Exploring language learning as uncertainty reduction using artificial language learning

Maša Vujović

A thesis submitted for the degree of
Doctor of Philosophy

University College London

2020

Declaration

I, Maša Vujović, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Acknowledgements

I would like to thank my supervisor, Liz Wonnacott, for timely and thorough feedback, for persistent enthusiasm for this work, which meant a lot to me in ever frequent times of doubt, and for her completely unfounded faith in me. I thank my collaborator-turned-supervisor, Michael Ramscar, for all the hours spent discussing data and working predictions out in front of the whiteboard, and for his ideas which have fundamentally challenged the way I think about language.

This work was funded by a 1+3 Advanced Quantitative Methods studentship from the Economic and Social Science Research Council, for which I am grateful. I quite literally could not have done this without them.

Language and Cognition Department at University College London is the most supportive, stimulating, friendly environment a PhD student could ask for. On that front, I want to thank: Vanessa Meitanis, for countless coffee breaks, elaborate lunches, and some spectacular squash rallies; Lena Blott, for all the chats and laughs; Anna Volkmer, for a legendary birthday cake; Daniela Singh, Jo Saul, and Claire Murphy, for going through PhD milestones with support and trips to the pub; and to Gwen Brekelmans, Ria Bernard, Claudia Bruns, and other residents of 202 for making it such a great place to work in. Thank you Merle Mahon and Richard Jardine, for seamlessly making sure we were all sane and reasonably happy. Thank you also to Caroline Newton, whose kindness, wit, and understanding have made the last year of my PhD considerably more bearable.

Language Learning Lab – Liz Wonnacott, Anna Samara, Cat Silvey, Gwen Brekelmans, Daniela Singh, and Hanyu Dong – have fostered a culture of sharing, academic honesty and scientific rigour, which has been a valuable source of support over the years. Thank you also to Courtenay Norbury's lab for the joint journal clubs and lunches.

While I wrote this thesis myself, it would be fair to say that bound between these covers would be two-hundred blank pages had it not been for an entire village to make this endeavour worthwhile. In my case it's more of a global village scattered across space – Montenegro, London, Edinburgh, Boston, Slovenia, Spain, France – and time: from video calls in-between brief holidays, coffees at airports, and lunches and trips to the beach between conference talks, to regular parkruns and theatre visits. To all of you – and you know who you are – thank you.

All mistakes remaining my own, this work, and particularly the parts that ventured into maths and/or required some crazy web-programming, would have been much poorer had it not been for Ben Davidson's brilliant mind, kind heart, and most instrumentally his perfect oat milk hot chocolates, which powered much of the writing of this thesis. Thank you for everything.

In my family, love never had to be earned, asked or thanked for. It was given generously and unquestioningly. Over the years I've come to see what a rare privilege that is, so I am going to break the family rule and thank my parents Tanja and Milan and my sisters Lara and Nađa Vujović for their unconditional love and unwavering support. If I ever achieved anything, it is because of you. *U slučaju neuspjeha, zborno mjesto ovdje.*

Mojem dedi Pavlu i babi Vjeri Vujović, koji više nisu tu, i dedi Novaku i babi Veri Vučinić

Abstract

How do children learn language in a way that allows generalization – producing and comprehending utterances that have never been heard before? A prominent view is that this is achieved through learning the statistical distributions of linguistic forms in the input. There is extensive evidence that humans are sensitive to the statistics of the input, but the exact nature of the learning mechanism that underpins it is unclear.

In this thesis, a discriminative approach to learning is taken, whereby language learning is a process of reducing uncertainty about the form and the meaning of the message by discriminating between informative and uninformative cues in the environment and in the utterance itself. This process is driven by key principles of learning theory – prediction error and cue competition, which are available to differing degrees in different learning contexts. Specifically, (1) learning suffixes provides greater cue competition than prefixing, which facilitates generalization via discriminative learning; (2) learning prefixes, on the other hand, facilitates the processing of upcoming parts of the utterance, because the prefix smoothes information content over the whole utterance, which promotes better learning of the utterance itself compared to suffixing.

This thesis tests these predictions in a series of artificial language learning experiments (with adult native speakers of English), which are either "suffixing" or "prefixing". Support for (1) was found across multiple experiments, but no consistent evidence for (2) was found. On the whole, the thesis demonstrates that discriminative learning provides a coherent theoretical basis for testing specific predictions about language, and identifies avenues for future work to address (2) more appropriately.

Impact Statement

Language learning has traditionally been described as learning a system of symbolic rules which operate across discrete items, which is "unlearnable" unless at least some linguistic knowledge is innate. An alternative has been that the productive system emerges as a consequence of learning the statistics of the input. A large body of evidence suggests that humans are sensitive to statistical information in the input, however, there has not been a testable model of the mechanism that underpins this statistical learning, and the conditions that facilitate and constrain it. As a result, while most researchers agree that some kind of statistical learning is critical for language learning, this has not had the fundamental effect on theories of language that it ought to have had.

This thesis proposes that discriminative learning theory, put forward by Ramscar and colleagues, provides a long over-due re-think about the nature of language learning. This theory sees language learning not as learning a system, but as reducing uncertainty about the meaning and the form of the utterance, by discriminating informative from uninformative cues in the environment. Critically, this theory makes specific predictions about when learning does and does not occur (which are based on empirical evidence and are captured formally in a mathematical model). This thesis tests the key predictions of the theory. While evidence was found for some and not for others, the thesis has demonstrated that the discriminative learning approach is a productive theoretical framework for re-thinking notions about language learning which have proven to be difficult to define and test in a way that can generate knowledge and advance the field.

While the immediate impact of developing a testable theory of language learning primarily concerns academic scholarship, where we still far from a coherent account (based in theory and empirically tested) of human language learning, the impact is ultimately far-reaching – for example, such an account can inform areas of artificial intelligence which have been notoriously challenging: speech recognition (particularly of atypical speech or speech in languages with little labelled data), sentiment analysis, and integrating context in language comprehension, to name a few. One of the key results from this thesis, however, does have immediate impact for second language learning in the classroom and atypical language therapy. This thesis shows that, as predicted by the discriminative learning theory, providing learners with more cues first, followed by simpler information later, facilitates learning. Using a simple example, this means that showing learners a picture of a scene/object followed by the word denoting that scene/object (rather than the other way round) helps the learning of the meaning of the word. This thesis is not the first to demonstrate this effect, however, it was demonstrated with over four hundred learners and therefore substantially contributes to our confidence in the effect.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

One of the key questions in cognitive science has been: how do children learn language productively, that is, in a way that allows them to produce and comprehend utterances that have never been heard before? This thesis takes a probabilistic view of language, whereby, rather than being governed by rules, human language is fundamentally probabilistic, and child language learning is underpinned by the ability to extract probabilistic structure from the input the child receives (Christiansen & Chater, 2008; Evans & Levinson, 2009; Seidenberg & MacDonald, 1999). In this introduction, I review evidence for the probabilistic approach, first from studies of natural language processing (Section 1.1), then from studies of language learning, both from child first language acquisition (Section 1.2.2) and from artificial language learning experiments (Section 1.2.3). I then turn to the critical question: if language *is* probabilistic, what kind of learning mechanism underpins probabilistic language learning and processing? A complete answer to this question must come from a theory which makes precise, testable predictions about when learning does and does not occur, and these predictions must be firmly grounded in psychological theory and empirical evidence. With these criteria in mind, in Section 1.3 I discuss some prominent approaches to modelling the underlying language learning mechanism. Although several approaches have successfully captured many language learning and processing phenomena, they have been criticised for a lack of grounding in psychological theory of learning, which limits the extent to which they advance our understanding of the underlying learning mechanism for language. In Section 1.4, I propose that discrimination learning framework for human communication, formulated by Ramscar and colleagues (Ramscar & Yarlett, 2007; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010; Ramscar & Dye, 2011; Ramscar, Dye, & McCauley, 2013) meets these criteria, and review this framework in more detail. The introduction ends with an outline of how the predictions of this framework regarding linguistic generalization and item-learning will be tested in this thesis (Section 1.5).

1.1 Language as a probabilistic system: evidence from natural language processing

A variety of psycholinguistic methods have demonstrated that human language processing is constrained by the probability distributions of different forms in the linguistic input, the so-called "statistics of the input", in complex and dynamic ways. In this section, I review some of the key work illustrating the probabilistic nature of language processing at different levels of linguistic organization, but I note that this review merely scratches the surface of a large and complex literature, and refer the reader to more extensive reviews (e.g., Ellis, 2002; Seidenberg & MacDonald, 1999).

1.1.1 Sub-lexical and lexical level

Part of human linguistic knowledge at sub-lexical level concerns the knowledge of phoneme co-occurrence within words, known as phonotactics. For example, in English, no word begins with the velar nasal, but when this sound does occur, it is mostly word-finally (e.g, /sɪŋ/), before velar stops (e.g., /fɪŋgər/), and very rarely before the unstressed schwa (e.g., /gɪŋəm/). While phonotactics have traditionally been described as a set of "rules" (L. Jones, 1967), psycholinguistic studies, an particularly studies of non-word processing, have provided evidence that human implicit knowledge of phonotactic constraints is probabilistic. Jusczyk, Luce, and Charles-Luce (1994) presented 9-month-old infants with nonwords which had phonotactic properties which were more or less frequent in English. Infants attended longer to nonwords with high-frequency phonotactic patterns than to nonwords with low-frequency phonotactics, suggesting that infants as young as nine months are sensitive to the probability of phonemes co-occurring in their first language input. Frisch, Large, and Pisoni (2000) reported the same result with adults: in their study, nonwords which had high-frequency phonotactic patterns were judged as more English-like, and were recognised faster than nonwords with low frequency phonotactics (see also Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997).

Edwards, Beckman, and Munson (2004) asked children aged 3-8 to repeat nonwords which contained two-phoneme sequences which were either high- or low-frequency in English. Children produced the same sound faster when it occurred in a high-frequency sequence than a low-frequency sequence (see Beckman & Edwards, 2000; Gathercole, Frankish, Pickering, & Peaker, 1999; Munson, 2001, for similar results with children). Importantly, this effect was stronger for children with smaller vocabulary size than children with larger vocabulary size or adults (see also Frisch, Large, Zawaydeh, & Pisoni, 2001; Gathercole, Hitch, Service, & Martin, 1997; Gathercole, Willis, Emslie, & Baddeley, 1992). This is to be expected – the more words one knows, the more likely one is to encounter low-frequency phonotactic patterns, and consider them more English-like. Nonwords containing these low-frequency patterns will therefore be similar to more words that the participant knows if the participant has a large vocabulary, and this will make the nonwords easier to process and repeat. The effect of vocabulary size on nonword processing is a powerful demonstration that linguistic knowledge is an emergent property of ever-growing interaction

with the input.

Frequency effects are not unique to the acoustic aspect of words. Humans are also sensitive to the frequencies of word meanings, for example. When presented with an ambiguous word, listeners/readers select an interpretation consistent with the more frequent meaning of the word. Evidence for this comes from the cross-modal priming paradigm, in which participants are presented with an ambiguous word in a sentence such as *The woman noticed a pen*, where *pen* could mean the instrument for writing or an animal enclosure. This is followed by the presentation of another word which is either semantically related to a high-frequency meaning of the word (e.g., *paper*), a low-frequency meaning of the word (e.g., *farm*), or to a completely unrelated word (e.g., *balloon*). When sentence context does not strongly favour either meaning of the word, participants are faster to respond to the high-frequency meaning than the low-frequency or the unrelated word, suggesting that listeners select the interpretation that favours the more frequent meaning (e.g., Rodd, 2017; Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982; Twilley & Dixon, 2000). However, when the preceding context strongly cues the less frequent meaning (e.g., *The woman noticed the pigs sleeping in a pen*), participants are faster to respond to the low-frequency meaning (e.g., MacDonald, Pearlmutter, & Seidenberg, 1994; Twilley, Dixon, Taylor, & Clark, 1994).

To summarize, adults and children as young as three have implicit knowledge of the probabilities of phonotactic patterns in the words they know, and this knowledge changes as they learn more words; humans are also sensitive to how phonologically similar a word is to other words, and whether the target word is similar to more or fewer words that they know. There is also evidence that processing is constrained by the statistical distributions of word meanings. This statistical information at lexical and sub-lexical levels is integrated in on-line language processing in complex ways.

1.1.2 The level of morphology

Much of the work discussed in this introduction – and in psycholinguistics in general – has been done with English. As English has little inflectional morphology compared to many other languages of the world, most of the studies of learners' sensitivity to the statistics of the input at the level of morphology focused on one of the few aspects of English where inflectional morphology is used – the marking of verb past tense. In English, the most frequent verb past tense inflection is the -ed morpheme. Children from about three years of age apply this form to novel verbs such as *wug*, producing *wugged* (Berko, 1958), and new verbs that enter the language are most likely to follow this pattern, such as *blogged*, *texted* or *spammed*. A rule-based approach to English inflectional morphology has described this in terms of a rule, whereby the past tense of a verb is generated by adding the -ed morpheme to any verb stem that belongs to the right VERB category (abstract syntactic category). However, many verbs in English do not follow the -ed pattern: *go – went, eat – ate, spring – sprang, blow – blew, feed – fed*, etc. – these are the so called "irregular" forms. According to the rule-based approach, while regular forms are generated by applying a symbolic rule over the appropriate stems, irregulars are listed in the mental lexicon as exceptions which

block the application of the rule; the child avoids overgeneralization by memorizing these forms (Pinker & Prince, 1988; Prasada & Pinker, 1993).

However, a closer look at adult and child knowledge of irregular forms reveals that irregulars are not simply unproductive exceptions listed in the mental lexicon: adults and children are likely to generalize these forms to novel verbs that are similar to existing irregular forms. Importantly, the way irregular patterns are generalized is probabilistic and it closely reflects the frequencies with which these patterns occur in the input (and as such cannot be straightforwardly captured by symbolic rules). For instance, adults and children from the age of three are more likely to generalize the irregular pattern (e.g., *spring – sprung*) to novel verbs such as *spling* and *krink* (giving *splung* and *krunk*, respectively) than to verbs such as *vin* (giving *vun*) (Albright & Hayes, 2003; Blything, Ambridge, & Lieven, 2018; Bybee & Moder, 1983). This is related to the fact that novel verbs such as *spling* and *krink* are phonologically similar to many existing forms that undergo this irregular pattern, such as *spring*, *string*, *drink* and *cling*, whereas verbs of the form consonant-ɪ-n/m are much less likely to follow this pattern (*dim – dimmed/*dum, pin – pinned/*pun*, etc.), showing that adults and children as young as three have an implicit knowledge of how likely certain inflectional patterns are. This is similar to the way knowledge of how likely certain phonotactic patterns are affects the processing of nonwords (Section 1.1.1). Similarly, acceptability judgment studies showed that adults (Albright & Hayes, 2003) and children aged 9-10 (Ambridge, 2010) rated past tense forms of novel verbs as more or less acceptable depending on their similarity to frequent existing forms. For example, *flept* was a more acceptable past tense form of *fleep* than was *fleeped*, due to similarity to existing forms such *sleep – slept, weep – wept*, whereas *nace – naced* was more acceptable than *nace – noce*, again, likely due to similarity to existing forms such as *face – faced* and *trace – traced*.

In this section we have seen evidence that the way adults and children generalize past tense inflections and judge the acceptability of novel past tense forms is affected by the frequency of similar forms in the input, suggesting that speakers generalize by analogy over existing exemplars, rather than by indiscriminately applying a rule to any word of the correct type. The studies discussed above suggest that, instead of necessarily being governed by the induction of deterministic rules, processing and learning of inflected forms may be driven by implicit knowledge of the probability distribution of inflected forms in the input.

1.1.3 Multi-word level

Studies of adult and child language processing have shown that listeners and readers are sensitive to the probabilities with which certain words occur in specific syntactic structures. Trueswell, Tanenhaus, and Kello (1993) studied whether the fact that verbs can occur in different structures with varying probability affects sentence processing. For example, many verbs occur both with noun phrase complements (e.g., *The chef found the recipe in a book*) and with sentence complements as in (e.g., *The chef found the recipe was complicated*), but the noun complement tends to be more frequent. The authors presented participants

with written sentences in which verbs that generally occur more frequently with noun phrase complements were actually followed by sentence complements, such as: The man accepted (*that*) the award would go to his brother, where the complementizer *that* was presented half of the time. Participants took longer to read the words following the noun phrase (*would go to...*) in sentences without the complementizer than in sentences with the complementizer. In the absence of the complementizer, participants interpreted *the award* as the direct object of *accepted*, demonstrating that participants used their implicit knowledge of how likely a verb is to occur in a specific syntactic context to guide processing. This effect has been demonstrated since with different syntactic contexts, with both adults and children (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Snedeker & Trueswell, 2004; Trueswell & Kim, 1998; Trueswell, Sekerina, Hill, & Logrip, 1999).

Further work in this paradigm demonstrated that distributional information in the input is combined with contextual information to constrain language processing in sophisticated ways. Snedeker and Trueswell (2004) presented adults and children with four objects: a stuffed toy frog holding a feather, a tiger, a feather, and a stick, and played sentences such as: "X the frog with the feather", X being a verb. The sentence could either mean (1) "Use the feather to X the frog" or (2) "X the frog that is holding the feather". The verb was either a verb that frequently occurs in an instrument phrase (e.g., *tickle, poke, hit*) and should favour the interpretation in (1), or a verb which rarely occurs in an instrument phrase (e.g., *hug, choose, find*) and should favour the interpretation in (2). Participants' actions and eye movements suggested that, in line with the literature, both adults and children chose the interpretations consistent with the verb frequency. However, a slightly different pattern of results emerged when participants were presented with referential cues which favoured an interpretation inconsistent with verb statistics. Specifically, on some trials, the tiger was replaced with another toy frog holding another object, which favoured the interpretation in (2) regardless of the statistics of the verb. For adults, these trials decreased the rate of instrument actions and duration of looks to the instrument (but did not entirely eliminate them). Children, on the other hand, still choose the interpretation consistent with the verb, but their eye movements revealed hesitation in looks to the instrument on referential trials. This suggests that children may start out by over-relying on distributional cues, but that they gradually learn to integrate them with other cues, possibly due to contextual cues being harder to master.

In summary, there is evidence that humans are sensitive to statistical distributions at the level of multi-word combinations. Together with the work reviewed in the other sections, this shows that language processing is driven by statistical information at all levels of organization – from phonemes and morphemes to syntactic constructions[1]. Importantly, it is not simply the case that more frequent words, constructions or word meanings are always easier to process, or that more frequent patterns will always be generalized. Rather,

[1]I use terms *phoneme, morpheme, syntactic construction* as linguistic descriptors. While there is evidence that humans are sensitive to statistical information at the level of what linguists describe as phonemes, morphemes and syntactic constructions, this is not necessarily to say that these units have psychological reality in the human mind (see Elman, 2009; Goldinger & Azuma, 2003; Ramscar & Robert, 2015, for discussions).

humans integrate information from multiple sources – distributional, contextual, referential – and from multiple levels of representation in a dynamic way, resulting in complex effects we observe empirically. We have seen in this section indications that this knowledge is developed and refined gradually during language learning. In fact, under the probabilistic approach, language learning is learning to process language with adult-like proficiency, which at least in part involves learning probabilistic constraints (Christiansen & Chater, 2008; Seidenberg & MacDonald, 1999). The next key question to consider is, how do humans accrue this knowledge of input statistics? In the next section, I discuss evidence from language learning literature – both naturalistic of early child language and lab-based experiments – which suggests that human language learning is underpinned by the ability to learn input statistics.

1.2 Language as a probabilistic system: Evidence from language learning

There is extensive evidence that child language learning is shaped by the statistics of the input, suggesting that the ability to learn input statistics may be the driving learning mechanism in language development. Before reviewing this evidence in Sections 1.2.2 and 1.2.3, I give a brief review of the theoretical background in Section 1.2.1, in which I place the probabilistic approach in the context of major theoretical approaches to language learning.

1.2.1 Theoretical background

Mainstream linguistics has been dominated by the *generativist/nativist approach* proposed by Chomsky (1957, 1965). This approach argues that human linguistic knowledge is a grammar of rules, and that humans are born with the knowledge of these rules. The approach is generativist in its view of grammar – generativist theories describe grammar as set of symbols and abstract logical rules which operate across those symbols, such that the grammar generates all and only grammatical sentences of a language. Importantly, this grammar allows humans to use language productively, that is, to produce and comprehend utterances that have never been heard before. Over the years, various theories have been put forward as to what exactly this grammar involves: a system of rules and representations (Chomsky, 1980); a set of linking rules for mapping thematic roles onto syntactic categories, as well as rules on how the linking rules are combined (Pinker, 2013); a set of principles that are common to all languages and a set of parameters on which the principles can vary (Chomsky & Lasnik, 1993). Most recently, as a result of a minimalist program which strives to significantly reduce the amount of structure required to describe grammar, the Bare Phrase Structure theory described grammar in terms of only two rules (Chomsky, 1994, 1995).

While different generativist theories may disagree as to what exactly generative grammar may involve, what makes them *nativist*[2] is that they argue that generative grammar

[2]In principle, a theory can be generativist but not nativist – such theory would argue that grammar is a set of rules, but that the knowledge of these rules is not innate, though no such theory is currently known

cannot be learned from the input that the child receives, and that therefore it must be innate. Specifically, language cannot be learned from the input because the input is impoverished, as the child only ever hears a small portion of all possible utterances in the language; the input is idiosyncratic, as learners of the same language are exposed to input that is too variable across learners; finally, the input does not contain the necessary negative evidence: the child only ever hears grammatical sentence, so she cannot know what not to say (the so-called "argument from the poverty of the stimulus"; Chomsky, 1965; Crain, 1991; Lightfoot, 1991). Therefore, all humans are born with knowledge of generative grammar which describes all human languages, and the task of learning a specific language becomes one of mapping the input from that specific language onto this innate universal knowledge (Chomsky, 1995; Roeper & Williams, 1987).

Importantly, the generativist/nativist approach argues that, while language processing may be constrained probabilistically (as discussed in Section 1.1), these constraints are separate and independent of grammar. Chomsky (1957) famously argued that the grammar of English is fundamentally distinct from a "statistical approximation of English" (p. 16):

1. Colorless green ideas sleep furiously.

2. *Furiously sleep ideas green colorless.

While the sentence in (1) has never been heard before (and therefore cannot be found in a statistical approximation of English), speakers of English would judge it as grammatical (albeit nonsensical); (2), on the other hand, is clearly ungrammatical. Chomsky (1957) concluded that probabilistic models would wrongfully reject both sentences as ungrammatical, and in doing so would fail to capture "basic problems of syntactic structure" (p. 17). However, researchers have demonstrated since that Chomsky's argument is a misinterpretation, and that probabilistic approaches can indeed distinguish between (1) and (2). Specifically, (1) is similar to *Revolutionary new ideas appear infrequently* (Abney, 1996), and the sensitivity to the existence of similar forms is integral to how humans generalize and judge the grammaticality of novel forms (see Section 1.1.2). Pereira (2000) captured this formally, by developing a model which computed similarities between sentences in English, and found that (1) was more probable than (2) by several orders of magnitude.

Rather than being a constraint that is "added on" to the innate grammar, an alternative is that the statistics of the input are inherent to human linguistic knowledge, and learning the input statistics is what enables humans to learn language. This alternative avoids postulating an independent innate module of the mind (the innate grammar), and is therefore more testable – if child language develops in a way that is consistent with the statistics of their input, this may suggest that language learning is input-driven. Various related approaches adopt the notion that human linguistic knowledge is not innate, and

to exist (Ambridge & Lieven, 2011). There are nativist theories of language learning in domains other than grammar which are not generativist: for example, the lexical principles approach to word learning (Markman & Wachtel, 1988), which assumes that children are born with a set of biases as to what new words are more likely to refer to.

that it emerges gradually in the course of learning how to comprehend and produce language. Where approaches differ is the exact description of the linguistic knowledge that emerges, discussed below.

One prominent approach, which is discussed in more detail in Section 1.3, rejects the notion of grammar as a collection of symbolic rules, and instead views linguistic knowledge as a network of mappings between forms and meanings, where grammar-like abstract representations (e.g., at the level of syntax and morphology) are emergent properties of learning from the input (see Christiansen & Chater, 1999; Elman, 1990; Rumelhart & McClelland, 1986). This approach is most commonly referred to as *connectionism*, though *emergentist* is also used. A related approach, the *constructivist approach*, views human linguistic knowledge as an inventory of stored constructions (Goldberg, 1995). Constructions are abstract form-meaning pairs which are constructed as the child learns to perform communicative functions. For instance, the English transitive construction *John kicks Bill* takes the form of: NOUN1 VERB NOUN2, and its meaning counterpart roughly corresponds to: "A acts on B". Critically, because constructions are form-meaning mappings, a construction grammar can generate novel utterances: if the child sees Mummy kicking a ball, the child can link the abstract representation "A acts on B" to its form pair, and produce *Mummy kicks the ball*.

