (LTF non-discriminating feature), *discriminating_feature_1* (HTF discriminating feature), *discriminating_feature_2* (LTF discriminating feature), *vowel1* (category1 discriminating feature) and *vowel2* (category2 discriminating feature). In the prefix condition, in order to avoid overplotting, a small number (0.03) was added to the weights of non-discriminating features *shape1* and *shape2* (black solid and black dash-dotted line, respectively), but in reality these are the same as the weights for the HTF and LTF discriminating features (dark grey solid line and dark grey dash-dotted line, respectively).

Figure 3.

Probability of Choosing the Correct Affix in the Simulations.

Note. The probability of the correct affix (normalized from raw weights using the choice axiom) for the HTF features of a randomly chosen affix category (*shape1*, *disrcriminating_features_1*, *vowel1*) (dark grey), and for LTF features (*shape2*, *discriminating_features_2*, *vowel2*) (white striped). The dashed line represents chance-level (0.5).

Figure 4.

Sample Training Set.

	Category 1: ge			Category 2: ma		
75%
	/mi:b gɛ/	/θi:p gɛ/	/si:g gɛ/	/ʃu:p mʌ/	/fu:g mʌ/	/θu:g mʌ/

	/ti:p gɛ/	/dʒi:d gɛ/	/fi:b gɛ/	/lu:d mʌ/	/tu:b mʌ/	/su:d mʌ/
25%
	/ʃi:d gɛ/	/li:b gɛ/		/dʒu:b mʌ/	/mu:p mʌ/	

Note. Images courtesy of Michael Tarr's lab at Carnegie Mellon University

(http://wiki.cmbc.cmu.edu/Novel_Objects). Within each category, nouns were assigned to pictures randomly on a participant-by-participant basis.

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

Figure 5.

Schematic Representation of a Single Training Trial in Each Condition.

Note. Labels were only auditory and were not presented orthographically.

Figure 6.

Performance on the Item Learning Test

Note. Each point is each participants' average accuracy on HTF items (black) and LTF items (grey). The violins plot kernel density, and the boxes represent 95% confidence intervals around the mean (horizontal lines). Dashed lines are chance-level performance (0.5).

Figure 7.

Performance on the Semantics and Phonology Generalization Test.

Note. Each point is each participants' average accuracy on HTF items (black) and LTF items (grey). The violins plot kernel density, and the boxes represent 95% confidence intervals around the mean (horizontal lines). Dashed lines are chance-level performance (0.5).

Figure 8.

Performance on the Trained Items Test.

Note. Each point is each participants' average accuracy on HTF items (black) and LTF items (grey). The violins plot kernel density, and the boxes represent 95% confidence intervals around the mean (horizontal lines). Dashed lines are chance-level performance (0.5). There are unequal numbers of trials contributing to proportions across the different frequency types.

Figure 9.

Performance on the Phonology Generalization Test.

Note. Each point is each participants' average accuracy on HTF items (black) and LTF items (grey). The violins plot kernel density, and the boxes represent 95% confidence intervals around the mean (horizontal lines). Dashed lines are chance-level performance (0.5).

Figure 10.

Sample Training Set.

	Category 1: ge			Category 2: ma		
75%
	/mi:b gɛ/	/θi:p gɛ/	/si:g gɛ/	/ʃu:p mʌ/	/fu:g mʌ/	/θu:g mʌ/
25%
	/ʃi:d gɛ/			/dʒu:b mʌ/		

Note. Images courtesy of Michael Tarr's lab at Carnegie Mellon University

(http://wiki.cmbc.cmu.edu/Novel_Objects). Within each category, nouns were assigned to pictures randomly on a participant-by-participant basis.

Appendix

Appendix A: Language Awareness Questionnaire

Below is the Language Awareness Questionnaire for the suffix condition. In the prefix condition, the questions were exactly the same except the wording was adjusted where necessary (prefix instead of suffix).

=====

Language Awareness Questionnaire

Please respond to all of the questions. If you answer YES to any of the questions, please provide a brief explanation. If you answer NO, please type in "none".

The names of the aliens in the language you have been learning consisted of two parts: the noun and the suffix. For example, you may have heard something like: *feep ge* (where *feep* is the noun and *ge* is the suffix), *kood da*, *jeeb ma*, *foog pe*, and so on. (Don't worry if you don't recognise any of these examples – each participant gets a slightly different version of the language, but the logic is the same!)

Do you think that the suffixes meant anything?

If YES, please specify. If NO, please write "none".

Did you notice any patterns in how the suffixes were used in relation to the nouns (i.e., the parts of the alien name that came before the suffix)?

If YES, please specify. If NO, please write "none".

If you did notice any patterns, how early in the experiment do you think you noticed them?

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

Were there any other patterns that you were considering along the way that turned out to be incorrect?

If YES please specify, if NO please write "none".

=====

Participants' responses were coded such that any response of the form "items with X went with one affix, items with Y went with the other" (this theoretically could be either the visual characteristics of the fribbles or the vowels in the nouns, or both; however no participants reported awareness of both) was coded as explicit awareness. These were divided into sub-groups based on what specific cue participants reported. Participants whose responses were not specific enough (e.g., "I knew there was a pattern but I couldn't put my finger on it", "it was based on the appearance of the alien", "different sounds went with different affixes") or described some other pattern (e.g., "position of the tail", "male or female") were coded as "other". Most participants reported noticing no patterns as to the usage of the affixes, while some reported the usage being conditioned on the shape and/or the colour of the fribbles, or the vowel in the nouns. No participants reported explicit awareness of the correct discriminating features. Participants were explicitly asked whether, if they thought they had noticed a rule, they also noticed exceptions to the rule. The results are given below:

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

Table 1.

Language Awareness Questionnaire Responses Experiment 1A

	Number of participants reporting explicit awareness			
	Shape and/or colour of the fribbles	Vowel in the nouns	None	Other
Prefix	11 (26.19%)	3 (7.14%)	24 (57.14%)	4 (9.52%)
Suffix	7 (16.67%)	1 (2.38%)	30 (71.43%)	4 (9.52%)

Table 2.

Language Awareness Questionnaire Responses Experiment 1B

	Number of participants reporting explicit awareness				
	Noticed exceptions	Shape and/or colour of the fribbles	Vowel in the nouns	Other	None
Prefix	4 (6.67%)	8 (13.33%)	1 (1.67%)	10 (16.67%)	37 (61.67%)
Suffix	4 (6.67%)	8 (13.33%)	1 (6.78%)	7 (11.67%)	40 (66.67%)

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

Table 3.

Language Awareness Questionnaire Results Experiment 2

Number of participants reporting explicit awareness						
	Noticed exceptions	Shape and/or colour of fribbles	Vowels in the nouns	Vowels and shapes of fribbles	Other	None
Prefix	8 (6.9%)	26 (22.41%)	2 (1.72%)	2 (1.72%)	17 (14.65%)	59 (50.86%)
Suffix	2 (1.75%)	24 (21.05%)	3 (2.63%)	0	17 (14.91%)	67 (58.77%)

Appendix B: Approaches to determining values to inform the H1

Wherever possible, we used independent data to inform the H1 (methods A—D); where using independent data was not possible, we used the approach of determining a plausible maximum, and using half of this value as the estimate for H1 (methods B—D), since the maximum for the half normal is equal to 2SD. The approaches are outlined below.

(A) Value for H1 comes from the estimate of the same effect with independent data

This approach was used wherever possible – this could be from a different experiment or from a different condition from the same experiment. For example, if we saw evidence for an effect in one experiment, we used that value as the H1 for the same or similar effect in another experiment; or if we saw evidence for an effect in one condition (e.g., prefix), we used that value as the H1 estimate for the other condition (e.g., suffix).

(B) Value for H1 for a main effect is based on the intercept estimate (from independent or current data)

Assumption for plausible maximum difference between the two conditions: participants in a_1 perform greater than chance and participants in a_2 show chance-level performance.

The main effect of a is given as follows:

$$a = (a_1 - \log(chance)) - (a_2 - \log(chance))$$

Applying our assumption for a plausible maximum to (1) therefore gives:

$$a = (a_1 - \log(0.5)) - (a_2 - \log(0.5))$$

$$a = (a_1 - 0) - (0 - 0)$$

$$a = a_1$$

The grand mean is given as follows:

$$\bar{e} = \frac{(a_1 - \log(0.5)) + (a_2 - \log(0.5))}{2}$$

$$2\bar{e} = a_1$$

$$2\bar{e} = a$$

Therefore, the main effect of a is equal to twice the grand mean ($2\bar{e}$). We therefore set x (the estimate for a) to half this value ($\bar{e}$), since the maximum for the half normal is equal to 2SD. As we use centred coding, this means that the intercept term corresponds to $\bar{e}$. Therefore, we use the intercept estimate as the value of x for a main effect a .