Connectionist and constructivist approaches are the most prominent non-nativist approaches, but other theories exist. For example, while constructivist approaches assume abstract representations, exemplar-based approaches argue against stored abstractions. Instead, the learner stores all exemplars in memory and generalizes via rapid, on-the-fly analogy across all stored items (e.g., Ambridge, 2018; Bybee, 1985, 2010; Croft, 2001). Finally, not all non-nativist approaches reject the notion of rules: for example, Marcus et al. (1992) argue that abstract symbolic rules are learned from the input.

Different non-nativist approaches model human linguistic knowledge in different ways. What they all have in common though is that they reject the idea of innate linguistic knowledge (while, of course, treating the ability to learn language as innate and uniquely human), and argue instead that language is learned from the input. Tomasello (2000) described learning from the input as a gradual process. Specifically, the child starts by memorising phrases directly from the input and grouping them together based on similarities in meaning into item-based schemas. For example, the child may initially use the transitive construction in a few phrases only (e.g., *I'm kicking it* and *I'm drinking it*). As the inventory of item-based schemas grows, the child forms analogies between similar schemas, which allows for more abstract constructions to emerge. The child gradually moves away from item-based usage of specific constructions to, ultimately, adult-like abstract generalizations. Importantly, this process is underpinned by the ability to extract statistical information – to track probability distributions of linguistic forms across the input. Tomasello (2000) refers to this mechanism as *pattern-finding*; elsewhere in the literature it is referred to as *statistical learning*. Before turning to empirical evidence for the role of statistical learning in language development, I want to point out that no theory argues that statistical learning alone explains all of language learning. Much of language

learning occurs through interaction with others, and Tomasello (2000) places particular emphasis on socio-cognitive cues and mechanisms such as joint attention, gaze-following and intention-reading as critical for language learning. There is extensive evidence that these socio-cognitive mechanisms are important in early language development (e.g., Akhtar, Carpenter, & Tomasello, 1996; Baldwin, 1993; Mundy, Sigman, & Kasari, 1990; Nadig & Sedivy, 2002; Tomasello & Haberl, 2003). The importance of socio-cognitive is not incompatible with the probabilistic approach I take in this thesis – quite the opposite, under the probabilistic approach, language is integrated with other aspects of cognition. The reason this review primarily discusses the role of learning input statistics is that my focus is on lower-level cognitive mechanisms. While these are different from, they also underly and feed into higher-level processes such as social cognition – the same low-level learning mechanisms such as statistical learning may underpin learning from socio-cognitive cues, too (see Section 1.2.3.3). Socio-cognitive approaches to language are not, therefore, mutually exclusive with the probabilistic approach to language, rather, these approaches are concerned with different levels of cognition.

1.2.2 Evidence from early child language

What might constitute evidence that human linguistic knowledge emerges gradually, from learning the statistics of the input? If we are born with a grammar of symbolic rules, and if language learning amounts to "working out" how these rules are realised in a particular language, then we should expect to see evidence of abstraction in children's early language – however, this is not what has been observed empirically.

Numerous corpus analyses of child speech as well as elicitation studies, in which researchers prompt children to produce certain forms, have shown that children's early use of grammatical structures is largely tied to specific items from the input, rather than abstracted and generalized. For example, Pine and Lieven (1997) studied a sample of 12 English-speaking children aged 2-to-3-years-old (from Lieven, Pine, & Baldwin, 1997), and found that, when the children first started using the determiners *a* and *the*, they used each determiner with almost entirely non-overlapping sets of nouns. This is inconsistent with generativist approaches, which suggest that children are born with a DET + NOUN rule, where any noun of the right category could be combined with any determiner. Instead, children's early use of determiners is idiosyncratic, suggesting a lack of abstract representations and rules at least until the age of three.

In the domain of verb morphology, Rubino and Pine (1998) analysed a three-year-old learning Brazilian Portuguese, and found that the child used the correct plural subject-verb agreement with only a few verbs, and these verbs were all highly frequent in the mother's speech. Similarly, Aguado-Orea and Pine (2015) studied two Spanish-speaking two-year-olds and found that children's use of verb forms was significantly more item-based than was the case with their parents, even when controlling for differences in vocabulary range, sample size and knowledge of different inflected forms. For example, for both children, over 50% of all 1st person singular uses were accounted for by the verb *querer* (to want), followed by *poder* (can) and *tener* (to have), and together these verbs made up over 75%

of all 1st person singular uses.

Generativist approaches would predict that the child had an abstract representation VERB, which is combined with morphological inflections according to a set of rules, and that these rules would indiscriminately apply to any item of the right category (subject to any constraints from working memory and sensorimotor constraints, which are external to grammar). This, however, does not seem to be the case empirically –children tend to use certain morphological forms with certain verbs only. Importantly, the individual verbs on which children base their early use of certain morphological forms is not random: in Aguado-Orea and Pine (2015) corpus, the verbs on which children based most of their use of inflected forms were all high-frequency verbs in Spanish; in Rubino and Pine's study, these verbs were frequent in the speech of the child's mother, and often occurred in the same form in mother's immediately preceding speech (Rubino & Pine, 1998).

While children's early use of certain subject-verb agreement forms was based on a few highly frequent verbs, children rarely made agreement errors – Rubino and Pine (1998) reported a 3% error rate. While low agreement error rates have been taken as evidence for innate knowledge (e.g., Wexler, 1998), Rubino and Pine (1998) found that, when the child in their study *did* make agreement errors, they were much more likely with low-frequency verbs and low-frequency agreement forms. Corpus and elicitation studies in English reported similar findings: both overregularization errors such as **goed, *drinked* or **fishes*, and errors of omission such as *she *like* are more likely with low-frequency than high-frequency verbs and nouns (e.g., Marchman, 1997; Maslen, Theakston, Lieven, & Tomasello, 2004; D. E. Matthews & Theakston, 2006). The effect of frequency on errors with inflectional morphology has since been reported in numerous languages other than English, many of which are much more inflected than English: in Finish (Räsänen, Ambridge, & Pine, 2016), Greek (Stavrakaki & Clahsen, 2009), Polish (Dabrowska, 2008), Serbian (Mirković, Seidenberg, & Joanisse, 2011), and Lithuanian (Savičiute, Ambridge, & Pine, 2018).

The effects of input frequency on child language are also observable beyond the level of individual inflected forms. At the level of multi-word strings, Bannard and Matthews (2008) asked two- and three-year-olds to repeat four-word sequences which were either highly frequent in child-directed speech corpora (e.g., *Up in the air*) or low-frequency (e.g., *Up in the bath*). In both age groups, children repeated high-frequency sequences more accurately, and three-year-olds also produced the final word faster in high-frequency than in low-frequency sequences (see also D. Matthews & Bannard, 2010; Arnon & Snider, 2010; Arnon & Clark, 2011), suggesting that children's early linguistic ability is affected by word co-occurrence statistics in their input. Similar findings come from the so-called *weird-word order studies*. D. Matthews, Lieven, Theakston, and Tomasello (2005) presented children aged 2;9 and 3;9 with English sentences describing simple scenes. The sentences used the SOV word order (not used in English), and contained either verbs which were high frequency in the Manchester corpus of child-directed speech (Theakston, Lieven, Pine, & Rowland, 2001), such as *Bear Elephant pushed*, or low-frequency, such as *Bear Elephant rammed*. When asked to describe the same scene, children aged 2;9 were more likely to use the "weird" word order for

low-frequency verbs, but reverted to the canonical SVO word order for high-frequency verbs. Children aged 3;9, however, used the SVO word order regardless of verb frequency. This finding was interpreted as evidence for the constructivist/usage-based approach, whereby English word order is learned gradually, from item-based schemas formed around frequent items to more abstract constructions, which develop as more input is experienced (see also Abbot-Smith, Lieven, & Tomasello, 2001; Akhtar, 1999). Finally, Ambridge, Pine, Rowland, and Young (2008) reported similar results with older children (aged 5-6 and 9-10), who were more tolerant of ungrammatical uses of low-frequency verbs than lhigh-frequency verbs – children judged **The magician vanished the rabbit* (low-frequency verb) as more acceptable than **The magician disappeared the rabbit* (high-frequency verb), though of course both were less acceptable than their grammatical counterparts (*The magician made the rabbit vanish/disappear*).

We have seen evidence that children's early language is built around high-frequency forms in the input, and that this knowledge is accrued gradually, and is refined as more experience with the input is gained. The facilitative/protective effects of high-frequency forms early in childhood suggests that the ability to track probability distributions of different forms must be incorporated into the underlying learning mechanism for language in a way that is more fundamental than is the case with generativist approaches, which keep probabilistic learning as separate from grammar. However, the studies discussed in this section used naturalistic or elicited data from the child's first language, meaning that the researcher does not have full control over the complexities of natural language. This means that we cannot rule out the possibility that something other than differences in input-frequency facilitates early learning of high-frequency items; this may be driven by other cues such as intonation or cues from the interactive context, such as eye gaze and other non-verbal cues. Even though researchers control for these cues in elicitation studies, it is still possible that they played a key role in the child's learning prior to being tested in the lab, and the researcher does not have control over the amount and kind of input that the child received. Therefore, while they provide rich and complex data, naturalistic studies of language development may not be best-suited for pin-pointing the role of statistical learning in language development.

Some of the issues with studying natural language can be remedied using the artificial language learning paradigm (sometimes also referred to as artificial grammar learning). In this paradigm, participants – infants, children, and adults – are exposed to miniature artificial "languages" with grammars in which different forms have different probability distributions, and are tested on novel instances of that language which may or may not follow the distributional patterns from the input. If learners respond to novel instances differently depending on their distributional properties, this is taken as evidence that they can generalize the probabilistic structure of the artificial language. Importantly, this paradigm allows the researcher to isolate statistical cues from other cues, by designing an artificial language where statistical cues are the only cues to grammatical structure available to the learner. Finally, using this method, the researcher can manipulate exactly how much input the learner receives.

In the next section, I review some of the key experiments that demonstrated that distributional information alone is sufficient for humans to learn and generalize patterns in artificial languages (see Aslin & Newport, 2014; Culbertson & Schuler, 2019; Erickson & Thiessen, 2015; Frost, Armstrong, & Christiansen, 2019; Romberg & Saffran, 2010; Wonnacott, 2013, for more extensive reviews). In this thesis I refer to this body of work as the "statistical learning literature", though note that artificial language learning experiments have been used to study other language learning phenomena, from learning biases (e.g., Culbertson & Newport, 2015; Culbertson, Jarvinen, Haggarty, & Smith, 2019), regularization of unpredictable variation (e.g., Fehér, Wonnacott, & Smith, 2016; Hudson Kam & Newport, 2005; K. Smith & Wonnacott, 2010), to emergence of communicative systems (e.g., Kirby, Cornish, & Smith, 2008; Kirby, Tamariz, Cornish, & Smith, 2015; Raviv & Arnon, 2018).

1.2.3 Evidence from artificial language learning experiments

Although the idea that distributional cues drive language learning is not new, and that the first uses of the artificial language learning paradigm can be traced back about fifty years (Braine, 1963; Reber, 1967; K. H. Smith, 1966), the paradigm was brought into mainstream focus in a highly influential study by Saffran, Aslin, and Newport (1996). They exposed 8-month-olds to a continuous speech stream: *bidakupadotigolabubidaku...*, the only cues to word boundaries being the transitional probabilities between syllables. For example, *bi* was always followed by *da*, meaning that *bida* was a "word" in this language, whereas *ku* was followed by *pa* only 33.3% of the time. After two minutes of exposure to the speech stream, the infants were exposed to the words from training (e.g., *bida*), as well as "part-words", which were generated by taking the final syllable of a word and the first two syllables of the following word, (e.g., *dakupa*). Therefore, part-words contained syllables which crossed word boundaries, such as ty#baby in pretty#baby. Using a familiarization-preference procedure (Fernald & Kuhl, 1987; Nelson et al., 1995), in which infant listening time is measured as the length of fixation on a blinking light, the authors found that the infants listened significantly longer to part-words compared to familiar words. This is taken as an indication that infants were able to pick up on word boundaries on the basis of the transitional probabilities between syllables alone. This finding has been replicated (e.g., Aslin, Saffran, & Newport, 1998; Perruchet & Desaulty, 2008) and extended to non-speech stimuli (Saffran, Johnson, Aslin, & Newport, 1999) and natural foreign language stimuli (Hay, Pelucchi, Estes, & Saffran, 2011; Pelucchi, Hay, & Saffran, 2009). Related work has also shown that learners can incorporate top-down distributional information at the lexical level during artificial speech segmentation (Lew-Williams, Pelucchi, & Saffran, 2011).

The finding that infants as young as eight months can segment artificial speech based on transitional probabilities between syllables alone is a powerful demonstration of statistical learning, and this work inspired a long debate about the nature of the learning mechanism required for language. However, if the argument is that statistical learning underpins language learning, we must be able to demonstrate that humans can learn statistical patterns beyond individual syllables, and generalize these patterns to novel instances. This is par-

ticularly important considering that much of the argument for innate grammar comes from the apparent impossibility to learn the complexities of syntactic structure (that is, beyond individual syllables or words) from the input. I discuss work in this area in the next section.

1.2.3.1 Beyond individual words: abstraction and generalization

Mintz (2002) demonstrated that learners can form category-like knowledge about groups of words on the basis of the distributional contexts in which they occur. Participants were exposed to six minutes of three-word artificial utterances, where groups of words occurred in the same "syntactic" frame: for example, *nex, kwob, zich,* and *pren* all occurred in the *bool_jiv* frame. In addition to this, all words from that "category" except one (e.g., *pren*) occurred in another frame, *sook_runk*. If participants have formed a syntactic category based on the occurrences of the words in the input, they should expect also to hear *sook pren runk* purely based on the distribution of *pern* in the input relative to other words. Note that no other cues were available to the learners -– the speech did not contain any useful acoustic cues (prosody, sound similarity, etc.), and no visual cues were available. At test, participants were exposed to *sook pren runk*, as well as to a control sentence in which *pren* occurred in a syntactic frame that was present in the input, but never occurred with the words of pren-category. Participants were asked to indicate on a scale how much the sentence was like the language they heard during exposure. As expected, participants judged *sook pren runk* to be more like the sentences heard in the input than, for example, *choon pren wug*, despite the fact that *pren* never occurred in either of the two frames in the input. Participants judged *sook pren runk* to be more like the language because other words which occurred in the same frame as *pren* also occurred in the *sook_runk* frame, suggesting that learners can track the distributional statistics of the input above-and-beyond transitional probabilities of individual syllables or words (see also Mintz, 2003), and can generalize this knowledge to unseen instances.

Related evidence for generalization beyond trained instances comes from Gomez and Gerken (1999), who exposed one-year-olds to two minutes of strings generated by a finite-state artificial grammar. At test, participants were exposed to the remaining grammatical strings as well as ungrammatical strings. Importantly, both the grammatical and ungrammatical testing strings were not heard during exposure – this is different from Saffran et al. (1996) among others, where ungrammatical strings were compared to trained strings, rather than to unheard grammatical strings. In a head-turn preference task, infants listened longer to grammatical than ungrammatical strings, despite the fact that none of the strings occurred in exposure, suggesting that infants' preference for grammatical strings cannot be explained by mere familiarity. Further evidence for learning beyond trained instances comes from Marcus, Vijayan, Rao, and Vishton (1999), who exposed 7-month-old infants to artificial utterances following an AAB or ABA pattern (e.g., *ga ga ti* or *li na li*). In a habituation procedure, infants were exposed to these strings until they failed to attend to the stimulus (taken to indicate habituation). At that point, infants were played strings that contained completely novel words, but that either followed the pattern (e.g.,*wo fe wo*) or did not (e.g., *wo fe fe*). Infants listened longer to ungrammatical strings, suggesting

that they were able to discriminate between novel stimulus based on whether or not it conformed to the pattern they heard during exposure. This pattern of results indicates that the infants were able to learn something about the structure of exposure stimuli that goes beyond individual words.

The studies reviewed in this section show that learners can generalize the statistics of the input to new instances. Strikingly, infants as young as seven months were able to do this in the absence of any cues other than the distributions of forms in the input. Can the same ability underpin natural language learning, and can learners do this without innate knowledge? One of the key arguments for innate linguistic knowledge is the supposed lack of negative evidence in the input: the idea that learners only ever hear grammatical forms, and from this evidence alone, it is not possible to learn when to generalize and when *not to generalize*. For example, how can the child learn that, while *John gave/sent/bought Bill the book* is correct, the same ditransitive construction does not generalize to certain other verbs: **John carried/donated/pushed Bill the book*? This learning problem, referred to as Baker's Paradox (Baker, 1979), received significant attention in the literature. A prominent view was that children avoid overgeneralization through the knowledge of verb semantics – for example, most verbs that occur in the ditransitive constructions are verbs of transfer (Pinker, 2013). However, correlating verb semantics to syntactic occurrence has been proven difficult in many cases; instead, researchers raised the possibility that the statistics of verb occurrence help the learner avoid overgeneralization[3] (Ambridge et al., 2012; Brooks & Tomasello, 1999; Gleitman, 1990; Goldberg, 1995). In the next section, I review artificial language learning experiments which show that the statistical structure of the input may provide the learner with sufficient cues as to how to constrain generalization.

1.2.3.2 Constraining generalization

In a series of experiments with adults, Reeder, Newport, and Aslin (2013) demonstrated that learners can constrain generalization based on the positive and negative evidence provided by distributional cues. Adult learners were exposed to sentences from an AXB grammar such as: *daffin tomber fluggit* ($A_1X_1B_1$), *glim tomber mawg* ($A_2X_1B_2$), but **daffin tomber glim* ($A_1X_1A_2$). As in other work, no other auditory or meaning-based cues were provided. At test, participants were presented with sentences from exposure, as well as untrained AXB sentences, and ungrammatical sentences of AXA and BXB type. Participants were asked to indicate on a 5-point scale how likely it was that the sentence came from the language. Grammatical sentences were rated more highly than ungrammatical sentences, but, importantly, trained sentences were not rated as more like the language than untrained grammatical sentences, suggesting that participants were able to abstract both categories of A, X and B words, and how these categories are combined into sentences, purely on the basis of the statistics of their occurrence in the input. However, when participants were exposed to a language with consistent gaps in the A_B contexts across X words (e.g., X_1

[3]In the context of avoiding overgeneralization with verbs, there is debate as to how exactly the child achieves this. Two competing learning processes have been proposed – *pre-emption* and *entrenchment*. I refer to Ambridge, Pine, and Rowland (2012) and Wonnacott, Newport, and Tanenhaus (2008) for discussions.

never occurred with A_3 or B_1, X_2 never occurred with A_1 and B_2, etc.), trained sentences were rated higher than untrained grammatical sentences, even though the untrained sentences conformed to the AXB grammar (though untrained grammatical sentences were still rated higher than ungrammatical ones). This suggests that learners treated these consistent non-occurrences of certain X words in certain contexts as relevant and meaningful cues to generalization, that is, for when not to generalize (see Reeder, Newport, & Aslin, 2017, for related findings). Returning to the example from the previous section, in the context of natural language learning, the child may pick up on the consistent non-occurrence of verbs such as donate, push or carry in the ditransitive construction, and from this learn not to use these verbs in that construction.

Reeder et al. (2013) also found that learners were able to apply their knowledge of input statistics not only to new combinations of existing words, but to new words altogether. In a series of follow-up experiments, a new X word, X_4, was added to the training set, but it only ever occurred in one context: $A_1X_4B_1$. How likely learners were to generalize X_4 to other A_B contexts depended not on the distribution of X_4 (it was only ever in one context), but on the distribution of other X words. When the other X words occurred with every other A and B word, the learners generalized X_4 to other contexts. When, however, the other X words consistently did not occur with certain A and B words, learners were less likely to generalize X_4 to novel contexts. This suggests that learners were able to track distributional information at category-level (specifically, that any X word can occur with any A and B word, or that X words are restricted in terms of which A and B words they occur with) and generalize this to novel instances.

Related evidence comes from Wonnacott (2011), who exposed 5-to-7-year-olds to an artificial language in which nouns were followed by one of two "particles". In the Lexicalist condition, three nouns occurred exclusively with one particle, and one noun occurred exclusively with the other; in the Generalist condition, on the other hand, nouns occurred with both particles, but one particle was three times more frequent. In both conditions, children were exposed to the nouns and particles in the first session. In the second session on the next day, children heard two new nouns in a few sentences only (the minimal-exposure nouns), and these nouns exclusively occurred with one of the particles (similar to the X_4 noun in Reeder et al., 2013). The key question was: can children learn the statistics of occurrence of the particles in the input, and, critically, can they extend this knowledge to the minimal exposure nouns? To test this, Wonnacott prompted the children to produce sentences with trained and minimal exposure nouns. Most children in both conditions learned the statistics of the input – they produced trained nouns with the particles in a way that reflected their distribution in the input. Importantly, children were able to extend these patterns to the minimal exposure nouns: children in the Lexicalist condition were more likely to produce the minimal exposure nouns with the particles that they occurred with during exposure, whereas children in the Generalist condition produced the minimal exposure nouns with both particles, even though these nouns only ever occurred with one of the two particles during exposure (see Wonnacott et al., 2008; Wonnacott, Brown, & Nation, 2017, for related findings). These findings demonstrated that children, much like adults

and infants, are able to track statistical information at the level of individual words as well as higher-order statistics. Critically, children can apply this knowledge to (relatively) novel instances in a way which allows correct generalization, while avoiding overgeneralization.

The work presented in Sections 1.2.3.1 and 1.2.3.2 showed that learners are sensitive to statistical information at multiple levels of abstraction, and can generalize this information to novel instances. Importantly, the statistics of the input provide the learner with cues for constraining generalization, that is, for knowing when not to generalize, something that has traditionally been assumed to be unlearnable form the input alone. Thus this work suggests that the ability to learn probability distributions of forms in the input may underpin many "hard problems" in language learning, such as learning how to abstract and generalize while avoiding over-generalization. While artificial language learning experiments have been instrumental in isolating the role of statistical learning in language learning under experimentally controlled conditions, it is now necessary to return to considering natural language as a whole, and how statistical learning may "scale up" to natural language learning, where distributional cues are just one kind of cues available to the learner. I turn to this in the next section.

1.2.3.3 Learning beyond "purely" distributional cues

In the studies reviewed so far, the useful cues available to the learners were "purely" distributional – the forms in the artificial languages shared no similarities in sound or meaning. In natural language, however, multiple cues are available to the learner. For example, phonological and semantic cues often predict class membership of a word (MacWhinney, Leinbach, Taraban, & McDonald, 1989; Zubin & Köpcke, 1981). In many gendered languages, linguistic gender correlates with phonological and semantic cues. In Serbian, most masculine nouns end in a consonant, and most feminine nouns end in -a (phonological cue), and, for example, nouns referring to vegetables tend to be masculine, while nouns referring to fruits tend to be feminine (semantic cue) (Mirković, MacDonald, & Seidenberg, 2005). There is evidence that children are sensitive to these phonologically and/or semantically-based patterns in their first language input. For example, Karmiloff-Smith (1981) found that French-speaking children as young as three are sensitive to phonological cues grammatical gender, such as the fact that nouns that end in -*elle* tend to be of feminine gender. Karmiloff-Smith taught the children nonwords denoting novel male or female alien characters. The children in their study used the feminine article *la* with nonwords such as *podelle*[4], suggesting that children can generalize this probabilistic tendency to novel nouns that share the critical phonological features.

The fact that learners are sensitive to semantic and phonological cues in generalization is not incompatible with a statistical learning account as, in theory, these cues can also be learned by tracking their statistical distribution in the input. However, while numerous artificial language learning experiments have investigated the learning of phonological and

[4]Interestingly, in this study, children up to ten years of age used the feminine article even when the alien was male, suggesting that children may show a bias towards phonological cues to grammatical gender (see also Gagliardi & Lidz, 2014; Pérez-Pereira, 1991). I return to the different contributions of semantic versus phonological cues in Study 2.

semantic cues in novel input (Brown, Smith, Samara, & Wonnacott, 2018; Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Culbertson, Gagliardi, & Smith, 2017; Culbertson et al., 2019; Lany & Saffran, 2011; Frigo & McDonald, 1998), few have investigated learning when these cues are probabilistic, rather than deterministic. This seems particularly important considering that the phonological and semantic cues in natural language are often probabilistic – grammatical categories correlate with semantic and phonological features, but there are many examples where semantic and phonological cues are not reliable (to second language learners, these cases are often taught as "exceptions to the rule"). Frigo and McDonald (1998) taught adult participants novel names for individuals in a "new culture" in which individuals used different greetings for daytime and for the evening. Individuals were grouped into categories based on which version of the daytime and evening greetings they used. Critically, in the marked condition, 60% of the individuals who used the same variants of the greetings had names which shared a phonological marker (e.g., 60% of category 1 names ended in *-ash* and 60% of category 2 names ended in *-gor*), whereas 40% were unmarked. In the unmarked condition, the names were unmarked. At test, participants were given a completely novel name and were asked to produce the correct greeting for that individual for a given time of the day. When the name alone was not a sufficient cue to the correct greeting (for unmarked names in the marked condition, and for all names in the unmarked condition), participants were given one of the correct greetings and asked to produce the other. Only participants in the marked condition were able to generalize the correct greeting to new names, and strikingly, this was the case not only for novel names that contained the phonological marker, but also for unmarked novel nouns, suggesting that even the partially consistent phonological cue boosted the overall learning of distributional patterns in the language. This provides evidence that learners can learn probabilistic phonological cues to grammatical categories, and that phonological cues are integrated with distributional cues in a sophisticated way.

Intriguingly, attempts to show probabilistic learning of semantic cues to class membership have been less successful. Brown et al. (2018) taught children aged 5-to-6 artificial languages in which animate nouns co-occurred with class marker 1, except one animate noun which occurred with class marker 2, and found that the children were unable to generalize class marker 1 to new animate nouns at test. While it may be that children found the exception items particularly challenging because they shared the semantic features of the opposite category, Schwab, Casey, and Goldberg (2018) found that, when exception items do not share features with the opposite category, 6-year-olds were still unable to generalize based on semantic cues. They exposed children to two novel classifiers, *dax* and *po*, which co-occurred with items that were stereotypically female or male, respectively; the exception items for each classifier were arbitrarily chosen inanimate objects. While adults used the "correct" classifier with new male and female items, children showed chance-level performance. As there is evidence that both semantic and phonological cues predict grammatical categories in natural language (in addition to the purely distributional information; Mintz, 2002), it is necessary to determine whether both adults and children can learn these cues from noisy input, that is, from input where these cues are not deterministic.

Cues from similarities in sound and meaning are not the only generalization cues available to the learner. Extra-linguistic cues have been known to condition distributions of linguistic forms, such as socio-economic status (Labov, 1966), gender (Labov, 1990), or ethnicity (Hoffman & Walker, 2010). Samara, Smith, Brown, and Wonnacott (2017) showed that the same ability to track co-occurrences in the input might underpin the learning of socially-conditioned variation. They exposed adults and 6-year-olds to an artificial language in which nouns were followed by particles, and where the choice of particle was associated with the gender of the speaker producing the artificial sentence. At test, participants were prompted to produce the particle for a novel noun on behalf of the male or female speaker. Both adults and children were able to condition their use of the particles on the gender of the speaker; importantly, they were able to do this even when the particle usage was only probabilistically correlated with speaker's gender. Rather than tracking co-occurrences between linguistic forms alone (as in previous work), learners in this study tracked co-occurrences between linguistic forms and social variables.

To summarize, artificial language learning experiments have been useful in studying how learners can track the statistics of occurrence of not only linguistic forms (words, particles, etc.), but of their phonological and semantic features, as well as socio-linguistic variation. The experiments presented in this section suggest that statistical learning might operate across information from various domains.

1.2.3.4 Evaluating artificial language learning

Artificial language learning experiments have provided evidence that adults, children, and infants can learn probabilistic patterns in the input. Strikingly, learners can do this with artificial input in which the only useful cues for generalization are distributional in nature – where the only cues to a word's category, for example, is the frequency with which it co-occurred with other words in the input. This kind of distributional information facilitated and constrained generalization in sophisticated ways: generalization was shaped both by the statistics of occurrence, and non-occurrence, at lexical level and beyond, and the more consistent this positive/negative evidence was, the more strongly did it affect generalization. I also presented some evidence that the same ability to track the probabilities of occurrence (or the lack of occurrence) in the input may underpin the learning of phonological, semantic, and even entirely extra-linguistic cues for generalization. Generalizing in this way has traditionally been assumed impossible without at least some aspects of linguistic knowledge being innate. As discussed in Section 1.2.1, the alternative is that, rather than being innate, linguistic knowledge emerges bottom-up, from tracking input statistics (informed by socio-cognitive mechanisms and other extra-linguistic cues). While the results from artificial language learning experiments are more consistent with this theoretical alternative, two things are important to consider. I discuss each below.

First, is the learning that occurs in artificial language learning experiments a good model of how humans learn language naturally? While it is clear that artificial languages are purposefully simplified "snippets" of a language, and that this is done in order to isolate a particular aspect of language in controlled conditions, it may be the case that the

way humans learn these languages is fundamentally different from how natural language is learned. Specifically, first language learning is believed to be largely implicit – adult native speakers of a language often cannot explicitly verbalise awareness of rules and patterns in their language, despite using them in processing and production. Statistical learning also occurs implicitly: no theory hypothesises that humans compute transitional probabilities explicitly, nor that they can explicitly verbalise statistical patterns in the input. However, the extent to which the learning of artificial languages is implicit is less clear. This paradigm often involves explicit, meta-cognitive tests, such as asking participants how likely it is that the sentence has come from the language, which may interfere with what was learned during exposure, and also may encourage explicit strategizing (see Siegelman, Bogaerts, Christiansen, & Frost, 2017, for a recent discussion). Adult learners might be particularly susceptible to these test effects, as they have more experience than children with explicit language instruction and test taking more generally. In addition, while the artificial language learning paradigm controls the amount of experience participants have with the artificial language, it does not control for the amount of experience with language as such. Adult participants have had considerably more experience than children with language, meaning that they come to the task of artificial language learning with strong (albeit implicit) intuitions about what language should look like, and which cues may be relevant for generalization – this is not the case during naturalistic first language learning.

There are several things researchers can do to control the nature of learning in artificial language learning experiments. Recent work has used post-exposure questionnaires to probe participants' explicit awareness of patterns in the language (Brown et al., 2018; Wonnacott et al., 2017), and in one case this has revealed that, what looked like implicit above-chance performance on a forced-choice test, was driven by participants reporting explicit awareness performing near-ceiling, whereas those who did not report explicit awareness showed no learning (Brown et al., 2018). Another, more challenging possibility, would be to develop age-appropriate versions of paradigms which showed statistical learning with adults, and use them with children and infants. There have not been many such robust age-appropriate paradigms to date, and much of the work in the literature still remains with adults (in my review, I prioritised work with children and infants). Finally, researchers could aim to incorporate more on-line, implicit measures of learning into experiments with adults (as opposed to "off-line", explicit post-exposure tests) which would not encourage explicit strategies, and may provide an insight into the trajectory of learning. Some recent work provided promising results in this direction (Siegelman et al., 2017; Siegelman, Bogaerts, Kronenfeld, & Frost, 2018).

A second, and more fundamental implication of artificial language learning experiments is that, while humans learn and generalize input statistics, this does not, in-and-of-itself, constitute evidence that there is no innate grammar. In theory, it could be that the learners of the artificial languages map this novel input to an innate grammar. Indeed, Yang (2004) argued that the fact that 8-month-olds in Saffran et al. (1996) could extract transitional probabilities from artificial speech after as little as two minutes of exposure, is in fact evidence for innate linguistic knowledge (Saffran et al., 1996, themselves claimed that their

findings do not rule out innate knowledge). The statistical learning literature, on the other hand, is vague with regards to the critical question: if not via innate linguistic knowledge, then how exactly do humans learn language from the statistics of the input? (see Frost et al., 2019, for a discussion). While there are both empirical and theoretical reasons to be sceptical of nativist/generativist approaches (not in the least because the notion of innate linguistic knowledge is difficult, if not impossible, to falsify), the lack of a precise theoretical alternative is partly why nativist/generativist theories remained highly influential. One way to formulate more precise hypotheses about the nature of language learning from the input is to build computational models of language learning, which I turn to in the next section.

1.3 Computational modelling of probabilistic language learning and processing

Artificial language learning experiments demonstrated that humans can learn probability distributions of forms in the input, and that this can shape generalization in sophisticated ways. As I discussed in the previous section, this does not necessarily mean that humans are not doing this on the basis of innate linguistic knowledge. The appeal of computational modelling is in that the same does not hold – any learning that we see in these models is not a result of "innate" linguistic knowledge (simply due to the fact that researchers did not build any into the model), but an emergent property of learning from the input. Therefore, if computational models can learn an aspect of language entirely on the basis of frequencies in the input, then learning that aspect of language does not necessitate innate knowledge (which is not to say that computational models do not make some important assumptions about the learner – I return to this in Section 1.3.1).

Mintz, Newport, and Bever (2002) tested the possibility that syntactic categories can be abstracted purely from the probability distributions of words in language. They exposed a simple learning model to utterances directed to children under the age of 2.5 years old, obtained from the CHILDES corpus of child-directed speech (MacWhinney, 2000). Specifically, for each word in the corpus, the model recorded its "context", that is, which words occurred immediately before and after the target word, and how frequently. This was the only information available to the model – it had no information about the syntactic categories of surrounding words. The model then compared words against each other based on how similar their contexts were, and this was fed to a clustering algorithm. Clustering algorithms group data points into "clusters", such that data points within one cluster are more similar to each other than to the data points in other clusters. This analysis revealed that the resulting clusters roughly corresponded to "noun" and "verb" grammatical categories: for example, *baby, box* and *train* were clustered together, whereas *bring, eat*, and *find* were in a different cluster (and exclamations such as *oh, oops*, and *uhhuh* in yet another). Recall that the only information the model was given about individual words was which other words it co-occurred with and how frequently. From this distributional information alone, it was possible to group words into nouns and verbs with high accuracy (the authors discuss other kinds of information which may further improve the model's performance, such as

semantic, morphological, and syntactic information). Similar findings were reported with other clustering models (Finch & Chater, 1992; Redington, Chater, & Finch, 1998).

Clustering models have shown that human linguistic input contains information from which syntactic categories can be inferred reliably – something that nativist/generativist approaches deemed impossible without innate knowledge of the categories. These findings are consistent with artificial language learning experiments discussed in Sections 1.2.3.1, and 1.2.3.2, which showed that learners can form categories based on distributional information alone. What the modelling allows us to say, however, is that the syntactic categories were learned entirely by tracking co-occurrence statistics and clustering words together based on this statistics (with human learners, on the other hand, other learning mechanisms, or indeed innate knowledge, cannot be ruled out entirely).

While clustering algorithms have been successful in modelling syntactic categories from distributional cues, human language learning has been modelled most extensively using another class of computational models – connectionist models (which I introduced briefly in Section 1.2.1). The next section focuses on these models more extensively.

1.3.1 Connectionism

Connectionist models, often also referred to as artificial neural networks, model human linguistic knowledge (and cognition more broadly) as a large network of processing units which are interconnected to varying degree, much like neurones in the brain. A unit receives input from other units in the network, and the activation of that unit is determined by the strength of the connection between that unit and the units from which it receives input. Learning occurs in the network via incremental changes in the strength of the connections between units in response to the input from the environment, until the configuration of weights is such that for every activation of a set of input units, the network activates a set of units at the output layer which are as close as possible to the correct response. Therefore, learning in connectionist models is fundamentally driven by error signal from the environment. In addition to this, representations in the networks are distributed, meaning that no single category is represented with a single unit – instead, individual categories are represented though parallel activation of many individual processing units (this property of the networks has important consequences, which will become more apparent when discussing particular models). And finally, as mentioned in Section 1.2.1, no symbolic rules are built into the network, so any rule-like behaviour observed in the network is an emergent property of the way the network learns from the input.

While connectionist models have no in-built rules, this is not to say they are a *blank slate*. Learning is constrained in connectionist models in several important ways. First, input is represented in neural networks in a certain way: for example, spoken input is represented at the level of phonological features (where the word "cat" may be represented in terms of discrete units each corresponding to the features of phonemes), and semantic input in terms of semantic features. These models therefore assume that humans encode the input in a comparable way, both in the mind and the brain (in individual neurones). Second, in most connectionist networks, between the input and the output layers there can

be many "hidden" layers of units. The number of units per hidden layer, and the number of hidden layers itself, known as the "architecture" of the network, critically constrains what the network can learn. Elman, Bates, and Johnson (1998) referred to this constraint as "architectural innateness", and argued that this is comparable to the way human learning is constrained by the density of different brain cells in each layer, as well as the number and thickness of layers. Finally, connectionism makes a core assumption that the learner is sensitive to the co-occurrences in the input to begin with; this sensitivity is built into the models by virtue of the way they learn from input.

In the next section, I review some key connectionist models. Before that, I should note that, even though the idea that neuron-like units can learn via associative learning from the input goes back more than seventy years (Hebb, 1949; McCulloch & Pitts, 1943; Rosenblatt, 1958), the origins of the connectionist approach are typically associated with the technological advances in 1980's. In particular, the development of a learning algorithm known as *back-propagation* (Rumelhart, Hinton, & Williams, 1986), an algorithm which can compute error minimization over a large number of weights efficiently, allowed researchers to build computational networks with multiple layers of units, as opposed to older work which, due to technological limitations, modelled single layers. This new, more complex architecture, drastically increased the capacity of the models and the complexity of problems they can learn, from morphology to long-distance syntactic dependencies, and other aspects of language such as spoken word recognition[5].

1.3.1.1 Morphology

In a famous model of the learning of English past tense, Rumelhart and McClelland (1986) demonstrated that a neural network – with no in-built linguistic rules – can generalize morphological patterns to novel forms probabilistically, in a way that reflects the frequencies of similar existing forms in the input, much like humans do (Section 1.1.2). Rumelhart and McClelland's (1986) model received as input the phonological form of a verb stem (represented as the activation of units corresponding to the presence or absence of different phonetic features of the stem – voicing, place of articulation, etc.) and produced the verb's past tense as output. The network's output was compared to the correct form, and the strength of connection between units was adjusted in a way that minimizes error, that is, the "distance" from the correct form. The model was exposed to a representative sample of English past tense forms, and as more experience with the correct forms occurred, the model's performance gradually improved. At test, the network generalized the -ed pattern to novel forms the way humans do: when given *blug*, the network produced *blugged.* Critically, this happened without an in-built past tense rule. Instead, rule-like behaviour emerged as a consequence of learning from the input – the network was sensitive to the fact that the most frequent correct past-tense form in the input involved a form of the -ed morpheme. The most important aspect of Rumelhart and McClelland's model, however, was that it was able to replicate probabilistic generalization of irregular past tense forms,

[5]Artificial neural networks have been used extensively in other domains of cognition, most notably in vision (see Schmidhuber, 2015, for a review).

which is observed with adults and children. For example, when the model was presented with a novel form such as *spling*, the response was probabilistic — sometimes the model produced *splinged* and sometimes *splung*, but it did not apply the i – u pattern to novel verbs such as *vin* (*vun* was an unlikely past tense form of *vin*). Probabilistic generalization is a natural consequence of the distributed representations in the network (recall from the beginning of the section). For example, the stem string is represented in the network via multiple discrete units: "#st", "str" ... "ing" (each letter represented as a cluster of phonological features). Therefore, -ing at the input layer will be associated with -ung at the output layer, not just when the model encounters string, but any other verb that follows the *ing-ung* pattern, such as *swing, fling*, and *cling.* This distributed nature of representations in the network means that the model can generalize via analogy across similar forms, such that any novel form which ends in -*ing* will be more likely to produce -*ung* than a form that does not end in -*ing*. Probabilistic generalization of irregular forms is inconsistent with a rule-based approach (Pinker & Prince, 1988; Prasada & Pinker, 1993), which claim that regular past tense forms were generated by applying the rule, whereas irregular forms were listed in the mental lexicon. The rule-based approach therefore cannot capture generalization of irregular patterns, nor any kind of probabilistic generalization.

While Rumelhart and McClelland (1986) were able to replicate the probabilistic nature of past tense generalizations in English with a model that learned through exposure to correct forms without explicitly in-built rules, it is unclear whether a connectionist network could capture the inflectional morphology of languages with, for example, more inflection than English (and there are many such languages). Not many connectionist models to date investigated this, with some notable exceptions – Hahn and Nakisa (2000) modelled German plurals; Mirković et al. (2011) modelled noun inflections in Serbian, a highly inflected language with three genders, two numbers and seven cases; Engelmann et al. (2019) modelled Finnish and Polish past tense, both highly inflected languages. These networks performed comparatively to adult judgments and children's productions, and captured effects such as frequency and *phonological neighbourhood density effects* observable with humans, suggesting that connectionist networks can learn at least some aspects of the morphology of languages more complex than English in that regard, though more extensive replication is required.

A more fundamental limitation of neural networks generally is their over-reliance on trained instances. While Rumelhart and McClelland's model generalized to novel forms, it also made errors with certain forms that could not be explained in terms of frequencies in the input: for example, the model produced *membled* as past tense of *mail*, or *treelilt* as past tense for a nonce verb *trilb* (Rumelhart & McClelland, 1986) This is presumably because these verbs were not similar to many other verbs in the model's training input. However, there is no evidence that children make such random errors. Subsequent models eliminated the problem of odd errors (e.g., Daugherty & Seidenberg, 1992; Plunkett & Juola, 1999) by adjusting the architecture of the networks and changing the way phonological information is represented. Note also that all the networks which learned the morphology of more inflected languages mentioned previously, involved multiple layers and relatively complex

architecture. Adding architecture to the model means that, if our goal is to model human learning, we are also proposing that a similar psychological function exists in the human mind, but it is not clear how psychologically motivated these changes to the networks are. I return to these limitations in Section 1.3.2.

1.3.1.2 Hierarchical structure

Another aspect of language learning that has been argued to be impossible without innate linguistic knowledge is the learning of long-distance dependencies, such as number agreement between the head-noun and main clause verb. For example, in the sentence *The boy who chases dogs which chase cats runs fast*, the verb runs agrees with the head-noun the boy – this is known as *hierarchical structure*, in that a phrase is contained within another phrase. Learning of long-distance dependencies has been traditionally thought to require some kind of domain-specific knowledge. Connectionist models, however, demonstrated that a system which keeps track of the probability of the occurrence of different linguistic forms can correctly process some long-distance dependencies. Elman (1993) trained an artificial neural network to predict the next word in a sequence – the current word was represented in the input units, and the next word was represented in the output units. As in all connectionist models, the model's response was compared against the correct response and this generated the error signal necessary for learning. The model also contained "context" units which contained a copy of the state of hidden units at the previous word – serving as "memory" of the network. At test, the model rejected **The boy who chases dogs which chase cats run fast* as ungrammatical, demonstrating that the model learned long-distance syntactic dependencies, entirely based on the statistics of word co-occurrence (see Servan-Schreiber, Cleeremans, & McClelland, 1991; John & McClelland, 1990, for related models of sentence processing). However, the extent to which these networks really learned hierarchical structure, as opposed to unstructured strings of categories, has been debated (Steedman, 2002). In addition, Kam, Stoyneshka, Tornyova, Fodor, and Sakas (2008) found that the same modelling approach – where the network simply remembers bigrams and trigrams of words – was not sufficient for learning question formation in Dutch.

1.3.1.3 Other domains

Connectionist models have been used successfully in other aspects of language. One famous example is the McClelland and Elman's TRACE model (McClelland & Rumelhart, 1986) of spoken word recognition, whcih was able to capture top-down effects in speech processing observed with humans. In this model, auditory, phonemic, and word-specific information is presented in a separate layer each. The network receives auditory input sequentially, and as this happens, activation is propagated to the phonemic layer, where the corresponding phonemes are activated, which in turn activate the unit corresponding to the word in question at the output layer. Importantly, activation at word-level can project back to activation at phonemic level. A natural consequence of this is that the model can capture the top down effects in speech processing, such as the phoneme restoration effect: when the model is presented with *b#eak* (where features of /r/ in break are replaced with random

noise), this partial input activates break at word-level, which in turn will project back to /r/ at phoneme level (represented as a cluster of phonological features) even though /r/ was never received in the input. Similarly, the model captures top-down effects in phonetic categorisation (see Ganong, 1980, for evidence with humans): when the model was presented with a phoneme between /d/ and /t/ in terms of feature activation, followed by -_ask, the model activated /t/ at phoneme-level. In TRACE, this occurred because thee acoustic input activated the word-level representation of task, which in turn reinforced the activation of /t/ at the phoneme level.

Connectionist models have been successful in capturing effects in other aspects of language, such as typical and atypical reading (Harm & Seidenberg, 1999, 2004) and verb morphology data from patients with brain lesions affecting language (Joanisse & Seidenberg, 1999).

1.3.2 Evaluating modelling approaches

Recall that when I introduced computational modelling of language learning as a window into the precise learning mechanism, I argued that the appeal of computational modelling is that the researcher has a complete insight into how learning occurs in the model, and can attribute learning to particular aspects of the model. Despite their wide-spread use due to their success at learning complex problems, the extent to which connectionist models allow the researcher to trace learning back to a component of the network is limited. Connectionist models often involve multiple hidden layers of units, and it soon becomes difficult to track the dynamics of weight updates as activation propagates through such large and complex networks. This makes it difficult to know how and why exactly the network chooses a particular output for a given input (Smolensky, 1988). As the field developed, connectionist modelling became increasingly more concerned with demonstrating what sufficiently complex networks can do, with insufficient emphasis on generating predictions about when and why learning does and *does not* occur (McCloskey, 1991). As M. Jones and Love (2011) described it, "the theoretical position underlying connectionism was thus reduced to the vague claim that the brain can learn through feedback to predict its environment, without a psychological explanation being offered of how it does so" (p. 172).

While connectionism may not be a theory of language learning in its own right, this is not to say that connectionism has not made a significant contribution to our understanding of language learning. Connectionist models showed that aspects of language were learnable without explicitly in-built rules and abstract categories, and were able to capture many psycholinguistic effects, such as probabilistic generalization on verb inflections or top-down effects in spoken word recognition. Seidenberg (1993) argued that a lot of the criticism of connectionism stems from disagreement about the goal and purpose of cognitive modelling. Rather than providing a theory of language, he proposed that connectionist models provide a set of computational principles, such as distributed representation, backpropagation, and the role of input-frequency, which constrain learning in systematic ways. Seidenberg argued that these principles, as long as they are neuropsychologically plausible (though see Crick, 1989, on neural implausibility of backpropagation), can be the necessary independent basis

for different exploratory theories, and that connectionist model can serve a particularly valuable role of providing a close link between theory and data.

Note that, while in this review I focused on connectionist models (largely due to their wide-spread application and impact), this is not the only approach for modelling language learning (and other aspects of cognition). One other framework, which is becoming increasingly popular, the Bayesian framework, views human learning as the process of approximating near-optimal rational inference, as captured by a mathematical identity known as Bayes' rule. Bayesian models have successfully captured human data in a wide range of contexts, from word segmentation (Goldwater, Griffiths, & Johnson, 2009), word learning (Xu & Tenenbaum, 2007; M. C. Frank, Goodman, & Tenenbaum, 2009) and verb argument structure (Perfors, Tenenbaum, & Wonnacott, 2010), to pragmatic reasoning (M. C. Frank & Goodman, 2012). While the details of this work are beyond the scope of this review, I point out that the evidence that human learning is the process of achieving near-optimal rational inference has been questioned (Bowers & Davis, 2012; Marcus & Davis, 2013) (see Griffiths, Chater, Norris, & Pouget, 2012; Goodman et al., 2015, for responses to the critiques, respectively), and similarly the theoretical motivation for many architectural aspects of Bayesian models has been questioned (M. Jones & Love, 2011). Therefore, not unlike connectionism, the explanatory power and theoretical contribution of Bayesian models in understanding the underlying mechanism of human language learning has been subject to criticism.

So far in this review, I presented evidence that human language use is probabilistic – that is, fundamentally shaped by the statistical distributions of linguistic forms. This notion is incorporated into several different theoretical approaches to language learning (Section 1.2.1), which argue that children learn language gradually, through repeated use and exposure to language, and that this process is underpinned by an ability to learn the statistics of the input. Studies of early child language, as well as lab-based studies of artificial language learning, provided evidence that learners use statistical information in the input – such as the probability of occurrence and non-occurrence of different forms – to generalize to novel forms. However, as I discussed in Section 1.2.3.4, while learners' generalization patterns may *look like* they are driven by probabilistic learning, in the absence of a model of the underlying learning mechanism (which makes specific predictions about when learning does and does not occur), it is difficult to attribute learners' generalization to a specific learning process and rule out alternatives. While computational models of language learning, and connectionist models in particular, have been able to show that different language tasks can be learned without in-built knowledge of rules and abstract categories, the architecture of the models makes it difficult to identify the underlying learning mechanism. Importantly, different modelling approaches have been criticised for insufficient theoretical grounding in the psychology of human learning, which makes it difficult to evaluate the extent to which learning in the models reflects human learning. Therefore, approaches to language learning reviewed so far in this introduction, which incorporate the idea that language is probabilistic, have not been specific enough about the underlying learning mechanism (the statistical learning literature), the mechanism does not allow the researcher to make pre-

cise predictions due to its complexity (connectionist models), or is not clearly based on empirical evidence about the psychology of learning (Bayesian approaches).