(C) Value for H1 for an interaction is based on a main effect estimate (from independent or current data)

Assumption for plausible maximum interaction $a.b$ between two main effects a and b with two levels each: the effect of b only occurs for one level of a , whereas the effect of b on the other level of a equals to zero:

$$a_1.b_1 - a_1.b_2 \neq 0$$

$$a_2.b_1 - a_2.b_2 = 0$$

The interaction $a.b$ is given as follows:

$$a.b = (a_1.b_1 - a_1.b_2) - (a_2.b_1 - a_2.b_2)$$

Applying our assumptions for a plausible maximum therefore gives:

$$a.b = (a_1.b_1 - a_1.b_2) - 0$$

$$a.b = (a_1.b_1 - a_1.b_2)$$

The main effect of a is therefore given by:

$$a = \frac{(a_1.b_1 - a_1.b_2) + (a_2.b_1 - a_2.b_2)}{2}$$

$$2a = (a_1.b_1 - a_1.b_2) + (a_2.b_1 - a_2.b_2)$$

$$2a = a.b + 0$$

$$2a = a.b$$

The interaction $a.b$ is thus equal to twice the main effect of a ($2a$). We therefore set x (the estimate for $a.b$) to half this value (a), since the maximum for the half normal is equal to 2SD.

Value for H1 for a three-way interaction is based on the estimate for a two-way interaction (from independent or current data)

Assumption for plausible maximum interaction $a.b.c$ between three main effects a , b and c with two levels each: the interaction $a.b$ only occurs for one level of c , whereas the interaction $a.b$ at the other level of c equals to zero:

$$a_1.b_1.c_1 - a_1.b_2.c_1 \neq 0$$

$$a_2.b_1.c_2 - a_2.b_2.c_2 = 0$$

The interaction $a.b.c$ is given as follows:

$$a.b.c = ((a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1)) - ((a_1.b_1.c_2 - a_1.b_2.c_2) - (a_2.b_1.c_2 - a_2.b_2.c_2))$$

Applying our assumptions for a plausible maximum therefore gives:

$$a.b.c = (a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1) - 0$$

$$a.b.c = (a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1)$$

The interaction $a.b$ is therefore given as:

$a.b$

$$= \frac{((a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1)) + ((a_1.b_1.c_2 - a_1.b_2.c_2) + (a_2.b_1.c_2 - a_2.b_2.c_2))}{2}$$

$$2a.b = ((a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1)) + 0$$

$$2a.b = (a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1)$$

$$2\ a.b = a.b.c$$

The interaction $a.b.c$ is thus equal to twice the $a.b$ interaction ($2a.b$). We therefore set x (the estimate of H1 for $a.b.c$) to half this value ($a.b$), since the maximum for the half normal is equal to 2SD.

Appendix C: Bayes Factor analyses

Test	Hypothesis	H1	Justification
Item Learning (Exp1b)	Prefix better than Suffix	Intercept from the same lme	B
	LF better than HF	Intercept from the same lme	B
	Prefix HF above chance	Prefix LF Intercept from Exp 2	A
	Prefix LF above chance	Prefix LF Intercept from Exp 2	A
	Suffix HF above chance	Prefix LF Intercept from Exp 2	A
	Suffix LF above chance	Prefix LF Intercept from Exp 2	A
Semantics & Phonology Generalization (Exp1)	Suffix better than Prefix	Intercept from Exp1	B
	Greater effect of type-frequency in Prefix than Suffix	Affix-by-type-freq. interaction	A
	HF better than LF in Prefix	Type-frequency effect in Suffix	A
	HF better than LF in Suffix	Type-frequency effect in Prefix	A
Semantics & Phonology Trained Items and Generalization (Exp1b)	Better performance on trained compared to novel items	Main effect of item-novelty from Exp2	A

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

	Greater affix-by-type-frequency interaction with novel compared to trained items	Affix-by-type-freq. interaction	D
Phonology Generalization (Exp1)	Suffix better than Prefix	Intercept from the same lme	B
	Prefix above chance	Suffix Intercept	A
	Suffix above chance	Prefix Intercept	A
Item learning (Exp2)	Prefix better than Suffix	Intercept from same lme	B
	LF better than HF	Intercept from same lme	B
	Prefix HF above chance	Prefix LF Intercept from Exp1b	A
	Suffix HF above chance	Prefix LF Intercept from Exp1b	A
	Suffix LF above chance	Prefix LF Intercept from Exp1b	A
Semantics & Phonology Generalization (Exp2)	Suffix better than Prefix	Intercept from Exp1 A	B
	Greater effect of type-frequency in Prefix than Suffix	Affix-by-type-freq. interaction from Exp1 A	A
	HF better than LF in Prefix	Type-frequency effect in Prefix from Exp1 A	A
	HF better than LF in Suffix	Type-frequency effect in Prefix from Exp1 A	A

LANGUAGE LEARNING AS UNCERTAINTY REDUCTION

Semantics & Phonology Generalization and Trained Items (Exp2)	Trained items better than novel items	Main effect of item-novelty from Exp1	A
	Greater affix-by-type-frequency interaction with novel compared to trained items	Affix-by-type-freq. interaction	D
Phonology Generalization (Exp2)	Suffix better than Prefix	Intercept from Exp1	B
	Prefix above chance	Suffix Intercept from Exp1	A
	Suffix above chance	Suffix Intercept from Exp1	A