For the reasons discussed above, this thesis adopts an alternative approach to language learning, the discriminative learning framework, put forward by Ramscar and colleagues (Ramscar & Yarlett, 2007; Ramscar et al., 2010; Ramscar & Dye, 2011; Ramscar, Dye, & Klein, 2013). This approach avoids the main pitfalls of the approaches discussed so far, in that it is theoretically based on core principles of learning, and, when implemented computationally, is simpler and more interpretable than traditional connectionist networks. The following section introduces the discriminative learning framework in detail.

1.4 The discriminative approach: Language learning as uncertainty reduction

The discriminative learning approach is based on insights from learning theory and information theory, and it views language learning a predictive process of reducing uncertainty about outcomes by discriminating informative from uninformative cues in the environment. Mechanistically, discriminative learning is driven by prediction error and cue competition. The notions of prediction error and cue competition date back to classical learning theory, where psychologists spent decades carefully studying learning in animals. To properly introduce the discriminative learning approach which is at the core of this thesis, it is therefore necessary to revisit this early work.

1.4.1 Principles of learning

Ivan Pavlov famously showed that, if a bell rings as dogs are given food, the dogs soon start to salivate when hearing the bell, even when no food is given (Pavlov, 1927). While highly influential, this seminal finding gave rise to a misconception that tracking simple frequencies of co-occurrences between cues (such as a tone, light, etc.) and outcomes (such as food, shock, etc.) is necessary and *sufficient* for learning. This is a fundamental misconception – so much so that Rescorla (1988) titled his paper, in which he argued against this view of learning, as: "Pavlovian conditioning: It's not what you think it is". Indeed, this simplified (mis)understanding of the dynamics of learning is one of the reasons why Chomsky (1959) famously rejected learning theory as a model of language learning. While some of the work discussed in the previous chapter (specifically, Sections 1.2.3.1 and 1.2.3.2) shows that participants in artificial language learning experiments do more than computing transitional probabilities, this was not explicitly incorporated into the statistical learning literature, and this fundamental misconception about the nature of associative learning remained largely unchallenged in the field of language learning.

However, in the area of animal learning, psychologists accumulated a coherent body of evidence which showed that the dynamics of learning are more complex than the mainstream view of "Pavlovian conditioning" suggested. This work showed that, rather than being driven by the probabilities of co-occurrence, learning is a predictive process of discriminating between informative (predictive or discriminating) and uninformative (unpredictive

or non-discirminating) cues. Importantly, this process has three critical and closely related characteristics:

1. **Learning is error-driven:** learning only occurs when events violate predictions.

This effect was first demonstrated by Kamin (1968), who trained rats on a pairing of two cues – a light and a tone – followed by a mild electric shock. Importantly, prior to the training, half of the rats were pre-exposed to only one of the two cues (e.g., tone) followed by the shock, and the other half received no such pre-exposure. At test, rats were exposed to one of the two cues in isolation. If the rats have learned to associate the cue with a shock, they should freeze in fear. The rats who were not pre-exposed showed the fear response, whereas those who were pre-exposed did not show the fear response to the cue they had not been pre-exposed to (e.g., light). This result, known as the blocking effect, shows that learning only occurs when a cue provides new information to the learner, or, as Kamin (1968) put it, when the event is somehow surprising to the animal. Rescorla and Wagner (1972) re-phrased this effect as: "organisms only learn when events violate their expectations" (p. 75). Had, for example, the presence of the light with the tone predicted the absence of the shock, this would be surprising to the animal expecting a shock (based on the tone), and learning about the light would occur. However, the light did not provide any additional, new information about the occurrence of the shock, and therefore no learning occurred. Importantly, any account of learning in which the statistics of co-occurrence alone predict learning would fail to explain this result: the shock had the same conditional probability following the tone and the light, that is, whenever the tone occurred, it was followed by the shock, and whenever the light occurred, it was followed by the shock. Therefore, learning is driven by the informativeness of a cue, not by the statistics of its occurrence with the outcome. Note that Kamin (1968) did not explicitly propose (and test) the conditions which do produce the surprisal (the example above about the light predicting the absence of the tone is mine). Importantly, however, this study was the first to demonstrate the blocking effect, which was the basis of much of the work that followed in the field.

2. **Background rate**: the informativeness or the predictive value of a cue is not determined only by the number of times the cue occurs with the outcome, but by that number relative to the number of times the cue occurs in the absence of the outcome.

This idea was first proposed by Rescorla (1967) in a discussion of the Pavlovian classical conditioning paradigm. In subsequent work, Rescorla (1968) demonstrated this empirically by training one group of rats on a tone followed by a mild electric shock. Another group of rats were exposed to the same number of tone-shock pairs, but these were interspersed by tones which were not followed by shocks. Only the rats in the first group showed conditioning. Importantly, both groups were exposed to the same number of tone-shock pairings. What differed between the conditions was the predictive value of the tone with respect to the shock. Only in the first condition was the tone predictive of the shock; in the second condition, the tone was not predictive of the shock, and therefore no conditioning occurred,

showing that learning is driven not by the frequency of cue-outcome co-occurrence, but by the predictive value of the cue. The critical point here is that the predictive value of a cue cannot be determined without negative evidence: a cue and an outcome may co-occur frequently (positive evidence), but if the cue also frequently occurs in the absence of the outcome (negative evidence), an association between that cue and the outcome will not be formed (note that this has been considered in the statistical learning literature, though not in the context of cue competition and prediction error; see Aslin et al., 1998).

3. **Cues compete for predictive value**: when the outcome occurs in the absence of the cue, this increases the predictive value of whatever other cues are present in the environment, at the expense of the target cue.

Wagner (1969) demonstrated this by training a baseline group (Group 1) of rabbits on a compound stimulus consisting of a tone and a flashing light, which was always followed by a mild electric shock to the eye. Two additional groups were trained on the exact same schedule, except that it was interspersed with trials on which the tone occurred alone, and either was followed by a shock (Group 2), or was not (Group 3). As expected, during the exposure phase, the number of blinks (indicating fear conditioning) to the tone alone was considerably higher in Group 2 compared to Group 3 – the rabbits in Group 3 learned that the tone alone was not predictive of the shock (much like the rats in Rescorla, 1968). The key question, however, was whether the reinforcement/non-reinforcement of the tone alone would affect the predictive value of the light. To test this, Wagner presented the rabbits with the light alone. A striking pattern of results emerged: relative to the baseline, the number of conditioned responses to the light was lower in Group 2, and higher in Group 3, despite the fact that every group received the same number of exposures to tone-light pairs and that the light was never encountered in isolation. This study is a compelling demonstration of cue competition. Specifically, the trials on which the shock occurred after the tone increased the predictive value of the tone at the expense of the light, as evidenced by a weaker conditioned response to the light in Group 2 compared to the baseline; the trials on which the tone was presented alone and a shock was expected but it did not occur decreased the predictive value of the tone to the benefit of the light, as evidenced by a stronger conditioned response to the light in Group 3 compared to the baseline. The study also provides further evidence for the two points raised above. First, that learning is fundamentally error-driven: the light was associated with the shock only in the condition in which the shock was expected but did not occur (Group 3). Second, temporal co-occurrence is not necessary for learning: the occurrence of the shock following the tone alone decreased the predictive value of the light (Group 2); in other words, the predictive value of a cue with respect to an outcome changed even though the cue did not occur with the outcome.

The work reviewed in this section constituted some of the earliest evidence that learning is an error-driven, discriminative, competitive process. Learning depends on the discrimination between informative and uninformative cues, and this is facilitated by prediction error and cue-competition. These key insights were demonstrated in numerous other studies in animal learning (Annau & Kamin, 1961; Egger & Miller, 1962; Konorski, 1948; Mackintosh,

1965; Reynolds, 1961; Wagner, Logan, & Haberlandt, 1968), and were captured formally in the Rescorla-Wagner mathematical model of learning (Rescorla & Wagner, 1972), which has become one of the most widely used models of learning in psychology (see Siegel & Allan, 1996, for a review). This model is also at the core of the discriminative learning approach. Below I give a brief conceptual overview of the model, but I introduce the model in technical detail in Experiment 1 (Section 3.2.1), and discuss its predictions at length throughout the thesis.

The essence of the Rescorla-Wagner model is that it learns to predict outcomes from input cues. As learning unfolds, cues become more or less predictive of the outcome; ultimately, only those few cues which are consistently present when the outcome is present (predictive cues) will have strong positive weights, whereas the cues which are consistently present in the absence of the outcome (unpredictive cues) will have strong negative weights. Critically, learning in the model is driven by prediction error and cue competition, as described in the studies of learning above. Error-driven learning is not unique to the Rescorla-Wagner model. In fact, the connectionist models discussed earlier (Section 1.3.1) all implement the same error-minimization algorithm (see Section 3.2.1. for detail). What sets the Rescorla-Wagner model apart is its simplicity: unlike more traditional connectionist models, the Rescorla-Wagner model is a simple two-layer network without hidden units. This fairly simple architecture has been able to capture numerous psycholinguistic phenomena, such as frequency and neighbourhood density effects in morphological processing (Baayen, Milin, Djurđević, Hendrix, & Marelli, 2011), frequency and similarity effects in lexical decision latencies (Milin, Feldman, Ramscar, Hendrix, & Baayen, 2017), and difficulties with the learning of grammatical gender with L2 learners (Arnon & Ramscar, 2012). However, these studies did not merely show that a computational model can replicate psycholinguistics effects with accuracy comparable to (or better than) existing models, but they re-interpreted these effects in terms of a parsimonious theory. This is possible because the Rescorla-Wagner learning rule is an implementation of a specific psychological theory, and its simple architecture allows the researchers to attribute its performance to components of that theory – something which I argued earlier in this review has been a stumbling block for other computational approaches.

1.4.2 Integrating learning theory with information theory

The key notions from learning theory are related to information theory (Shannon, 1948), a mathematical theory of communication. Under this theory, communication occurs in inherently noisy channels; the goal of communication is therefore for the decoder to be able to reproduce, with the least amount of error, a source message that was transmitted by the encoder through the noisy channel. In other words, the goal of optimal communication is to reduce the uncertainty about the source message. Importantly, the amount by which the uncertainty about the source message is reduced is quantified using entropy, a mathematical measure of the amount of information contained in a message (in bits), expressed as the negative logarithm of any probability mass function. The logic behind entropy is that low-probability events carry more information (and therefore have higher entropy) than

high-probability events, which are more predicable and therefore less informative.

In a re-analysis of learning theory in information-theoretic terms, Gallistel (2002) reformulates Rescorla's insights (Rescorla, 1967, 1968) about the role of predictive value in learning (rather than the statistics of co-occurrence) as follows: what determines learning is not a frequent individual pairing of events, rather, how information is structured in a flow of events. Gallistel (2002) further argues that Kamin's notion of surprisal (Kamin, 1968), that is, the amount of information a cue provides, can be quantified using entropy. While information theory and learning theory and highly compatible and overlapping, insights from these two literatures have not been explicitly and formally combined in a unified theory of human language which would encompass learning, processing, and human communication more broadly.

However, there *has* been a revival of interest in information theory in cognitive science in the last two decades. For example, numerous studies have demonstrated that the amount of information (entropy) conveyed by a word predicts the amount of cognitive effort involved in processing the word (Hale, 2003; Levy, 2008), as evidenced by distinguished event-related potentials (S. L. Frank, Otten, Galli, & Vigliocco, 2015) and longer reading times (Boston, Hale, Vasishth, & Kliegl, 2011; S. L. Frank, 2013; N. J. Smith & Levy, 2013) for words and parts-of-speech with high information content. In the domain of language production, it has been shown that speakers distribute information relatively uniformly across the utterance, so that peaks and dips in the amount of information are smoothed over the signal, in order to facilitate optimal processing. For example, a number of studies found that those elements that are less predictable in context are articulated more slowly, and with more phonological and phonetic detail than those that are more predictable (Aylett & Turk, 2004; Pluymaekers, Ernestus, & Baayen, 2005; Van Son & Van Santen, 2005). Aylett and Turk (2004) interpret this finding in terms of entropy smoothing. Specifically, they argue that elements which are less likely in a given context have high linguistic entropy; elements which are articulated slowly have low acoustic entropy. By articulating low-probability words more slowly, the speaker (implicitly) compensates for the high linguistic entropy with low acoustic entropy. This means the entropy is smoothed over the linguistic item itself, and the speed of articulation, making the otherwise low-probability item easier to process. Similarly, in an analysis of a corpus of spontaneous speech, Jaeger (2010) found that speakers choose to include or omit the *that* complementizer in relative clauses depending on the information content of the utterance – the complementizer is more likely to be included when the subsequent relative clause has high information content (high uncertainty) (see also A. F. Frank & Jaeger, 2008; Jaeger & Levy, 2007). A related idea has been that linguistic structures self-organise in a way that distributes information (amount of uncertainty) evenly across the linguistic signal. For example, Piantadosi, Tily, and Gibson (2011) demonstrated across ten languages that, rather than by frequency, word-length is more strongly predicted by the amount of information the word conveys, such that words with high entropy tend to be longer than words with low entropy (see also Futrell, Mahowald, & Gibson, 2015). Similarly, distributions of canonical word order across languages of the world have been interpreted to be optimised in terms of uncertainty reduction (Gibson et al., 2013; Maurits, Navarro,

& Perfors, 2010).

The information-theoretic approaches to language have offered principled analyses of various phenomena in human language use, and have demonstrated that humans are sensitive to, and implicitly manage information structure in language. Under these approaches, language learning is conceptualised as the process of converging on a shared probabilistic model of language to be used to optimally encode meanings as linguistic signal, and decode indented meanings from the signal. However, it is important to note that applying information theory to language directly skips an important step – information theory is not a theory of language, and the extent to which human communication is like the communication systems Shannon (1948) described has not been determined systematically (see Shannon, 1956; Ramscar, 2019). In addition, information-theoretic approaches have not yet developed a testable theory of how exactly the learner comes to encode and decode messages in these optimal terms – note that this is similar to the pitfalls of the statistical learning literature I discussed in earlier sections. On the other hand, the discriminative learning framework explicitly incorporates the information-theoretic idea of uncertainty reduction into human language learning and use, but unlike information-theoretic or other probabilistic approaches, it is underpinned by the core principles of learning theory (outlined above), where learning is an inherently predictive, discriminative, competitive process. Viewed in this way, rather than using direct form-to-meaning mappings to decode meanings from the forms of the messages (as assumed by applications of information theory to language), learners use linguistic forms to reduce the uncertainty about the indented meaning, by discriminating between informative and uninformative cues both in the environment and in the utterance itself[6]. The next section reviews how the discriminative learning framework has combined learning theory and information-theoretic concepts to generate specific predictions about human language learning.

1.4.3 Applying discriminative learning to language learning: order effects

Learning theory, the basic principles of which are modelled in the Rescorla-Wagner model (Rescorla & Wagner, 1972) is sensitive to the fact that the order in which events occur over the course of learning has important consequences for what is learned (see Milin, Divjak, Dimitrijević, & Baayen, 2016, for a discussion). This has allowed researchers within the discriminative learning framework to use this model to formulate precise predictions about order effects in language learning. I review this work in this section, although I note here that, even though this section focuses on the work of Ramscar and colleagues, who formulated a comprehensive theoretical framework, they are not the only ones to use notions of cue competition and prediction error in language learning (see Ellis, 2006).

In the context of word learning, Ramscar et al. (2010) demonstrated that the order in which learners encounter different cues has drastic consequences for what is learned.

[6]Form-to-meaning mappings, and related notions such as compositionality and concepts, have been difficult to explain in a principled way by linguists studying meaning (Ramscar, 2010; Ramscar & Robert, 2015). The discriminative approach avoids resorting to these notions.

Specifically, they started by pointing out that word learning can logically take one of the two directions: (i) predicting features in the environment from verbal labels (feature-label learning) or (ii) predicting verbal labels from features in the environment (label-feature learning). Whilst these appear similar on the surface, qualitatively different learning occurs under these scenarios because only (ii) allows cue competition. 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 snou*t; behaviours: for example, *barks, wags tail*, etc.) compete for predictive value for the label *dog.* Over time, uninformative cues will lose predictive value: whenever the learner encounters a furry animal that is not a dog, s/he is provided with negative evidence which down-weights the cue *furry*, and, in turn leads to an increase in predictive value of other, more informative cues (as demonstrated by Wagner, 1969, see Section 1.4.1). In contrast, if the learner first hears the label *dog* and then predicts its "meaning", there is no comparable cue competition over the features in the environment. Therefore, 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.

Ramscar et al. (2010) tested these predictions in an artificial language learning experiment, in which adult learners were taught category labels for novel items. In one condition, participants first saw a novel object and then heard the 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 (feature-label condition): only in this condition could participants have the opportunity to view a wug, and expect to hear *dep* (because the exemplar had same body shape as most of the deps), and then, upon hearing *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 (label-feature condition), uninformative cues cannot be "unlearned" as there is no negative evidence: every time the label *wug* is heard, 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 feature-label 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., low type-frequency wugs). This result was also confirmed in a computational simulation which implemented the Rescorla-Wagner model (Rescorla & Wagner, 1972), and was provided with the same input structure as human learners.

Ramscar et al. (2010) further corroborated the finding that feature-label versus label-feature ordering affects word learning in an experiment in which two-year-olds were taught colour labels. Colour labels were either taught using post-nominal constructions such as

"This cup is blue", where multiple features of the cup compete for predictive value for blue, or with prenominal constructions, such as "This is a blue cup", where no comparable cue competition was available. When asked to select a blue object from an array of objects at test, the two-year-olds were better at this if they had been trained with post-nominal constructions, demonstrating that effects of prediction error on learning can be observed in children as young as two. Note that this finding is consistent with other work which observed facilitatory effects of prediction error in child language development. For example, Reuter, Emberson, Romberg, and Lew-Williams (2018) showed that 1-to-2-year-olds' ability to update predictions about the upcoming stimulus is predicted by vocabulary size. Children were shown an object in the centre of the screen which reliably preceded a target object appearing on the periphery of the screen – after some time, children showed anticipatory looking to the periphery of the screen following the fixation object. Half-way through the experiment, however, the target started appearing on the opposite side of the screen. Children's ability to switch anticipatory looks to the other side of the screen was positively correlated with their vocabulary size. Importantly, as the task was entirely non-linguistics, this gives some indication about the direction of the relationship – prediction-error may facilitate vocabulary learning, rather than the other way round (otherwise in a linguistic task, the child may be better at predicting the upcoming word because of better vocabulary). Related evidence comes from Havron, de Carvalho, Fiévet, and Christophe (2019), who exposed French speaking 3-to-4-year-olds to sentences in which nouns were either followed by another noun or another verb. The nouns of interest were those which in French were more frequently followed by a noun, but in the experiment were followed by a verb. The key question was: can children update their expectations about the category of the upcoming word following limited exposure to evidence which violates their expectations? The study found that they can – when trained words were followed by novel words (and therefore children had no evidence as to whether these words are nouns or verbs), children looked at the depictions consistent with verb-interpretations, even though in general that trained noun is more frequently followed by a noun in their language. This suggests that children can rapidly update predictions and use them to facilitate processing of novel instances. This work is in line with broader work on the predictive nature of early language development (e.g., Gambi, Pickering, & Rabagliati, 2016; Thothathiri & Snedeker, 2008; Waxman, Lidz, Braun, & Lavin, 2009).

Returning to the finding of Ramscar et al. (2010), whereby feature-label learning facilitates generalization, this raises the question of the role of discriminative learning when labels are predicted not from features in the environment (objects, events), but from the preceding parts of the utterance. Dye, Milin, Futrell, and Ramscar (2017) addressed this question in the context of gender paradigms in natural language. They demonstrated that nouns, being the most diverse part of speech (in most languages), pose a particularly challenging discrimination problem – at any point when a noun occurs, the number of possible alternatives is higher than for any other part of speech, meaning that individual nouns tend to have higher entropy than other parts of speech. In a corpus study of article+noun pairs in German, Dye et al. (2017) showed that the entropy of the individual nouns is reduced by

the gendered article. 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 article. Therefore, total uncertainty is smoothed over the article+noun pair, making the noun easier to process (see Section 1.4.2 for related work in processing). Eye-tracking experiments show that this entropy-reducing property of gendered articles is readily exploited during language processing by both adults (Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Lew-Williams & Fernald, 2010) and children as young as three (Lew-Williams & Fernald, 2007), who, as soon as they hear the gendered article, direct their gaze towards the item of that gender from an array of items. For example, Dahan et al. (2000) showed speakers of French arrays of objects on-screen, and played sentences in which the prenominal article either provided gender information (e.g., *cliquez sur le button* – click on the button (masculine)) or did not. When the prenominal article did not provide gender information (e.g., *cliquez sure les buttons* – click on the buttons), participants fixated pictures whose names shared initial sounds as the target picture (e.g., *bouteilles* (feminine)) more so than those which were phonologically unrelated – the so-called "cohort" effect (Marslen-Wilson, 1987). However, when the article ruled out gender inconsistent pictures, participants did not fixate those pictures despite the fact that they shared the same initial sound (thus eliminating cohort effects), suggesting that participants were able to use preceding grammatical information to constrain processing in real-time. Note that languages without gendered articles have other means of reducing uncertainty of the nouns. Dye, Milin, Futrell, and Ramscar (2018) showed that in English, a language without gender agreement, pre-nominal adjectives serve the same purpose of uncertainty reduction. For instance, 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.

To summarize, a diverse body of experimental, computational, and corpus-based evidence suggests that the order in which information is presented during learning and processing enables qualitatively different learning, which is explained in terms of uncertainty reduction through discriminative learning mechanisms – cue competition and prediction error. This thesis further tests the effects of order on language learning.

1.5 This thesis

In this thesis, I consider the effects of order on learning, as framed by the discriminative learning approach, with respect to a test-case which has historically gained attention from various approaches to language – differences between suffixes and prefixes in inflectional morphology. More detail is given in the introduction of Study 1 (Section 3.1), but here I note that, in this thesis, *suffixing* is taken to be analogous to feature-label learning described in Ramscar et al. (2010), whereas *prefixing* corresponds to label-feature learning. From this I predict that, in a series of artificial languages in which adult learners are taught languages which are "suffixing" or "prefixing", participants in the suffixing condition will show better

generalization of correct informative features to new instances, similar to better feature-label learning in Ramscar et al. (2010). Participants in the prefix condition, on the other hand, will be better at learning the trained items, because the prefix smoothes entropy over the utterance, as proposed by Dye et al. (2017, 2018). These predictions are discussed in more detail in Study 1 (Section 1.3).

The predictions will be tested using artificial language learning (introduced in Section 1.2.3). In all experiments in the thesis, adult native speakers of English will be exposed to artificial languages in which "nouns" are preceded by an "affix" (the prefix condition), or followed by an "affix" (the suffix condition). Each noun+affix bigram is uniquely matched to a novel referent (a picture of an alien-like creature or a novel object). In Studies 1 and 2, nouns and referents are matched with each other and with the affix based on semantic and phonological cues (these cues are deterministic in Study 1, and probabilistic in Study 2). In Study 3, nouns and referents are such that there are no obvious similarities in the visual and acoustic qualities of the nouns and referents which occur with the same affix. Across these three studies, however, the interest is in how well participants can learn: (1) which nouns and affixes occurred with which referents in the exposure set, and (2) whether the learners can generalize the patterns in the input to novel items (Studies 1 and 2). Specifically, based on the discriminative learning theory and previous work, I predict that the suffix condition will provide more cue competition, and thus greater error signal, and that this will allow the learner to discriminate informative from uninformative cues. This should be evidenced by better generalization in the suffix condition compared to the prefix condition: when presented with a new noun and/or a referent, participants in the suffix condition will be able to choose the correct affix based on the key informative features that the novel noun/referent will contain. In the prefix condition, on the other hand, the prefix will reduce the entropy of the upcoming noun and referent, which will make the noun easier to process than in the suffix condition. This should be evidenced by better item-learning in the prefix condition compared to the suffix condition: when presented with a trained noun, participants in the prefix condition will be more accurate at selecting the correct referent than in the suffix condition.

1.5.1 Methodology

Throughout the thesis, learning was studied using artificial languages, however, across the studies, different methods of exposing participants to the artificial language as well as different ways of testing learning was explored. As is usually the case in artificial language learning experiments, participants were not given feedback nor any kind of explicit instruction at any point in the experiment in order to mimic first language acquisition.

Note that, as I discussed in Section 1.2.3.4, there are some methodological concerns with artificial language learning experiments, one being that adult learners may be using explicit strategizing to learn the patterns in the input. While this is not something that can be completely inhibited, steps were taken in study design to at least discourage explicit learning and gain some control over the nature of participants' learning. First, participants were instructed not to try and "work out" the rules in the language, and instead to focus on the

sounds and the pictures (while this is intended as a precaution, there is no guarantee that participants will follow these instructions). To gain some insight into what strategies (if any) participants used in the experiment, all studies included a post-experiment questionnaire which probes participants' explicit beliefs about the structure of the artificial language. The responses from the questionnaire were used to supplement the analysis of generalization tests in an exploratory way – here, of interest was to ensure that any differences between conditions that we may observe were not an artefact of there being more explicit learners (presumably due to chance sampling) in one condition than another. The questionnaire is, however, interpreted with caution, as it is an indirect indication, rather than a precise and robust measure of explicit learning. Second, in Studies 2 and 3, speeded training was used, as well as referents which are harder to verbalise – both of these measures should discourage explicit learning. Another concern with our methodology is that this work involves adult learners, who bring implicit and explicit knowledge of how language works, which is strongly informed by years of experience with their first language, and possibly also with formal instruction of a second language. This should be borne in mind given that the predictions tested in this thesis are made with respect to a naïve learner, rather than participants with intuitions about how informative different cues are likely to be. I return to this point in Chapter 7, but here point out that evidence for the key predictions would suggest that the effects of order on generalization and item-learning are robust to other (possibly conflicting) biases that learners might bring to the task. There are other limitations to this method, as with all methods; these are discussed at length in Chapter 7.

All but one experiment in this thesis were done on-line – all participants were recruited via Prolific Academic, a third-party online recruitment platform, and they completed the experiment on-line, remotely, rather than in the lab. All the online experiments were hosted on Gorilla Experiment Builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2019). Online methods were chosen primarily to allow the recruitment of large samples. This is particularly important given recent methodological concerns in the field regarding much of the publish work being underpowered (Maxwell, Lau, & Howard, 2015). Appropriately sampled studies are therefore necessary to minimize Type I error resulting from low statistical power. Online data collection allows us to reach large numbers of participants, and do so quickly – this is particularly valuable given the time constraints associated with doctoral programs. In addition to this, online data collection can give access to monolingual native speakers of English, which has proven difficult though the UCL Subjects Pool which has access to UCL students, many of whom are international, as well as to local residents who also tend to be multilingual.

1.5.2 Outline of the thesis

This thesis uses a method of statistical inference which is not widespread in the psychological literature – the Bayes factor. The method for computing Bayes factors in this thesis is also relatively novel, and therefore the next chapter (Chapter 2) introduces this statistical approach in detail.

Chapter 3 presents Study 1, in which two relatively small sample experiments were

carried out to test the learnability of a paradigm which was developed to be used with child learners in future work. This work showed evidence for better item learning in the prefix condition; however, generalization was also better in the prefix condition, which was the opposite of what is predicted both by the theory, as well as by the performance of two computational simulations presented in Experiment 1 of that chapter.

Several aspects of the design of Study 1 were identified which may have caused the unpredicted prefixing advantage in generalization, and therefore in Chapter 4 (Study 2), a different paradigm was used to test the same predictions. Specifically, as Study 1 showed a high degree of explicit learning, the paradigm was made faster in order to discourage explicit learning. While this was a necessary modification, it made the paradigm unsuitable for child learners, and the plans for carrying out work with children were placed on hold until a robust paradigm that works with adults in the first instance has been developed. In addition to the change in the speed of stimuli presentation, Study 2 used a different design, in which some of the highly frequent, salient cues were actually unpredictive. Three large-scale experiments showed that the suffix condition was better at generalization when this required the "unlearning" of highly frequent but unpredictive cues; in the prefix condition, on the other hand, this unlearning did not occur (as indicated by chance-level performance), due to a lack of prediction error from cue competition in that condition. However, the final experiment in that study found that learners in the suffix condition also performed better on generalizations that did not require the "unlearning" of frequent, unpredictive cues. Together with the prefixing advantage in generalization from Study 1, these findings were inconsistent with the predictions of the discriminative learning model. This is why in Chapter 5, I revisited the model, and identified several ways in which the original predictions may have been incorrect. While the modelling in Chapter 5 indicates that the unexpected findings in Studies 1 and 2 may in fact be predicted by the discriminative learning model, this of course is a post-hoc interpretation of the findings. I note this and suggest that future work should replicate Study 1 and Experiment 6 from Study 2 in order to test this possibility.

Study 2 showed no evidence for better item learning in the prefix condition. This is contrary to the prediction, and therefore in Study 3 (Chapter 6) I follow up on this in two artificial language learning experiments in which within-category variation between individual items was increased to make the items more distinctive and thus boost item learning. Again, there was no effect of the order in which nouns and affixes were presented on item-learning. This is why in the third and final experiment of Study 3, a cross-situational learning paradigm was implemented, as similar work using this paradigm found a facilitatory effect of preceding utterances on the correct identification of the referent. This experiment, however, found no support for such an effect either. This is potentially important considering how widespread the effect of preceding parts of the utterance on processing is in the literature (see Section 1.4.3). However while all the experiments in Study 3 found relatively strong noun-picture learning, the learning of noun-affix correspondences was weaker. This is important given that the facilitatory effect of prefixing on noun-picture (item) learning is only predicted when there is noun-affix learning. It may therefore be that the key effect

is only found in a paradigm inducing better noun-affix learning and the chapter concludes with a discussion of alternative methods which could induce this.

Finally, the General Discussion in Chapter 7 considers the findings across the three studies together with respect to the two main hypotheses tested in this thesis: (1) suffixing advantage in generalization, due to greater cue competition in that condition; (2) prefixing advantage in item-learning, because the prefix reduces the entropy of the following noun, and makes it more learnable. This thesis found support for hypothesis (1) under certain conditions, but not for hypothesis (2). While the lack of evidence for (2) may, at least to some extent, be attributable to the limitations of the paradigm which may require further fine-tuning for testing this hypothesis, it may also be the case that hypothesis (2) is not true – that low entropy does not help learning. What this thesis does show is that the discriminative learning framework allows the researcher to generate principled, testable predictions; when no support for a prediction is found, this may be traced back to a specific component of the theory, which may be difficult under alternative approaches to language learning.

Chapter 2

Statistical Approach

2.1 Bayes factor: an overview

The primary method of statistical inference in this thesis is the Bayes Factor. Unlike p-values, which represent the likelihood of the data given the null hypothesis, the Bayes factor is the measure of the strength of evidence that the data provide for the theory (H1) over the null (H0) (or vice versa). The strength of evidence or the Bayes Factor represents the amount by which our prior confidence in H1 over H0 should change given the data. Notably, quantifying evidence of H0 is not possible using frequentist p-values. A non-significant p value ($p > 0.05$) does not provide evidence for the null (in reality, in some cases there may be more evidence for H1), despite this common misinterpretation in the literature (Dienes, 2008, 2014). The Bayes factor, on the other hand, differentiates evidence for no effect (H0) from ambiguous or inconclusive evidence. Apart from the theoretical and methodological appeal of being able to quantify evidence for the null, using Bayes factors may also be important given the growing methodological concerns in the field regarding publication bias, whereby work with ambiguous/null results is much less likely to be published (Ferguson & Heene, 2012; Masicampo & Lalande, 2012) or even submitted for publication in the first place (Franco, Malhotra, & Simonovits, 2014). A robust method which allows one to differentiate between an ambiguous result and a result that supports the null hypothesis, as well as to quantify the strength of support for the null, might make null results more valued in publication.

Throughout the thesis, Bayes factors are computed using the calculator presented in Dienes (2008) (implemented in R by Baguley & Kaye, 2010). This method requires a summary of the data and a model of the H1 – the theory. In the remainder of this section, I outline the logic for summarising the data and modelling the H1 at a conceptual level, and in Sections 2.1.1 and 2.1.2 I provide more technical detail of the approach.

Let us consider one of the predictions of this thesis: the suffix condition is better than the prefix condition in generalization. Specifically, we predict that in a test of generalization, participants in the suffix condition will on average have a higher score than participants in the suffix condition. To test this hypothesis statistically with Bayes factors, we need a model of the data, and a model of the H1. We model the data using the sample mean (e.g., mean difference between conditions, or the difference between a mean and a baseline)

and the standard error (SE) of that mean – in our example, therefore, we take the mean difference between the two affix conditions and the SE of the mean. Given that the Bayes factor calculator requires that the data are normally distributed, rather than using the average difference between the raw scores in the two conditions, we fit a logistic regression model (where the outcome variable is participants' accuracy at each generalization test trial) and use the coefficient for the main effect of affix as the mean, and the SE of that coefficient as the SE (see Section 2.1.1 for further detail). Using the log-transformed values therefore allows us to meet the normality assumption.

The next step is to model the H1 or the theory, that is, the predicted difference between conditions. This step involves specifying a probability distribution of the values of the parameter in question - in our example, this is a probability distribution of the difference between the suffix condition and the prefix condition. While there are different ways of modelling this distribution (see Dienes, 2014, for details), throughout the thesis we model the H1 as a half-normal distribution with a mean of 0 and an SD of x – a rough estimate of the mean difference. The key question is: what should this estimate be? (Note that it is possible to use a default value, but this approach is unprincipled.) One of the most straightforward cases is when comparable data are already available from previous work. For example, if we had previously ran the same or a similar experiment, or if a relevant experiment already exists in the literature, we could use the regression coefficient for a main effect of affix from that experiment as the SD of our model of H1. However, independent data are often not available, and in such cases the researcher needs to determine the value of x in some principled way. One possibility is to determine a plausible maximum effect and use that maximal value as an estimate of x. For example, if we predict better generalization in the suffix condition than in the prefix condition, a plausible maximum effect would be that there is no learning in the prefix condition (chance-level performance). In that case, all learning observed in the experiment across all participants would come from the suffix condition, and therefore the effect would correspond to the grand mean. Because we center the data before fitting the logistic regression, the grand mean corresponds to the intercept coefficient, and so we can use this value as the SD for our model of H1 (recall that this also allows us to meet the normality assumptions of the Bayes factor calculator). This approach of determining the plausible maximum is used throughout the thesis where independent data were not available (see Dong, Clayards, Brown, & Wonnacott, 2019; Samara, Singh, & Wonnacott, 2019; Wonnacott et al., 2017, for similar approaches). Note that the choice of the value from within the data set depends on the effect in question, that is, whether it is a main effect (as in the current example where we look at the difference between the two conditions), a two-way interaction, etc. For each of these possibilities, we developed a method for determining the plausible maximum, and Section 2.1.2 provides mathematical justifications for each method. In the results section of each experiment, for each Bayes Factor, I indicate which method was used to compute it.

We now have all the required values: a summary of the data and a model of the H1. The Bayes factor calculator then computes the likelihood of the data (as described by our summary values) given the model of the H1, and the likelihood of the data given the null

(where the null is zero difference between the conditions). The resulting Bayes factor is the ratio of these two likelihoods (H1 over H0), and it is interpreted 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). Note that in this thesis, in addition to the Bayes factor, we also report the p-value for each hypothesis following Dienes (2008) recommendation of reporting "a p for every B". Importantly, although p-values are reported for reader's interest, they are not interpreted – all statistical inferences in the thesis are based on Bayes factors only.

2.1.1 Data summary using mixed-effect models

In this thesis, the values required to summarize the data were obtained by fitting mixed effect regression models (Baayen, Davidson, & Bates, 2008) (using the lme4 R package Bates, Mächler, Bolker, & Walker, 2014; R et al., 2013). The regression coefficient for the effect of interest was used as the mean and the SE of the coefficient was used as the SE of the mean in the summary of the data. Logistic regression mixed-effect models were fitted when the dependent variable was a correct/incorrect response (which was the case in all but one analysis which involved reaction time data, and here linear regression was used). Using the values from the logistic mixed effects models, rather than the raw means and standard errors of participants' test accuracy allows us to work with log-transformed values and therefore meet normality assumptions.

In all models, all experimentally manipulated variables and all interactions between those variables were included as fixed factors. 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, Levy, Scheepers, and Tily (2013). Note that in theory, by-item slopes can also be included. This is in fact recommended in work with natural languages, where the researcher generalizes an effect of interest to all items in the language, that is, to the whole language (H. H. Clark, 1973). However, this is not relevant for artificial languages, which are used throughout this thesis (in addition, the assignment of words to pictures in artificial languages in this study is randomised on a participant-by-participant basis to limit item-specific effects, see Section 3.3.1.2 for details). All reported models converged with Bound Optimization by Quadratic Approximation (BOBYQA) optimization (Powell, 2009).

Importantly, throughout the thesis, I only report and interpret those fixed effects relevant to specific hypotheses, along with comparisons to chance for individual cells in the experimental design. The latter were used to shed light on whether there was evidence for "no learning" in some circumstances.

2.1.2 Modelling the H1

Recall that, following Dienes (2014), our approach is to model the H1 using a half-normal distribution with a mean of 0 and an SD of x, which is a rough estimate of the mean

difference. We developed five methods for determining x – each method is explained below, along with a mathematical justification.

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 there was evidence for an effect in one experiment, the beta coefficient and the SE from that experiment was used as the H1 for the same or similar effect in another experiment. Similarly, if there was evidence for an effect in one between-subjects condition (e.g., prefix), those values were used to model the H1 for the other condition (e.g., suffix).

Where using same or similar effects to estimate x was not possible, a plausible maximum value was determined from within the data or from independent data. In such cases, x is set to be half this plausible maximum value, since H1 is modelled as half normal and the maximum for the half normal is equal to 2SD.

B: Value for H1 for a main effect is based on the estimate for the grand mean (from independent or current data). For a main effect a with two levels, *a1* and *a2*, a plausible maximum difference between two conditions (main effect) corresponds to a situation in which one level performs at chance, and therefore the whole effect is carried by the other level. This corresponds to twice the grand-mean in log odds space. Since we use centred coding in the mixed-effects models (see below), twice the grand-mean corresponds to twice the intercept estimate. Therefore, x is equal to the coefficient for the grand mean.

A main effect a with two levels, *a1* and *a2*, is given as follows:

$$a = (a_1 - log(chance)) - (a_2 - log(chance)) \tag{2.1}$$

Applying our assumption for a plausible maximum difference to (2.1) therefore gives:

$$\begin{aligned} a &= (a_1 - log(0.5)) - (a_2 - log(0.5)) && (2.2)\\ &= (a_1 - 0) - (0 - 0) && (2.3)\\ &= a_1 && (2.4) \end{aligned}$$

The grand mean is given as follows:

$$\begin{aligned} \overline{e} &= \frac{(a_1 - log(0.5)) + (a_2 - log(0.5))}{2} && (2.5)\\ 2\overline{e} &= a_1 && (2.6)\\ &= a && (2.7) \end{aligned}$$

Therefore, the main effect of a is equal to twice the grand mean ($\overline{e}$). We therefore set x (the estimate for a) to half this value, since the maximum for the half normal is equal to 2SD. Since we use centred coding in the mixed-effects models, twice the grand-mean corresponds to twice the intercept estimate. Therefore, x is equal to the beta coefficient

for the intercept from the mixed-model.

C: Value for H1 for an interaction is based on a main effect estimate (from independent or current data). A two-way interaction $a.b$, where each of a and b have two levels, a_1, a_2 and b_1, b_2, respectively is given as follows:

$$a.b = (a_1.b_1 - a_1.b_2) - (a_2.b_1 - a_2.b_2) \tag{2.8}$$

A plausible maximum interaction corresponds to a situation in which there is no effect of b on one level of a, a_2, and the interaction is carried by the effect of b on the other level of a, a_1, given by:

$$a_1.b_1 - a_1.b_2 \neq 0 \tag{2.9}$$
$$a_2.b_1 - a_2.b_2 = 0 \tag{2.10}$$

Applying our assumptions for a plausible maximum therefore gives:

$$a.b = (a_1.b_1 - a_1.b_2) - 0 \tag{2.11}$$
$$= a_1.b_1 - a_1.b_2 \tag{2.12}$$

The main effect of a is therefore given by:

$$a = \frac{(a_1.b_1 - a_1.b_2) + (a_2.b_1 - a_2.b_2)}{2} \tag{2.13}$$
$$2a = (a_1.b_1 - a_1.b_2) + (a_2.b_1 - a_2.b_2) \tag{2.14}$$
$$= a.b + 0 \tag{2.15}$$
$$= a.b \tag{2.16}$$

The interaction $a.b$ is thus equal to twice the main effect of a ($2a$) (or b, depending on what is of most theoretical interest). 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.

D: Value for H1 for a three-way interaction is based on the estimate for a two-way interaction (from independent or current data). For the maximum corresponds to a result where the *a:b* interaction (or *a:c* or *b:c*, depending on what is of most theoretical interest) occurs at one level of c only and not at the other. This corresponds to twice the β coefficient (from the logistic mixed effects model) for *a:b*, and therefore x is set to half this value.

A three-way interaction $a.b.c$, where each of a, b and c have two levels, a_1, a_2, b_1, b_2 and c_1, c_2, respectively, is given as follows:

$$a.b.c = ((a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1)) - \tag{2.17}$$
$$((a_1.b_1.c_2 - a_1.b_2.c_2) - (a_2.b_1.c_2 - a_2.b_2.c_2)) \tag{2.18}$$

A plausible maximum $a.b.c$ interaction corresponds to a situation in which 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:

$$(a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1) \neq 0 \tag{2.19}$$

$$(a_1.b_1.c_2 - a_1.b_2.c_2) - (a_2.b_1.c_2 - a_2.b_2.c_2) = 0 \tag{2.20}$$

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

$$a.b.c = (a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1) \tag{2.21}$$

The interaction a.b is given as:

$$a.b = \frac{((a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1)) + ((a_1.b_1.c_2 - a_1.b_2.c_2) - (a_2.b_1.c_2 - a_2.b_2.c_2))}{2} \tag{2.22}$$

$$2a.b = (a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1) + 0 \tag{2.23}$$

$$= (a_1.b_1.c_1 - a_1.b_2.c_1) - (a_2.b_1.c_1 - a_2.b_2.c_1) \tag{2.24}$$

$$= a.b.c \tag{2.25}$$

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.

2.1.3 Robustness regions

Given the novelty of this approach, and the fact that the choice of value for H1 is subject to debate (Dienes, 2008), for every Bayes factor in this thesis, "robustness regions" were computed. Specifically, for each analysis in question, we calculated the Bayes factor using not just the value of H1 that was derived using our method, but for all other plausible values of H1. These values ranged from 0 in log odds to whichever value in log odds space corresponded to odds/odds ratio equivalent to one group being at chance and the other having almost perfect performance at 99% accuracy (bearing in mind that log odds corresponding to 100% accuracy cannot be computed). For example, when chance-level corresponded to 12.5%, the range of values was from 0 in log odds to 6.54. The values were examined in increments of 0.01, and the results were reported as ranges of values for which the data would support the same conclusion. For example, if we found evidence for the null in one of the analyses, we would report a robustness region [*x1, x2*] with *x1* being the smallest and *x2* the largest estimate of H1 (i.e. the value used as the SD of the theory) which would also lead to evidence for the null. 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. 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. These calculations were not used for formal inference, however, we did use them to support the interpretation of the data – for example, if there was support for a hypothesis with the majority of the values, this was taken as indirect support for the hypothesis.

2.2 Replication and optional stopping

Most of the experiments in this thesis were replicated. This was decided given concerns in the field about "replication crisis" (on-line methods were chosen in part to facilitate fast replication), whereby efforts to replicate previously published effects often fail to show the same result. For example, in a recent large-scale replication attempt (OpenScienceCollaboration et al., 2015), fewer than half of 100 effects from high-impact journals were replicated (see also Camerer et al., 2018; Nieuwland et al., 2018). Maxwell et al. (2015) suggested that multiple replications are required to have enough power to identify true effects, and the reason many replication attempts fail is because the original studies were often underpowered in the first place. This is further exacerbated by publication bias, whereby replications and negative results are difficult to publish (Fanelli, 2010; Makel, Plucker, & Hegarty, 2012; Sterling, 1959).

Therefore, the approach to replication was as follows: an initial study was carried out with a pre-specified sample size, and a replication was carried out regardless of the results of the original study (the only experiment that was not replicated was the final experiment of the thesis, Experiment 9, where a replication was not possible due to time constraints). This was done in order to eliminate researcher bias, which might lead to only replicating studies which show an effect without concern for the power to detect the effect (a "failed" experiment may in fact be Type II error, whereas a positive result may be Type I error). When the original study found evidence for an effect of interest, the effect from that study was used to model the H1 in the replication study. When the original study did not find the effect, it was not appropriate to use the estimate of that effect in the replication study, as in the absence of a true effect, the estimate is likely to be noise. In that case, an alternative methods for modelling the H1 was used, depending on the type of effect (see Appendix 1). Note that, unlike p-values, Bayes factors are interpretable when combining data sets that have been inspected independently (Dienes, 2016).

Note that the time and resource constraints associated with a doctoral programme meant that obtaining sufficient power to detect the effect had to be balanced against testing as few participants as possible in the given amount of time. This is why the optional stopping procedure was generally used (in all studies except in Chapter 3) – we would start out by testing 20 participants per condition, and continue testing until there was evidence for either the null or for the H1 (regardless of the prediction). While optional stopping would not be appropriate with a frequentist approach ("data peeking", Francis, 2012), unlike p-values, Bayes factors allow optional stopping (Dienes, 2016; J. N. Rouder,

2014). The robustness of Bayes Factors to optional stopping was disputed by de Heide and Grünwald (2017), however, J. Rouder (2019) demonstrated that optional stopping is an issue with Bayes Factors only when unjustified default priors are used, which is not the case in this thesis. Another advantage of the robustness of Bayes factors to optional stopping was that it was possible to combine the original study data and the replication data. Therefore, after analysing each data set separately, and reporting those results, we also combined the two data sets and analysed them in a separate series of analyses. The benefit of this was that the combined sample was larger than the individual samples, making the data more robust (Dienes, 2016), therefore the key conclusions were drawn based on combined data, where applicable (in some cases this was not possible, for example, when the replication study involved additional tests of learning not used in the original study).

Two later studies in this thesis – Studies 2 and 3 – used pre-registration, which has been proposed as a solution to the replication crisis (Nosek & Lakens, 2014). The idea is that researchers report their study design and hypotheses prior to data collection, and thus prevent post-hoc changes to the original predictions as a result of seeing the data; any additional analyses or hypotheses tested which were not part of the pre-registered report are therefore unambiguously exploratory. There are several ways in which researchers can pre-register. In recent years, many journal articles have introduced the Registered Report, a format in which the authors write-up the introduction and the method of their paper, including hypotheses and planned statistical analyses, and submit to the journal before data are collected. The Registered Report is then peer-reviewed and, if accepted for publication, publication is guaranteed regardless of whether they study finds a positive or a negative result, thus removing publication bias (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Nosek & Lakens, 2014). Another possibility, used in this thesis, is to pre-register the key hypotheses and planned analyses by publishing them on-line, on a free public repository such as the OSF (which offers multiple formats for pre-registration, from a detailed questionnaire to an open-ended verbal description; https://osf.io) or RPubs (https://rpubs.com). This option is not related to publication and is not tied to any journal, but it provides a public, permanent record of pre-registered hypotheses and methods that authors can refer to in journal manuscripts. For Study 2, I wrote "analysis plans" – R markdown documents which combined a verbal description of the study method and all hypotheses to be tested, with R code to be used for analysis, including specifying in advance which method will be used to compute Bayes factors. These analysis plans were posted on RPbus, where they are time-stamped and publicly available. For Study 3, I wrote verbal descriptions of study design and predictions, and reported values to be used to model the H1 or, when no previous values were available, I specified the methods which would be used to obtain the values from the data. These were published on OSF as Open-Ended Registrations. For each chapter, I provide links to the pre-registrations, as well as the collected data and R analyses scripts.

Chapter 3

Study 1: Experiments 1 – 2

3.1 Introduction

One of the key insights from learning theory is that cue competition is the driving mechanism of learning. As discussed in Section 1.4, Ramscar et al. (2010) demonstrated that the order in which learners are presented with linguistic labels and referents affects cue competition, and as a result, learning. In their study, participants only showed appropriate generalization when they were first presented with visual referents and then with their corresponding labels; when labels were presented first, participants showed poorer generalization. The authors interpret this result to suggest that visual referents provide greater cue competition than verbal labels, and that seeing referents first should lead to better discrimination of informative cues, and generalization on the basis of those cues.

I explore the effects of order on learning using artificial languages which are "prefixing" or "suffixing". Specifically, these languages consist of "nouns", which uniquely map onto visual objects, and these nouns are either preceded by an "affix" (prefix condition) or followed by an "affix" (suffix condition). Before going into more detail about the predictions of the discriminative learning framework with respect to order effects, it is necessary to clarify what exactly is meant by suffixing and prefixing in the context of this thesis, and how it relates to the psycholinguistic study of morphology. First of all, the suffixing and prefixing artificial languages in this thesis are simplified representations of affixing, and are not designed to reflect natural language morphology – the only difference between the conditions is the order in which an artificial language noun and affix are presented to the learners. While suffixing and prefixing are used in this thesis as a test-case for studying the effects of order on learning, the differences in learning and processing suffixing and prefixing in natural language has been of interest to psycholinguists. For example, E. Clark (2007) reported that children learning predominantly suffixing languages learn the morphology of their language faster than children learning predominantly prefixing languages. In terms of processing, Hawkins and Gilligan (1988) suggested that prefixed words take longer to process, as the information necessary to uniquely identify the lexical comes after the prefix during processing (whereas in suffixing, the lexical item is identifiable sooner). Similarly, other approaches have argued that prefixed words take longer to process because they are indexed in the mental lexicon by their stems, rather than by the prefix

(Taft, 1988). This causes processing delays with pseudo-prefixed words such as *reindeer* (where *re-* is misinterpreted as a prefix), where the listener searches the mental lexicon for the pseudo-stem *indeer*; as *-indeer* is not available as an entry point in lexical access, a new search has to be initiated for the complete form *reindeer*, which results in longer processing times (note that Milin et al. (2017) recently demonstrated that such delays are predicted parsimoniously by a discriminative learning model). This thesis, however, does not directly speak to these theories, and any differences in learning between the suffix and prefix condition will be interpreted in terms of the effects of order on discriminative learning – any such findings may be of interest to researchers studying learning and processing of natural language morphology, but would have to be replicated with artificial languages designed to more closely match particular aspects of natural language morphology. Finally, the potential suffixing advantage in processing has been proposed as an explanation for a cross-linguistic suffixing bias, that is, the fact that, across languages of the world, strongly suffixing languages are much more common than strongly prefixing languages (Dryer & Haspelmath, 2013; Greenberg, 1963) (see Hawkins & Gilligan, 1988; Hupp, Sloutsky, & Culicover, 2009; St Clair, Monaghan, & Ramscar, 2009, for such proposals). However, linking structural tendencies across languages to aspects of human cognition is controversial (Evans & Levinson, 2009); Dunn, Greenhill, Levinson, and Gray (2011) demonstrated that what on surface may appear as a cross-linguistic bias is much weaker once language relatedness is controlled for. Therefore, this thesis does not speak to the different cross-linguistic distributions of suffixing and prefixing.

Returning to the discriminative learning framework laid out by Ramscar et al. (2010), we can think of affixes as labels and the stems that they attach to as features. To illustrate with an example from our study, in the suffix condition, the participant hears a noun *foop* and its corresponding picture, followed by a suffix ma. Therefore features in this context are the phonological features of foop and the visual features of the associated picture; the label is *ma*. The discriminating features are all the features that predict the occurrence of the affix ma most reliably, across exemplars such as *foop, moog, joob*, and the visual (semantic) features of their corresponding pictures. In the prefix condition, on the other hand, the participant hears the affix ma, followed by *foop* and a picture, where the noun and the picture are features, and the affix is the label. It follows from the predictions of the discriminative learning framework that suffixing is more conducive of discriminative learning: only when lexical items precede affixes (which is the case with suffixing) can cue competition over the features of those lexical items occur. St Clair et al. (2009) demonstrated this experimentally using artificial language learning. They taught adults 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 particular suffixes (suffixing condition) or prefixes (prefixing condition). During training, participants heard noun + affix pairs, and at test were presented with novel pairs which were either compatible (categoryA noun + categoryA affix) or incompatible (categoryA noun + categoryB affix) to training pairs. At test, participants were instructed that half of the pairs would be

"similar" to the training language and half "dissimilar", and to press "Y" for similar and "N" for dissimilar. While both conditions performed above chance, participants in the suffix condition were significantly more accurate than those in the prefix condition. A similar finding is reported by Hupp et al. (2009), who found that adult learners were more likely to extend a modified version of a target label if the label was suffixed as opposed to prefixed. For example, participants heard *this is a tate*, and saw a heart-shaped object; after this, they heard the same label but with an added affix, and had to choose from a heart-shaped and a star-shaped object. If the label was suffixed (i.e., which one is a *tate-be*?), participants were more likely to choose the heart-shaped object than they were if the label was prefixed (i.e., which one is a *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 than the prefixed label.

The work of St Clair et al. (2009) and Hupp et al. (2009) suggest that suffixes promoted the learning of the informative features – that is, the common dimensions which grouped nouns with the same affixes. Ramscar (2013) further points out that that, rather than suffix learning being generally "better" than prefix learning, the two different orderings promote different types of learning. Specifically, while suffixing promotes the abstraction of dimensions across features which group multiple items with the same suffix, prefixing should promote learning of the trained items. The latter is predicted under the account laid out by Dye and colleagues (2017, 2018) and discussed in Section 1.4.2 – that is, encountering a prefix (similarly to a gender marked article or pre-nominal adjective) reduces the entropy of the upcoming noun, aiding its processing. The processing benefit associated with reduced entropy may also help learning. 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 (i) participants were first exposed to the nouns without their articles, and then to nouns with their articles versus (ii) participants where 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 nouns in isolations (with implications for foreign language teaching). However, they also report higher learning of noun-object mappings in that condition, which they argue is because in the early exposure to the article-noun mappings, the article made the nouns easier to process, which in turn made it easier to learn the "meaning" of the noun (its corresponding picture). Therefore, while suffixing promotes generalization (due to greater cue competition and prediction error in this condition compared to the prefix condition), learning prefixing helps item-learning (as the prefix reduces the entropy of the noun, while the suffix does not). Ramscar (2013) tested this in an artificial learning experiment, where adult learners were trained on prefix + noun + suffix strings denoting pictures of everyday objects. 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. Items occurring with the same suffix, however, were judged more similar to each other than items occurring with the same prefix, because learning an item-suffix

mapping involved learning shared dimensions which grouped items together with the same suffix.

The studies reviewed above suggest that prefixing and suffixing facilitate qualitatively different learning, with prefixing leading to better learning at the individual item-level (Ramscar, 2013), and suffixing facilitating the learning of discriminating features – semantic and/or phonological features which group items together (Hupp et al., 2009; St Clair et al., 2009; Ramscar, 2013). However, 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. In the work presented in this chapter, both learning of trained items, and generalization to novel items will be tested, providing a more realistic learning context. Not only does this design provide a more controlled test of the core principles of discriminative learning with respect to linguistic generalization – which is the main aim of this thesis – we may also gain more insight into the relationship between item-based learning and generalization. Note this would speak to the usage-based theory language learning (Tomasello, 2000), which argues that generalization emerges gradually and that early language use is item-based (Section 1.2.1).

This chapter includes a computational modelling experiment implementing the Rescorla-Wagner mathematical model of learning (Rescorla & Wagner, 1972) (Experiment 1) and two artificial language learning experiments (Experiments 2a and 2b) with adult learners. The computational models were exposed to abstract representations of the input that was used with human learners, and the performance of the models was used as a basis for formulating the relevant predictions for the experiments with humans. Both the simulations and human participants were "trained on" artificial languages (in the case of the simulations, on abstract representations of the languages) which consisted of visual referents and corresponding noun+affix pairs. Each noun was associated with a unique visual referent – a novel "alien" character – and each noun+affix pair was associated with one of the two affixes. Affix usage was predicted by both semantic and phonological cues: all the aliens which had the same body shape co-occurred with the nouns which had the same vowel, and with the same affix. Therefore, the body shapes and the vowels were the discriminating features. There were also other features of the items, such as the number of eyes and feet of the aliens, as well as consonants in the nouns, which varied across affixes and were not predictive of affix occurrence – these were the non-discriminating features.

The Rescorla-Wagner model (Rescorla & Wagner, 1972) learns associations between cues and outcomes. During training, on a given trial in the prefix simulation the model was first exposed to the prefix (the cue), followed by features representing the noun and the visual referent (outcomes); the order was reversed in the suffix simulation, where the model was first exposed to the noun and the visual referent (cues), followed by the suffix (the outcome). In the experiments with humans, participants in the prefix condition were exposed to the prefix, after which the visual referent appeared on-screen and the noun

was played; in the suffix condition, the timing was reversed: the visual referent appeared on-screen first, as the noun was played, after which the affix was played (see Section 3.3.13 for details). Following exposure, participants were tested on their recall of trained items and on generalization, that is, on their ability to choose the correct affix for a novel item which contains the discriminating features predictive of that affix (see Section 3.3.1.3). In the simulations, on the other hand, learning was evaluated by activating the features corresponding to a novel "alien" item in the network, and determining the probability of the network selecting the correct affix (or the incorrect affix) (as outlined in Section 3.2.3). With respect to item-learning, we predicted better learning in the prefix condition. This prediction is based on the idea that the prefix reduces the entropy of the following noun, and that this makes it easier to learn the "meaning" of the noun (its associated picture). Therefore in Section 3.2.6., we present a mathematical analysis of entropy which demonstrates that the probability of choosing the correct picture for a given noun in our experiments is greater in the prefix condition compared to the suffix condition.

Finally, note that the behavioural experiments were designed with the view of developing a paradigm that could be used with child learners in future work. To this end, discriminating features in this design were deterministic – this was considered appropriate for future use with child learners, as previous work (Brown et al., 2018; Schwab et al., 2018) found that children were unable to generalize probabilistic semantic cues. Unlike previous relevant work (Ramscar et al., 2010; Ramscar, 2013; St Clair et al., 2009), in this study, semantic and phonological cues were equally predictive of the affix. This was done in order to provide participants with as many possible helpful cues for generalization, which may be particularly important for child learners. Also, semantic and phonological cues to grammatical function tend to correlate in natural languages (e.g. Mirković et al., 2005), which makes this design more naturalistic. In addition to this, the exposure phase was designed to be more child-friendly, as participants moved between trials at their own pace (timed exposure might be overwhelming for child learners or make it hard for them to engage with the stimuli).

The data and the R analyses code for Experiments 2a and 2b is here: https://osf.io/tqp5a/ and the simulation code for Experiment 1 is here:
https://github.com/masavujovic/RescorlaWagner

3.2 Experiment 1 (Computational Model)

In this experiment, we implement the Rescorla-Wagner model of learning (Rescorla & Wagner, 1972), which models the basic principles of learning theory, introduced in Section 1.4. The output of the model 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, under different conditions. In this experiment, two conditions are compared: suffixing and prefixing, with the goal of testing one of the predictions of the discriminative learning theory (Ramscar et al., 2010), specifically, that predicting affixes from semantic and phonological features (suffixing) results in better discrimination

between informative and uninformative cues, compared to predicting features from affixes (prefixing). The section below introduces the model formally.

3.2.1 Rescorla-Wagner learning rule

We define a single discrete learning trial as a vector of cues:

$\phi(S) = (\phi_1(S), \phi_2(S), \ldots, \phi_n(S))$, where $\phi_i(S) = 1$ if the cue is present in the trial, and 0 otherwise. The predicted response is generated by aggregating the associative strengths of all the cues present in the trial. Therefore, if an n-dimensional vector of associative strengths (weights) is θ, the predicted response is given by:

$$V(S, \theta) = \theta\phi(S)^{\mathbf{T}} \tag{3.1}$$

The predicted response is then compared with the outcome y. If the outcome occurs, the value of $\mathbf{y}$ is 1, otherwise it is 0. The prediction error (the distance between the predicted response and the outcome) is computed as follows:

$$\frac{1}{2}(y - V(S, \theta))^2 = L(S, \theta) \tag{3.2}$$

The next step is to compute the amount by which to update the vector of weights θ^k (where k is the current trial) with the ultimate goal of finding the weights that make the prediction error as close to zero as possible. For this we used the Rescorla-Wagner learning rule. Note that this rule is essentially the same as the Least Mean Square or the Widrow-Hoff learning rule (Widrow & Hoff, 1960), which was developed in mathematics for solving sets of linear equations. The same learning rule is known in the machine learning community as the Delta rule, a special case of the backpropagation algorithm which uses gradient descent to minimize prediction error. The updated vector of weights θ^{k+1} is therefore given as follows:

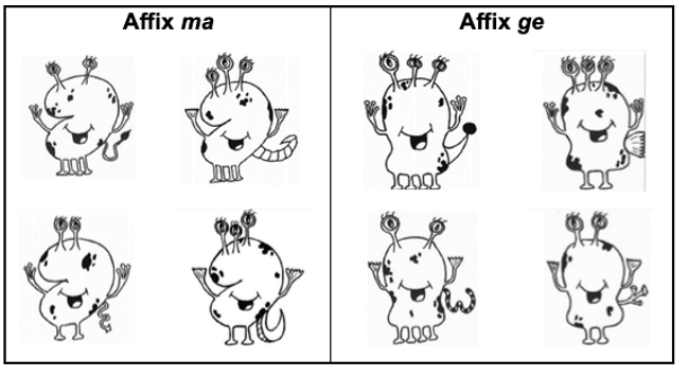
$$\theta^{k+1} = \theta^k - \alpha_i\beta_j\nabla_\theta L(S_i, \theta^k) \tag{3.3}$$

$$= \theta^k + \alpha_i\beta_j(y_j - V(S_i, \theta^k))\phi(S_i) \tag{3.4}$$

where α is a hyperparameter corresponding to saliency of individual cues, and β is the learning rate. In this experiment, default values of 0.01 were used for both parameters following Ramscar et al. (2010).

3.2.2 Simulations

The two simulations were designed, one for the suffix condition and one for the prefix condition – the design was comparable to Ramscar et al. (2010) who modelled word learning in feature-label and label-feature conditions, analogous to the suffix and prefix conditions in this experiment, respectively.



Affix	Discriminating		Non-discriminating					
	shape1	shape2	eyes1	legs1	hands1	eyes2	legs2	hands2
ma	1	0	1	1	1	0	0	0
	1	0	0	1	0	1	0	1
	1	0	1	0	1	0	1	0
	1	0	0	0	0	1	1	1
ge	0	1	0	1	1	1	0	0
	0	1	0	0	1	1	1	0
	0	1	1	0	0	0	1	1
	0	1	1	1	0	0	0	1

Figure 3.1: The images used as the training set for modelling (top panel), and the same training set represented as a matrix on which the models were trained (bottom panel).

In both simulations, the models were trained on a representation of the input that was used with human learners in Experiments 2a and 2b. Semantic cues were modelled in terms of individual discrete verbalizable features (e.g., "body shape 1" – *shape1*, "body shape 2" – *shape2*, "has two legs" – *legs1*, "has four legs" – *legs2*, etc.). For simplicity, phonological cues were modelled in terms of individual discrete phonemes. However, we note that this is likely not a realistic model of the way humans perceive continuous speech, and that it does not capture potential differences between the learning of phonological and semantic cues we may observe with human learning (I return to this point in Chapter 5, in which the modelling is revisited with insights from the behavioural experiments across Studies 1 and 2). A representation of the input set is given in Figure 3.1, which shows the visual stimuli and the same training set represented as an 8x8 matrix on which the models were trained. Note that the figure is simplified for illustration purposes, as phonological cues are left out. The simulations included two vowel features (vowel1 (*ee*) – affix1; vowel2 (*oo*) – affix2) and an additional eight consonants (k, f, m, j, p, b, g) which were combined to produce a unique label for each exemplar: e.g., *keeb, jeed, foog, moop*. These were left out from the diagram as an 8x18 matrix would be difficult to present clearly. The networks were trained on 2000 trials each.

3.2.3 Evaluating the model

Recall that our goal is to assess whether the model can discriminate between informative and uninformative cues under the two different conditions – in the prefix simulation and in the suffix simulation. While this can be done in different ways, we followed the procedure outlined in Ramscar et al. (2010). Specifically, after the final learning trial, learning was assessed by activating a set of features corresponding to a randomly chosen exemplar, and summing the associative strengths (connection weights) of those features for each of the affixes. Note that summing the weights is directly comparable to how the model generates a response during learning (as outlined in Section 3.2.1). Following Ramscar et al. (2010), the sums of raw weights w for the correct and the incorrect affix were normalized into a probability distribution using the Luce's 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:

$$P(w_i) = \frac{w_i}{\sum_j w_j} \tag{3.5}$$

Finally, in addition to evaluating the models on the final set of weights, we will plot the weights between cues and outcomes over time, in order to gain an insight into the dynamics of learning as it unfolded.

3.2.4 Results

Figure 3.2 shows the associative strength of each feature for a randomly chosen affix "ma" in both conditions in a single model run, and Figure 3.3 shows the results of the evaluation method, in which we compare how strongly a randomly chosen test exemplar is associated with the correct affix in each condition. As can be seen in the left panel in Figure 3.3, raw weights suggest that the test item is associated more strongly with the correct than the incorrect affix in both conditions. However, when the raw sums of weights were normalized into a probability distribution, as described in the previous section, the probability of choosing the correct affix was greater in the suffix condition than in the prefix condition. This is shown in Figure 3.3, where the right panel plots the probability of the network choosing the correct affix (for simplicity the probability of the incorrect affix was not plotted, but this corresponds to 1 – p(correct)).

3.2.5 Discussion

Beginning with a discussion of raw weights between individual features and labels over time, our results show that, in both conditions, the highest associative strength was assigned to the relevant informative/discriminating features. However, the total amount of associative strength was distributed differently across the cues in the two conditions. In the suffix condition, the same amount of associative strength was assigned to the discriminating features, but the strength was positive for the discriminating feature for the given affix, and

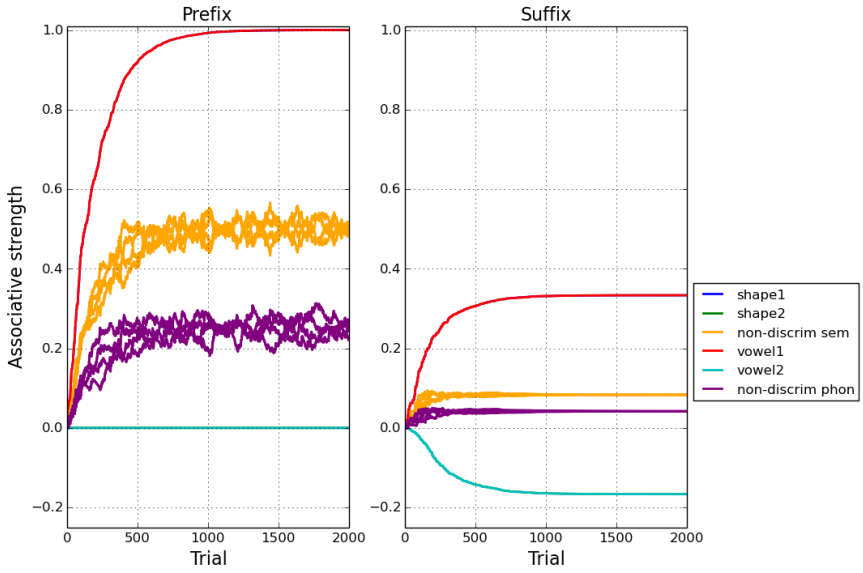
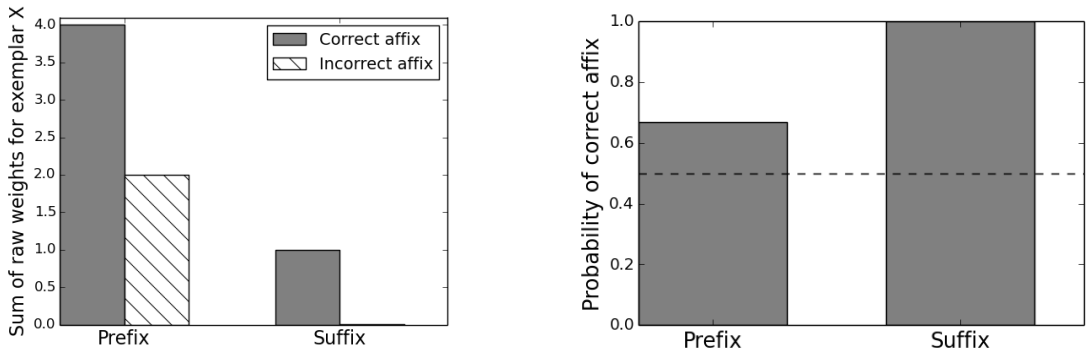


Figure 3.2: Associative strength of each feature for a randomly chosen affix "ma" over time in a single model run, in the two conditions. The red line represents the discriminating phonological feature for affix "ma", *vowel1*. This feature was learned exactly the same as the discriminating semantic feature *shape1* (dark blue line), and thus the line has been over-plotted. The bright blue line is the feature *vowel1*, the discriminating phonological feature for the opposite affix "ge" (which has been learned identically as the discriminating semantic feature *shape2*, corresponding to the over-plotted dark green line). The orange lines are the non-discriminating "semantic" features (*eyes1, eyes2, legs1, legs2, hands1, hands2*), and the purple lines are the non-discriminating "phonological" features (*k, m, j, f, p, b, g, d*.



(a) (b)

Figure 3.3: (a) The sum of raw associative strengths of featured corresponding to a randomly chosen test item, for the correct affix (grey bar) and the incorrect affix (white striped bar) in each affix condition. (b) The sums for correct affixes normalized into probability of choosing the correct affix out of two options (dashed-line is chance, 0.5).

negative for the discriminating feature for the opposite affix. This is a clear demonstration of the background rate and cue competition principles of learning theory – on the trials in which the affix *ma* (for example) occurred in the absence of shape2 and vowel2, these cues lost predictive value to the other cues present in the trial. Because the affix *ma* always occurred in the absence of shape2 and vowel2, this resulted in negative weights for this feature. Critically, by comparison, this did not happen in the prefix condition, where there was no learning about the feature which never occur with the target affix (such as shape2 and vowel2, where the weights were constantly at zero).

Turning to the features which did occur with both affixes, and were thus not predictive – that is, the non-discriminating features – the weights for these features were more stable over time in the suffix condition, where the model converged on relatively low weights after fewer than 500 trials. In the prefix condition, on the other hand, the weights of these features did not reach asymptote. This is indicative of a difficulty in the prefix simulation with "unlearning" these uninformative, non-discriminating features across training trials. The consequences of these differences between the conditions became apparent when we tested the model, which we turn to next.

In order to assess the learning of the model, we activated a set of features corresponding to a randomly selected testing item through the final set of weights, and observed whether the testing item would be associated more strongly with the correct or the incorrect affix. Looking at the raw sums of weights, in both conditions the test item was associated more strongly with the correct affix than the incorrect, however, the associative strength for the incorrect affix was higher in the prefix condition than in the suffix condition. This is because in the prefix condition, non-discriminating features and opposite-category discriminating features were not "unlearned"; in the suffix condition, on the other hand, these had low or negative weights. To get a measure of model's accuracy in selecting the correct affix, the raw weights were normalized into a probability distribution. This revealed that, while both conditions were above chance in selecting the correct affix (as suggested by the raw weights), the probability of the correct affix was greater in the suffix condition than in the prefix condition.

It is worth pointing out that the metric for assessing the simulations is not built into the Rescorla-Wagner model, rather, is an experimenter choice. Normalizing raw weights into probability distributions via the Luce's choicenaxiom was chosen as it was used by Ramscar et al. (2010), whose work we aimed to replicate. However, other metrics are used with neural networks, such as the sigmoid or the softmax function. The discriminative learning theory does not explicitly model how humans may be using associative strengths to generate responses, and therefore it is not clear which method is most appropriate. This important point is re-visited in Chapter 5, where I re-assess the models following insights from human experiments in Chapters 3 and 4. For now, however, I follow the previous relevant work and make the predictions based on the results from the test metric presented above.

To summarize, the simulations presented in this section showed via basic principles of discriminative learning that suffixing leads to better discrimination of informative versus

uninformative cues. As a consequence of this, when tested, the suffixing network was more likely to select the correct affix for a test item than was the prefixing network. This is consistent with Ramscar et al. (2010). Note that their simulations included a type-frequency manipulation, which was not the case in the present work. Our results are comparable to their high type-frequency results, where they found above-chance performance in both simulations, but higher performance in the FL simulation (analogous to the suffix condition in this work) than in the LF simulation (corresponding to the prefix condition).

3.2.6 Predictions for human learning based on the model and entropy calculations

Based on the performance of the two simulations (as well as previous work), we predict that, in the behavioural experiment, learners in the prefix condition will show poorer learning of discriminating features than learners in the suffix condition, and this will be evidenced by significantly poorer performance on tests of generalization. Regarding item-learning, the simulations in this experiment do not capture the effect of entropy-smoothing on real-time processing and, consequently, learning, and therefore our prediction is not based on the Rescorla-Wagner model. We predict better item-learning in the prefix condition on the basis of previous work, as well as our calculations of entropy given below:

We know that for a joint probability distribution of nouns n and affixes a, P(n, a), the marginal probability of a noun n is given as follows:

$$P(n) = \sum_{n} P(n, a) \tag{3.6}$$

The conditional probability of a noun n given a is:

$$P(n|a) = \frac{P(n, a)}{P(a)} \tag{3.7}$$

$$= \frac{P(n, a)}{\sum_n P(n, a)} \tag{3.8}$$

The uncertainty of a probability distribution is measured as entropy. Let p(X) be any finite probability mass function, such that $\sum i = 1^N p(X_i) = 1$ and $p(X_i) and$p(Xi)$\geq$ 0 for all i. Then the entropy (expressed in bits) p(X) is given as:

$$H(p) = -\sum_{i=1}^{N} p(X_i log p(H_i)) \tag{3.9}$$

Therefore, the entropy of the probability distributions above (the marginal probability of nouns and the conditional probability of the noun given the affix), can be computed using the entropy equation.

Below we demonstrate this with a specific example. In the first experiment in this

chapter, Experiment 2a, participants are exposed to 8 nouns and two affixes, where half of the nouns occur with one affix, and the other half with the other. Formally, the joint probability distribution of nouns n and affixes a for this experiment is therefore given as:

$$P(n = i, a = j) = \begin{cases} \frac{1}{8} & \text{if } i \leq 4, j = 1 \\ 0 & \text{if } j \leq 4, j = 2 \\ 0 & \text{if } i > 4, j = 1 \\ \frac{1}{8} & \text{if } i > 4, j = 2 \end{cases}$$

Across the two conditions – the suffix and the prefix condition – we manipulate the order in which the nouns and the affixes are presented to participants, and we are interested in probability of a noun n_i just before it is heard.

In the suffix condition, the nouns are played before the affixes. Therefore the probability of n_i just before it is heard corresponds to the marginal probability of n defined in (3.7). In our experiment, this equals to:

$$P(n_i) = a \sum_{a} P(n_i, a) = \frac{1}{8} \tag{3.10}$$

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

$$P(n_i|a_j) = \begin{cases} \frac{1}{4} & \text{if } j = 1 \\ 0 & \text{if } j = 2 \end{cases}$$

Therefore, in idealized conditions, where we assume that participants have learned the probability distribution perfectly, and that they generate their expectations about the identity of the upcoming noun according to the distribution, we can see that in the suffix condition, the greatest chance of successfully predicting the noun is equal to chance, or 1 in 8, whereas in the prefix condition, this is twice as high – it equals to 1 in 4.

Using the entropy equation defined above, we can calculate the entropy of the marginal probabilities of n, as well as the entropy of the conditional probability of n 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 n is 4 bits, but the entropy of n given a is 3 bits. This shows that in the prefix condition, at the point at which the noun is heard, the entropy is lower than in the suffix condition.

To summarize, from the performance of the computational models presented in Experiment 1, we predict better generalization in the suffix condition compared to the prefix condition. From the entropy calculations presented above, we demonstrated that the entropy of the noun is lower in the prefix condition compared to the suffix condition. Based on this, we predict better item-learning (better learning of which noun occurred with which picture) in the prefix condition than in the suffix condition.

3.3 Experiment 2a

3.3.1 Rationale and hypotheses

In Experiment 2a, in the prefix condition, the affix was played first, followed by the noun, which was played at the same time as the picture was displayed. The picture remained on-screen until the participant moved on to the next trial. In the suffix condition, on the other hand, the noun and the picture were presented first, and this was followed by the affix. As our aim was to ensure that the picture is visible for the same amount of time in both conditions, this did mean that in the suffix condition the picture remained on-screen while the affix was played. In retrospect, it became clear that the key aspect of stimulus presentation are that the nouns and the pictures (features) are separate from the affixes (labels), as this represents the most direct, controlled test of the effects of the order of features and labels on learning. Therefore, in Experiment 2b, the suffix condition was timed such that the referent disappeared once the noun finished and before the suffix was played.

Following exposure, participants' generalization and item learning was tested via a battery of tests. The purpose of the generalization tests is to use novel nouns to assess whether participants have learned the relationships between noun properties (semantic and phonological) and the affixes. Ideally, we would test semantic and phonological cues in isolation (i.e. to assess whether they are learned independently), however, although it is possible to test a novel noun without presenting its semantics (i.e. if it is heard with no visuals), it is not possible (or at least is very unnatural) to do the reverse and test only semantics without phonology. Therefore, our first set of generalization tests novel nouns with both semantic and phonological properties. A further concern is that, in addition to learning (i) the relationships between the semantic properties of the nouns and the affixes and (ii) between the phonological properties of the nouns and the affixes (our key questions of interest) participants might (iii) directly associate the phonological and semantic features of the noun with each other. Whilst it is difficult to test these independently, we designed three tests, such that each tested two of (i), (ii) and (iii) (so that if participants only learned one of these types of relationships, they would be at chance at one of the tests). Specifically, generalization was tested by asking participants to either match a novel noun+affix label to one of two novel fribbles (Test 1), or to match a novel fribble to one of two novel noun+affix labels, manipulating whether the foil label differed with respect to noun phonology (Test 2) or affix (Test 3). Here, we predict better performance in the suffix condition compared to the prefix condition. In Experiment 2b, in addition to the three generalization tests described above, we included a final test of generalization without visual stimuli which looked at the learning of phonological cues in isolation (since, as noted above this is possible since it is possible to present novel nouns and affixes without semantics). Here we also predicted that the suffix condition would outperform the prefix condition.

Item learning was tested by presenting participants with a trained label and the pictures of all trained items arranged in a grid. In order to do well in this test, participants needed to have learned idiosyncratic associations between individual items and noun-labels.

Following Ramscar (2013), we predicted that the prefix condition will outperform the suffix condition, as evidenced by a main effect of affix in that direction (also consistent with Arnon & Ramscar, 2012; Dye et al., 2017, 2018; Lew-Williams & Fernald, 2007, 2010). In Experiment 2b, we also included an additional test with trained items, which was equivalent in design to the generalization test. Our predictions here are less clear because participants could perform above-chance in this test either by making appropriate abstractions without item knowledge (in which case the test is equivalent to the generalization test), or with appropriate item-knowledge and no abstraction. However, we tested the hypothesis that participants would act as in the generalization test and thus be better in the suffix condition and with an interaction with type-frequency. We do bear in mind that any such effects may be attenuated by better learning of items under prefixing.

To summarize, across the two experiments with humans, the two key hypotheses are: (1) that participants in the prefix condition would show better item-learning than in the suffix condition – that is, they would be more in recalling the "names" (labels) of individual aliens (item-learning); and (2) that participants in the suffix condition would be better than participants in the prefix condition at generalization – that is, better at learning that it is the body shape and the vowel (and not other visual/auditory features) that consistently predict the affix.

3.3.2 Method

3.3.2.1 Participants

Thirty-two participants (16 per condition) were recruited through the UCL Subjects Pool (SONA). All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for participation.

3.3.2.2 Stimuli

Audio stimuli consisted of 16 nouns (8 training, 8 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 and two Category B affixes (following St Clair, 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, Pagel, Pierret, Bataille, & Van der Vrecken, 1996) with a male British English voice.

Visual stimuli were hand-drawn pictures of two categories of novel "aliens". There were

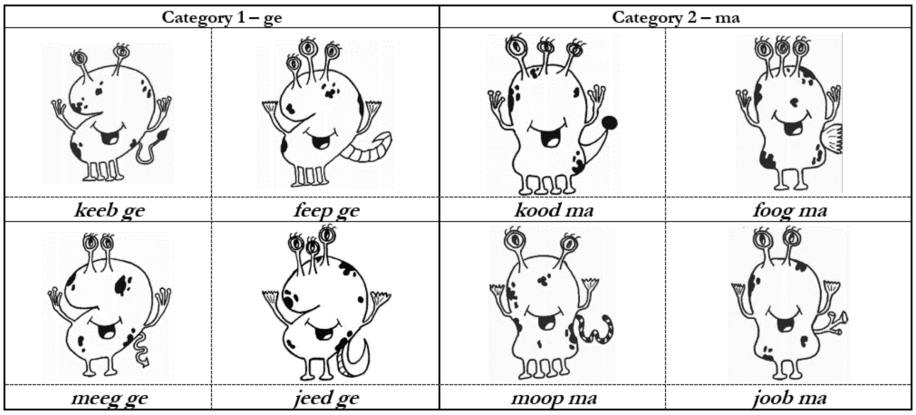


Figure 3.4: Sample training set. Note that nouns were assigned to individual aliens randomly on a participant-by-participant basis

16 items in total (8 training, 8 testing - with assignment to training/testing randomized across participant), eight per category. The two categories of aliens were distinguished by a salient visual feature, body shape. In addition, there were four non-discriminating visual features: number of eyes (three or two), number of feet (four or two), shape of hands (duck-like or frog-like), and tail (each alien had a unique tail). There were 16 aliens, 8 in each category (4 training, 4 were reserved for testing. The nouns were assigned to individual aliens randomly on a participant-by-participant basis. See Figure 3.4 for a sample training set.

3.3.2.3 Procedure

Participants were told that they were about to "learn an alien language" and that they would see pictures of aliens and hear "how they are referred to in the alien language". In order to discourage explicit learning, they were told not to focus on learning the rules of the language, but to instead focus on the pictures and the sounds.

Training The training session consisted of 4 blocks of 24 trials each (256 trials in total, 12 exposures per item). Participants were instructed to repeat outloud the sound they hear, and to press the space bar to move on to the next trial. 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 alien. The picture and the noun were presented at the same time. After the participant pressed the space bar, a blank screen was displayed for 1000ms and a new trial began. In the suffix condition, the picture was presented for 450ms, after which the noun and the affix were played. The picture remained on-screen until the participant pressed the space bar (see Figure 3.5).

Testing

Item learning test We tested participants' knowledge of the associations between individual nouns and individual visual items using an eight-alternative forced-choice test. All of the eight training items were displayed on-sceen, and a noun+affix bigram was played, which corresponded to one of training items. Participants had to click on the item matching the bigram. Each trained item served as a target once, giving a total of eight trials.

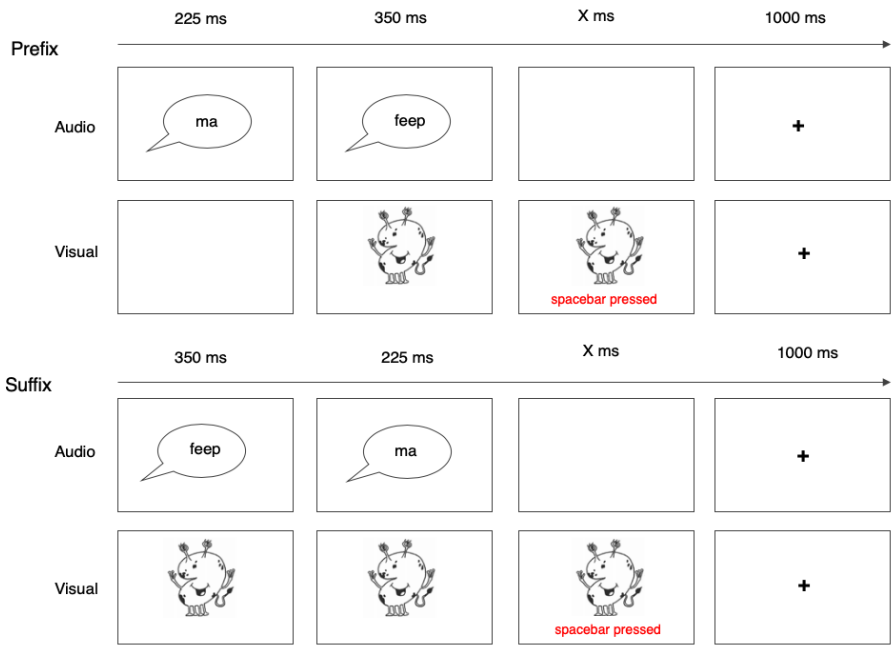


Figure 3.5: Schematic representation of the timing of stimuli presentation in the prefix condition (top panel) and the suffix condition (bottom panel). Verbal labels were presented as sound only, but are written here for illustration.

Generalization tests. Generalization on the basis of semantic and phonological cues was tested using three test-types.

Test 1: semantics-and-affix association and/or semantics-and-phonology association. Pictures depicting two previously unseen aliens were presented on-screen. After 500ms, participants heard a novel noun and the affix that matched the noun in terms of the discriminating phonological feature. Participants were asked to click on the picture that matched the label. The target item had the semantic features associated with the affix and with the phonological features of the noun. The foil item was a novel item from the opposite category but from the same frequency-type. The position of the target and the foil items on-screen was randomized on a trial-by-trial basis. In order to perform above chance in this test, participants need to have learned the association between the noun's semantic cues and the affix, and/or the association between the noun's phonological cues and the noun's semantic cues. Each novel picture (alien) appeared twice, once as a target and once as a foil, giving a total of 16 trials.

Test 2: phonology-and-affix association and/or semantics-to-phonology association. A previously unseen alien was presented on-screen and participants heard two novel noun+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. The foil label matched the target image in terms of the affix, but not in terms of

the vowel. 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 *keed ge* (vowel B + affix A). In order to perform above chance in this test, participants need to have learned the association between the phonological cues and the affix and/or the association between the nouns semantic cues and the nouns phonological cues. Each testing alien appeared as target once, giving a total of eight trials.

Test 3: semantics-and-affix association and/or phonology-and-affix association. Tthis test was identical to Test 2, except that the foil label matched the target image in terms of the vowel, but not 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 affixs, and/or the association between phonological cues and affixes. Each testing alien appeared once, giving a total of eight trials. Note that in tests 2 and 3 we used two different nouns for the target and the foil labels on the grounds that using the same noun might have drawn explicit attention to the affix as the only variable aspect of the trial.

Testing consisted of 40 trials in total: eight trials with all trained items presented in a grid (item-learning test), 16 trials in which they chose between two pictures (generalization test1) and 16 trials in which they chose between two sounds (generalization tests2 and 3 combined).

Language Awareness Questionnaire (LAQ) Participants were asked questions about explicit awareness of the relationship between the aliens and the affixes (semantics and affix association), and between the nouns and the affixes (phonology and affix association) (see Appendix 1). The questionnaire was administered by the experimenter, who wrote down participants responses verbatim.

3.3.3 Results

A total of 1280 data-points were collected, one per each of the 40 test trials from each of the 32 participants.

3.3.3.1 Item learning

The data are shown in Figure 3.6 and the inferential statistics are in Table 3.1. There was strong evidence for the predicted benefit of prefixing. Fitting separate intercepts for each condition showed strong evidence for above-chance performance in the prefix condition, and ambiguous evidence for above-chance performance in the suffix condition.

3.3.3.2 Semantics and Phonology Generalization

The data for all three tests are shown in Figure 3.7 and inferential statistics are in Table 3.2. We predicted better performance in the suffix condition than in the prefix condition. In tests 1 and 2, the evidence for this was ambiguous, whereas in test 3 the evidence was

Table 3.1: Experiment 2a: Item Learning Test Statistics.

Hypothesis	Contrast in the lmer	Mean difference	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect of affix	1.33	0.54	1.00[1]	10.12	[0.29 : >6.54]	.013
Prefix above chance	Intercept-Prefix	1.67	0.37	0.34[2]	159.61	[0.11 : >6.54]	<.001
Suffix above chance	Intercept-Suffix	0.34	0.39	1.67[3]	0.52	[0 : 2.69]	.391

[1]Intercept from the same lme (method B)
[2]Suffix-Intercept from the same lme (method A)
[3]Prefix-Intercept from the same lme (method A)

in the opposite direction - there was strong evidence for better performance in the prefix condition compared to the suffix condition.

Comparing each affix condition to chance showed evidence for above-chance performance in both conditions in test 1; in test 2, the evidence was ambiguous in both conditions, and in test 3, there was evidence for chance-level performance in the prefix condition, and ambiguous evidence in the suffix condition.

3.3.3.3 Language Awareness Questionnaire

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 alines 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 hands", "male or female") were coded as "other".

Twelve participants in the prefix condition (75%) and five participants in the suffix condition (31.25%) reported explicit awareness of the relationship between the aliens/nouns and affixes. This raises the possibility that participants' ability to generalize depended on explicit awareness of the patterns in the language. We plot participants' performance broken down by whether or not they reported awareness (see Figure 3.8). Informal visual inspection of the plot suggests that, for test 1, only those participants who reported explicit awareness performed above chance in both conditions; for test 2, there does not seem to be a relationship between awareness and performance; in test 3, while participants who reported awareness were, on average, near-ceiling in both conditions, it was only in the prefix condition that participants who reported no awareness performed above-chance on average. Due to small samples, particularly for the "unaware" group, statistical analysis was not appropriate here.

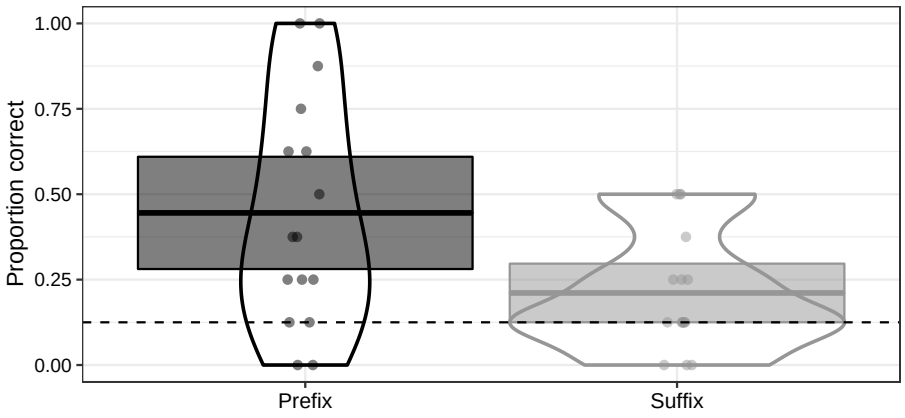


Figure 3.6: Epxeriment 2a: Proportion of correct responses on the Item learning test. Points show mean scores by participant, and violins show the kernel probability density of participants' mean scores. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line is chance-level performance.

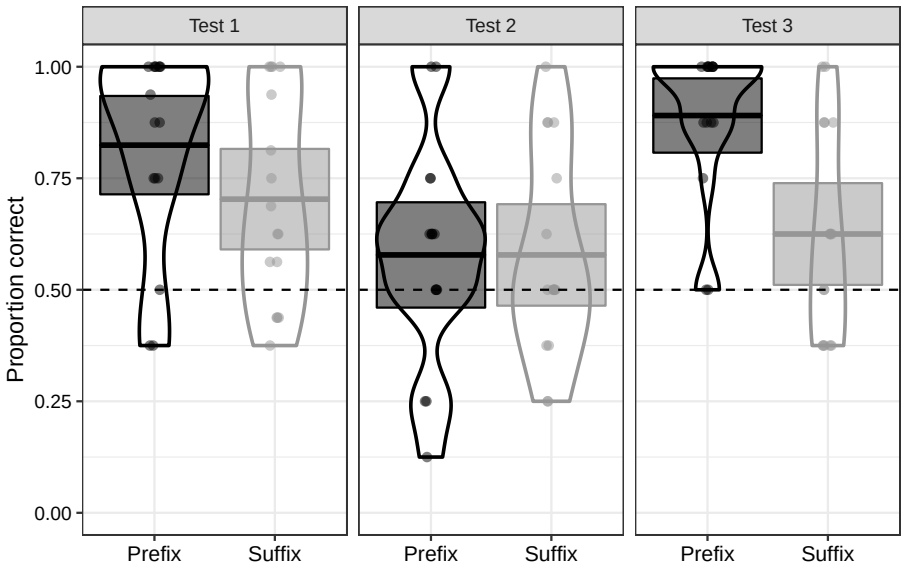


Figure 3.7: Experiment 2a: Proportion of correct responses on the Semantics and Phonology generalization tests 1 to 3 (left to right). Points show mean scores by participant, and violins show the kernel probability density of participants' mean scores. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line is chance-level performance.

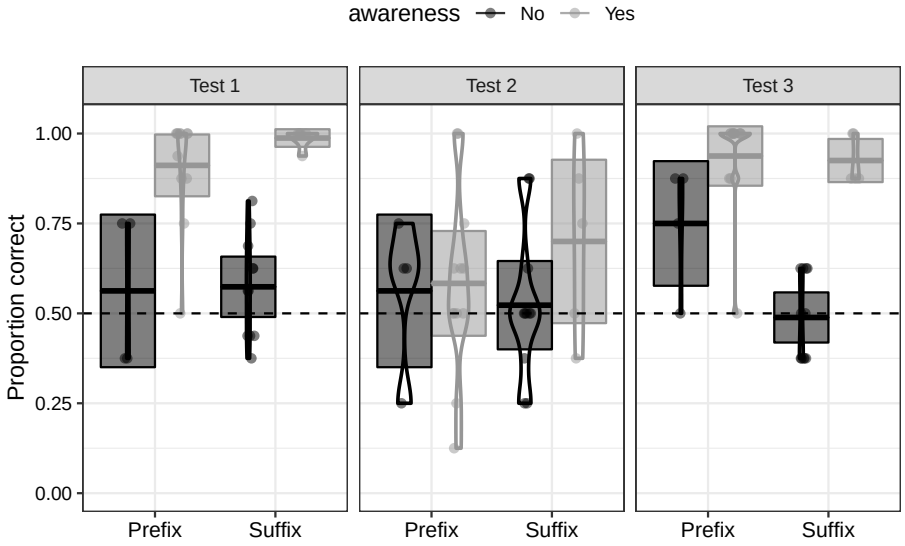


Figure 3.8: Experiment 2a: Proportion of correct responses on the Semantics and Phonology generalization tests 1 to 3 (left to right). Points show mean scores by participant, and violins show the kernel probability density of the mean scores of participants who did not report explicit awareness (black) and for those who did (grey). Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

Table 3.2: Experiment 2a: Semantics and Phonology Generalization Test Statistics.

Hypothesis	**Contrast in the lmer**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Test 1							
Suffix better than Prefix	Main effect of affix	1.07	0.69	1.86[1]	1.87	[0 : >4.591]	.12
Prefix above chance	Intercept-Prefix	2.40	0.54	1.33[2]	3993	[0 : >4.591]	<.001
Suffix above chance	Intercept-Suffix	1.33	0.48	2.4[3]	16.24	[0: >4.591]	.005
Test 2							
Suffix better than Prefix	Main effect of affix	-0.00	0.37	0.35[1]	0.72	[0 : 1.031]	.996
Prefix above chance	Intercept-Prefix	0.35	0.26	0.36[2]	1.86	[0 : 3.541]	.173
Suffix above chance	Intercept-Suffix	0.36	0.26	0.35[3]	1.89	[0 : 3.631]	.171
Test 3							
Suffix better than Prefix	Main effect of affix	1.89	0.54	1.57[1]	150.98	[0 : >4.591]	<.001
Prefix above chance	Intercept-Prefix	2.51	0.46	0.62[2]	22767	[0 : >4.591]	<.001
Suffix above chance	Intercept-Suffix	0.62	0.32	2.51[3]	1.54	[1.111 : >4.591]	.054

[1]Intercept from the same lme (method B)
[2]Suffix-Intercept from the same lme (method A)
[3]Prefix-Intercept from the same lme (method A)

3.3.4 Discussion

In Experiment 2a, participants were exposed to an artificial language in which affix occurrence was conditioned on the visual features of the items and the phonological features of the nouns. We predicted better item learning in the prefix condition, and better generalization in the suffix condition.

As predicted, item learning was better in the prefix condition than in the suffix condition. Contrary to the prediction, we did not find evidence for better generalization in the suffix condition compared to the prefix condition, with ambiguous evidence for tests 1 and 2, and strong evidence in the opposite direction (better performance in prefix compared to suffix) for test 3. We begin with discussion test 2, before turning to tests 1 and 3.

Performance was overall poor in test 2, with no clear evidence for above-chance performance in either condition. Recall that to perform above chance in this test, participants had to have learned the associations between the phonological features of the noun and the affix, and/or the association between the phonological features and the semantic features. Poorer performance on this test compared to the other two tests, which tested the learning of the semantic cues, suggests that participants in both conditions were better at learning the semantic cues than phonological cues.

There was evidence for better performance in the prefix condition compared to the suffix condition in test 3, whereas in test 1, even though the means were in the same direction, the evidence was ambiguous. An informal analysis of the post-experiment questionnaire, however, suggested that participants' generalization reflected explicit learning of the patterns in the language. For each test, we computed the average performance and confidence intervals for participants grouped based on whether or not they reported explicit awareness. In the suffix condition, those participants who reported explicit awareness of the relationship between the shape and the affix on average performed to ceiling, whereas those participants who did not report awareness averaged near-chance. In the prefix condition, this was true for test 1, whereas in test 3 the four participants who did not report explicit awareness on average performed above chance. Across the two tests and the two conditions, however, it appears that generalization may be contingent on explicit awareness. Moreover, participants who reported explicit awareness make up a large majority, which suggests that our paradigm may have been particularly conducive of explicit learning. Note, however, that to the best of our knowledge, most artificial language learning experiments do not tend to probe explicit awareness, and this large proportion of participants in our study who did report explicit awareness might be typical in such paradigms. Nevertheless, adult artificial language learning experiments are typically not self-paced, which raises the possibility that the self-paced training encouraged explicit learning, which is an issue for the current work, where we make claims about first language learning in a natural setting, which is driven by implicit learning mechanisms. A high number of participants reporting explicit awareness makes it difficult to draw conclusions that may be relevant for first language learning.

Note that we interpret these informal findings with caution, as we were unable to run statistical analyses. We also 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 pattern, but were unable to verbalize them (due to the objects being novel) and/or unwilling, although this would suggest that the current findings may be an underestimate, rather than an overestimate of explicit awareness. It is also possible that explicit awareness emerges though implicit learning mechanisms, and our current method does not allow us to pull apart different learning mechanisms with confidence. Finally, the sample size in the current study is small. Therefore, before we can draw conclusions as to our generalization hypothesis, we believe it is necessary to replicate this study with a larger sample. We may find that greater statistical power is required in order to see better generalization for those participants without explicit awareness, and/or a more ballanced ratio of explicit vs. implicit learners. Therefore, we will replicate Experiment 1a with an appropriately powered sample. In the replication experiment, we will also include new tests which assess the generalization of the phonological cue (*affix* → *vowel*) alone, in order to shed light on the poor performance in test 2, which suggested weak learning of the phonological cues. We also add a test analogous to the generalization test 3, but with trained items - this will further inform our understanding of any learning we may observe. Finally, we will also slightly adjust the training timings to make the conditions more balanced. In Experiment 1a, in the suffix condition, the picture remained on-screen while the affix was played, whereas in the prefix condition, the affix was played before the picture appeared. While we have no theoretical reason to believe that this difference between the conditions resulted in no suffixing advantage, it does mean that the two conditions did not receive the exact same training (the only difference being the order in which individual components were presented). Therefore in the next experiment, the prefix condition will remain the same, but in the suffix condition, the picture will disappear before the affix is played (see Section 3.4.1.3 for detail).

The next experiment, as well as the remainder of this thesis, will be web-based: participants will be recruited through a third-party web service, they will complete the experiment on-line and their data will be stored on a third-party web platform. One of the methodological reasons for using on-line experiments in this thesis was to increase sample sizes, however, the sample in the next experiment is still relatively small (22 participants per condition), as this was an early test of the paradigm.

3.4 Experiment 2b

3.4.1 Method

3.4.1.1 Participants

Fourty-four participants (22 per condition; Mage = 34.54, SD = 9.56, 22 female) were recruited through Prolific Academic, and participated on-line. An additional ten participants were tested, of which: two were excluded due to taking longer than 40 minutes to complete the study, and the remaining eight for failing attention check trials (see Section 3.4.1.3).

All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated

to one of the two affix conditions. They provided informed consent and were paid for participation.

3.4.1.2 Stimuli

Same as Experiment 2a.

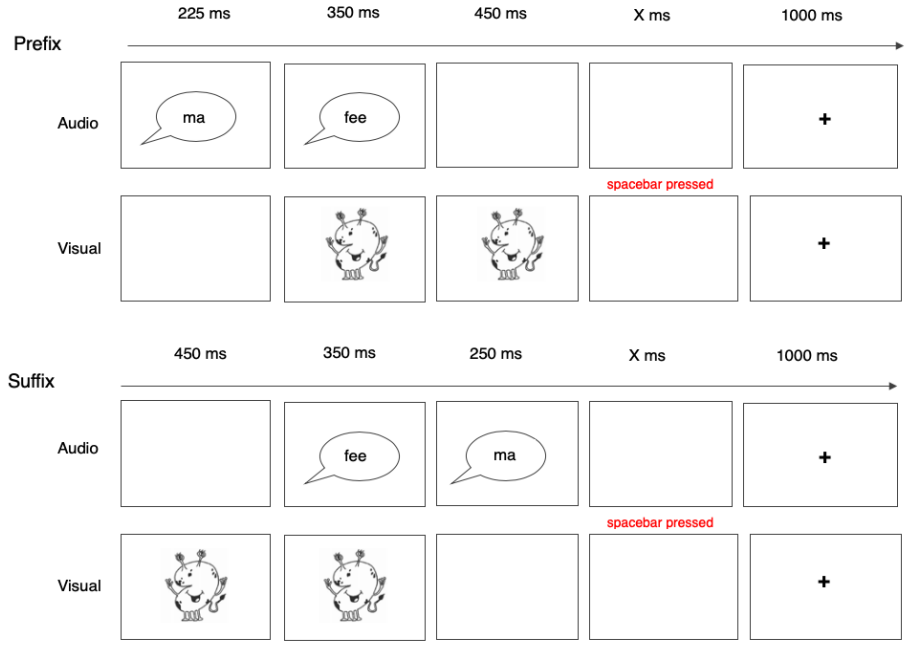


Figure 3.9: Schematic representation of the timing of stimuli presentation in the prefix condition (top panel) and the suffix condition (bottom panel). Verbal labels were presented as sound only, but are written here for illustration.

3.4.1.3 Procedure

Same as Experiment 2a, except that the timing of the training procedure was adjusted (see below), and two new tests were added at the end of the experiment.

Training This was same as Experiment 2a in terms of the number of trials and exposures per trial, but the timing of stimuli presentation changed. In the suffix condition, the picture was no longer present when the affix was played. However, this would mean that the picture was only present for the duration of the noun (350ms), which was shorter than in the prefix condition. To avoid this, we first presented the picture for 450ms at the beginning of the trial; after this, the noun was played. Once the noun ended, the picture disappeared and the suffix was played. To balance the two conditions in terms of the amount of exposure to the picture, in the prefix condition, the picture was presented after the prefix was played, remained on-screen for the duration of the noun, and for an additional 450ms. After this, the picture disappeared, and the participant could press the space bar to move on to the next trial (see Figure 3.9). Therefore, 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.

Testing

Semantics and Phonology Test: Trained items. The set-up was identical to Semantics and Phonology Generalization test 3, except that (i) we used items that had occurred in training (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 noun-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. Notice 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 eight trials.

Phonology Generalization Test. We tested the learning of phonological cues in isolation by playing two novel labels (using two speech bubbles, as described in Section 3.3.2.3), but without 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. There were two test-types depending on the foil - both test knowledge of the relationship between the phonological cue in the noun (the vowel) and the affix:

Test 1: Participants heard two affix+noun bigrams one after the other with a 500ms silence between them. The noun was identical in each case but the affix differed. For example, if the target label was *feep ge* (categoryA vowel + categoryA affix), the foil label was *moop ge* (categoryB vowel + categoryA affix). Note that this is identical to Semantics and Phonology Generalization Test 2, but without a picture. The target nouns were the same eight nouns used in to Semantics and Phonology Generalization Test 2, with a total of 8 trials.

Test 2: Participants heard two affix+noun bigrams one after the other. 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 3, but without a picture. The target nouns were therefore the same eight nouns used in to Semantics and Phonology Generalization Test 2, with a total of 8 trials.

Attention check trials. At the end of the whole experiment, as part of the Phonology Generalization test, four trials were included in which the foil label was clearly not a label from the language. The four labels were: *lopipa*, *gemule*, *mugugo*, and *nemuba*, presented in random order with one of the testing labels selected at random. These trials were introduced to screen out inattentive participants. Participants who did not respond correctly to all of the four check trials were excluded.

Language Awareness Questionnaire The questionnaire was the same as the one used in Experiment 1a, except that, rather than being administered by the experimenter, participants responded themselves on their computers. The questionnaire was presented on participants 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.

3.4.2 Results

A total of 1408 data-points were collected, one per each of the 64 test trials from each of the 44 participants. Recall that data from eight participants were excluded due to failing the attention-check trials. Of these, three participants did not respond correctly to any of the four trials (one in the prefix condition and two in the suffix condition), three responded correctly to 50% of the trials (one in the prefix and two in the suffix condition), and two participants, both in the suffix condition, failed one of the four trials.

We present analyses both with the data from Experiment 2b, as well as, where appropriate, combined data from Experiments 2a and 2b. In both cases, values from within the data were used to inform the H1 (for Experiment 2b, this was done as we believed the internal data would provide a more appropriate constraint of the plausible maxima, as the performance in Experiment 2a may have been more strongly driven by explicit learning).

3.4.2.1 Item learning

The data are presented in Figure 3.10 and inferential statistics are in Table 3.3. As predicted, there was strong evidence for better item learning in the prefix condition compared to the suffix condition, and this was also the case with joined data. There was strong evidence for learning in the prefix condition, and, unlike in Experiment 2a, there was evidence for no learning in the suffix condition (true for combined data, too).

3.4.2.2 Semantics and Phonology Generalization

The data for all three tests are shown in Figure 3.11 and inferential statistics are in Table 3.4. We predicted better performance in the suffix condition than in the prefix condition.

In tests 1 and 3, there was evidence for the opposite, that is, for better performance in the prefix condition compared to the suffix condition (and this was also true for combined data). In test 2, the evidence was ambiguous.

Comparing each affix condition to chance showed evidence for learning in the prefix condition in tests 1 and 3, whereas in test 2 the evidence was ambiguous, and the same pattern was observed with combined data from Experiments 2a and 2b. In the suffix condition, the evidence was ambiguous in tests 1 and 2, whereas in test 3 there was evidence for the null. For combined data, the evidence remained ambiguous for test 2, but for test 1 there was now evidence for learning, whereas in test 3 the evidence was now ambiguous.

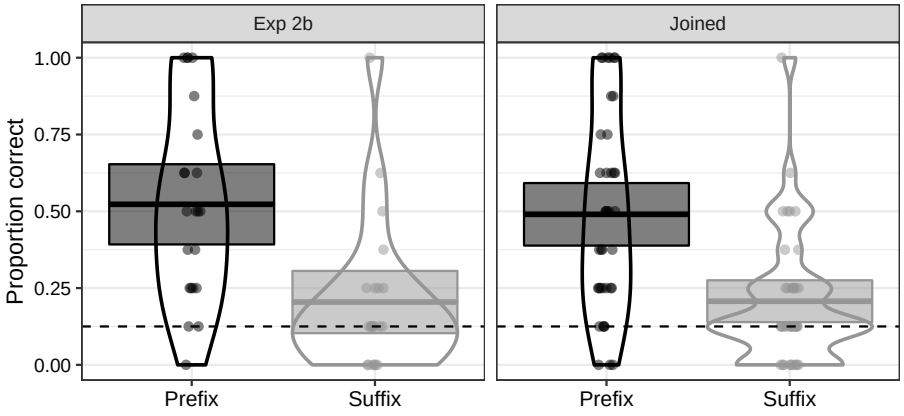


Figure 3.10: Experiment 2b: Proportion of correct responses on the Item learning test in Experiment 2b (left) and with joined data from Experiments 2a and 2b (right). Points show mean scores by participant, and violins show the kernel probability density of participants' mean scores. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

Table 3.3: Experiment 2b: Item Learning Test Statistics.

Hypothesis	**Contrast in the lmer**	**Data**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Prefix better than Suffix	Main effect affix	Exp 2b	1.96	0.53	1.33[1]	252.31	[0.17 : >6.54]	<.001
		Joined	1.69	0.38	1.07[1]	3725	[0.1 : >6.54]	<.001
Prefix above chance	Prefix Intercept	Exp 2b	2.11	0.35	0.15[2]	28.28	[0.07 : >6.54]	<.001
		Joined	1.92	0.26	0.23[2]	3.16×10^5	[0.04 : >6.54]	<.001
Suffix above chance	Suffix Intercept	Exp 2b	0.15	0.39	2.11[3]	0.26	[1.591: ∞]	.704
		Joined	0.23	0.28	1.92[3]	0.31	[1.831: ∞]	.42

[1]Intercept from the same lme (method B)
[2]Suffix-Intercept from the same lme (method A)
[3]Prefix-Intercept from the same lme (method A)

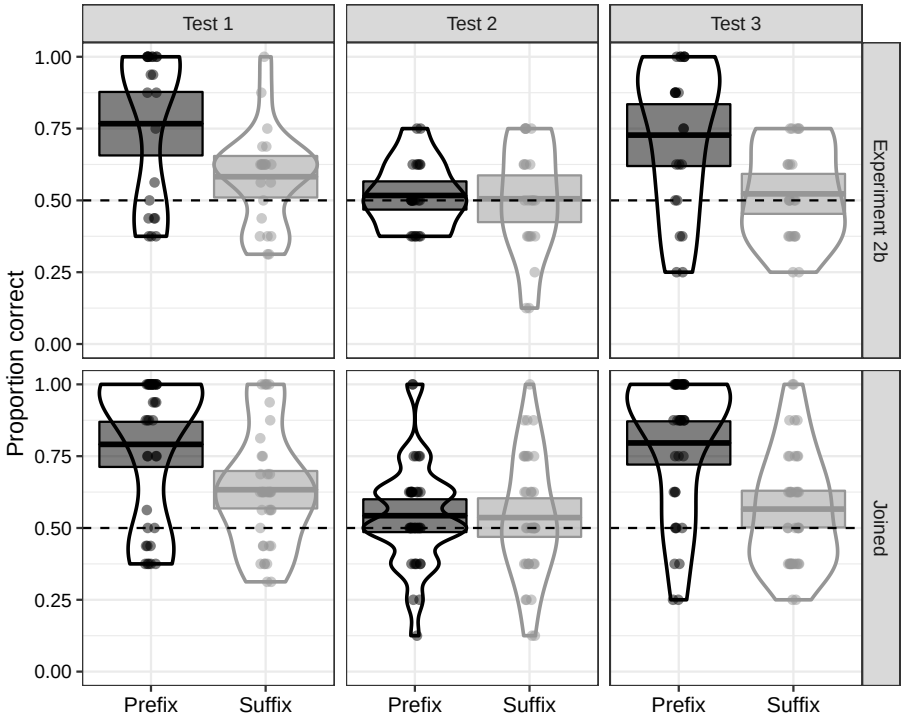


Figure 3.11: Experiment 2b: Proportion of correct responses on the Semantics and Phonology generalization tests 1 to 3 (left to right) in Experiment 2b (top) and with joined data from Experiments 2a and 2b (bottom). Points show mean scores by participant, and violins show the kernel probability density of participants' mean scores. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed lines indicate chance-level performance.

Table 3.4: Experiment 2b: Semantics and Phonology Generalization Test Statistics.

Hypothesis	Contrast in lme	Data	Mean difference	SE	H1	B	Robustness region	p
Test 1								
Suffix better than Prefix	Main effect affix	2b	-1.38	0.45	1.10[1]	0.10	[0.29 : ∞]	.002
		Joined	-1.29	0.39	1.40[1]	0.06	[0.23 : ∞]	.001
Prefix better than Suffix	Main effect affix	2b	1.38	0.45	1.07[1]	40.79	[0.18 : >4.591]	.002
		Joined	1.29	0.39	1.40[1]	94.77	[0.14 : >4.519	.001
Prefix above chance	Prefix Intercept	2b	1.79	0.34	0.41[2]	4220.80	[0.08 : >4.591]	<.001
		Joined	2.04	0.29	0.76[2]	1.53×10^{9}	[0.05 : >4.591]	<.001
Suffix above chance	Suffix Intercept	2b	0.41	0.30	1.79[3]	0.75	[0 : 4.14]	.171
		Joined	0.76	0.26	2.04[3]	16.99	[0.11 : >4.591]	.004
Test 2								
Suffix better than Prefix	Main effect affix	2b	-0.04	0.21	0.04[1]	0.95	[0 : 0.5]	.831
		Joined	-0.03	0.18	0.16[1]	0.69	[0 : 0.45]	.882
Prefix better than Suffix	Main effect affix	2b	0.04	0.21	0.04[1]	1.00	[0 : 0.5]	.831
		Joined	0.03	0.18	0.16[1]	0.81	[0 : 0.58]	.882
Prefix above chance	Prefix Intercept	2b	0.07	0.15	0.02[2]	1.05	[0 : 0.64]	.651
		Joined	0.18	0.13	0.15[2]	1.89	[0 : 1.76]	.172
Suffix above chance	Suffix Intercept	2b	0.02	0.15	0.07[3]	0.96	[0 : 0.49]	.88
		Joined	0.15	0.13	0.18[3]	1.50	[0 : 1.29]	.247
Test 3								
Suffix better than Prefix	Main effect affix	2b	-1.01	0.32	0.60[1]	0.12	[0.2 : ∞]	.002
		Joined	-1.31	0.29	0.97[1]	0.05	[0.13 : ∞]	
Prefix better than Suffix	Main effect affix	2b	1.01	0.32	1.89[1]	40.03	[0.13 : >4.591]	.002
		Joined	1.31	0.29	0.97[1]	6279	[0.08 : >4.591]	<.001
Prefix above chance	Prefix Intercept	2b	1.11	0.24	0.10[2]	8.32	[0.06 : >4.591]	<.001
		Joined	1.62	0.23	0.31[2]	2.37×10^{7}	[0.04 : >4.591]	<.001
Suffix above chance	Suffix Intercept	2b	0.10	0.21	1.11[3]	0.29	[0.95 : inf]	.644
		Joined	0.31	0.19	1.62[3]	0.81	[0 : 4.05]	.104

[1]Intercept from the same lme (method B)
[2]Suffix-Intercept from the same lme (method A)
[3]Prefix-Intercept from the same lme (method A)

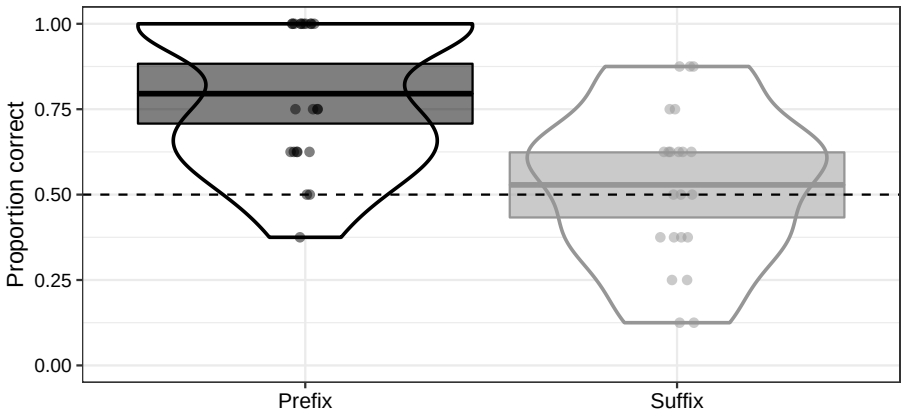


Figure 3.12: Experiment 2b: Proportion of correct responses on the Semantics and Phonology trained items test. Points show mean scores by participant, and violins show the kernel probability density of participants' mean scores. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed lines indicate chance-level performance.

Table 3.5: Experiment 2b: Semantics and Phonology Trained Items Test Statistics.

Hypothesis	**Contrast in the lmer**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Prefix better than Suffix	Main effect affix	1.45	0.37	0.85[1]	460.52	[0.11 : >4.591]	<.001
Prefix above chance	Prefix Intercept	1.58	0.28	0.13[2]	25.06	[0.06 : >4.591]	<.001
Suffix above chance	Suffix Intercept	0.13	0.24	1.58[3]	0.25	[1.15 : ∞]	.592

[1]Intercept from the same lme (method B)
[2]Suffix-Intercept from the same lme (method A)
[3]Prefix-Intercept from the same lme (method A)

3.4.2.3 Semantics and Phonology: Trained items

The data are shown in Figure 3.12 and inferential statistics are in Table 3.5. There was evidence for better performance in the prefix condition compared to the suffix condition. There was evidence for learning in the prefix condition, whereas in the suffix condition there was evidence for the null.

3.4.2.4 Phonology Generalization

The data are shown in Figure 3.13 and inferential statistics are in Table 3.6. The evidence for better performance in the suffix condition was ambiguous in both tests. Breaking down by affix condition showed ambiguous evidence for learning in the suffix condition in test1, and evidence for no learning in test2. Due to no clear evidence for learning in the suffix condition, it was not appropriate to use these values to model the H1 for the prefix condition. Therefore, in the prefix condition we computed Bayes Factors for the whole range of plausible values, and used this for indirect inference - the evidence was ambiguous in test1 for any value up to 1.08 (corresponding to), whereas in test2 there was evidence

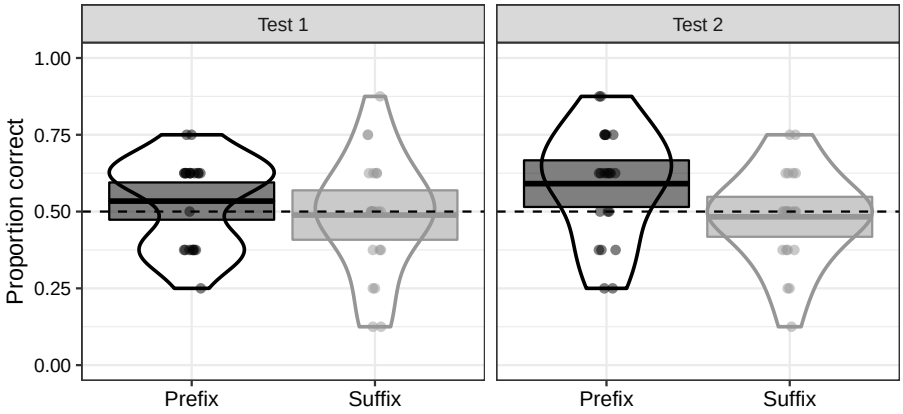


Figure 3.13: Experiment 2b: Proportion of correct responses on the Phonology generalization tests 1 and 2 (left to right). Points show mean scores by participant, and violins show the kernel probability density of participants' mean scores. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed lines indicate chance-level performance.

for learning for any value between 0.09 and 1.77.

3.4.2.5 Language Awareness Questionnaire

Participants' responses were coded as described in Section 3.3.3.3.

Twelve participants in the prefix condition (54.5%) and seven participants in the suffix condition (31.8%) reported explicit awareness of the relationship between the aliens/nouns and affixes. We were therefore interested in whether the prefix benefit in generalization, observed in tests 1 and 3, was driven by those participants in the prefix condition who reported explicit awareness (Figure 3.14). Removing participants who reported explicit awareness in both conditions, the evidence for the prefix benefit was ambiguous in both tests (test1: $\beta = 0.126$, $SE = 0.279$, $B = 12.8$, $p = .651$; test3: $\beta = 0.488$, $SE = 0.359$, $B = 2.59$, $p = .175$).

3.4.3 Discussion

Experiment 2b was a replication of Experiment 2a. Recall that in Experiment 2a, contrary to the original prediction, there was evidence for better generalization in the prefix condition compared to the suffix condition. One possibility was that this was caused by the differences in the timings of stimuli presentation during training: specifically, in the suffix condition, the picture was not present when the affix was played, whereas in the prefix condition it was, which suggested that the prefixing benefit may be attributed to a reduced perceptual load. The timings were therefore adjusted in Experiment 2b, such that the affix was presented alone (without the picture present) in both conditions. The results from Experiment 2b rule out the possibility that the prefix benefit observed in Experiment 2a was an artefact of the timings – in 2b, there was evidence for a prefix advantage in test 3, and also in test 1 (in Experiment 2a this was ambiguous). The evidence remained ambiguous in test 2, where, like in Experiment 2a, learning was overall poor, with ambiguous evidence for learning in

Table 3.6: Experiment 2b: Phonology Generalization Test Statistics.

Hypothesis	Contrast in the lmer	Mean difference	SE	H1	B	Robustness region	p
Test 1							
Suffix better than Prefix	Main effect affix	-0.18	0.21	0.046[1]	0.853	[0 : 0.32]	.394
Prefix above chance	Prefix Intercept	0.14	0.15	–	–	ambig: [0 : 1.08], H1: [1.09: ∞]	.366
Suffix above chance	Suffix Intercept	-0.04	0.15	0.137[2]	0.638	[0 : 0.33]	.763
Test 2							
Suffix better than Prefix	Main effect affix	-0.44	0.21	0.15[1]	0.397	[0 : 0.18]	.043
Prefix above chance	Prefix Intercept	0.37	0.15	–	–	H1: [0.09 : 1.77], ambig: [1.78 : >4.591]	.016
Suffix above chance	Suffix Intercept	-0.07	0.15	0.368[2]	0.282	[0.31 : ∞]	.651

[1]Intercept from the same lme (method B)
[2]Prefix-Intercept from the same lme (method A)

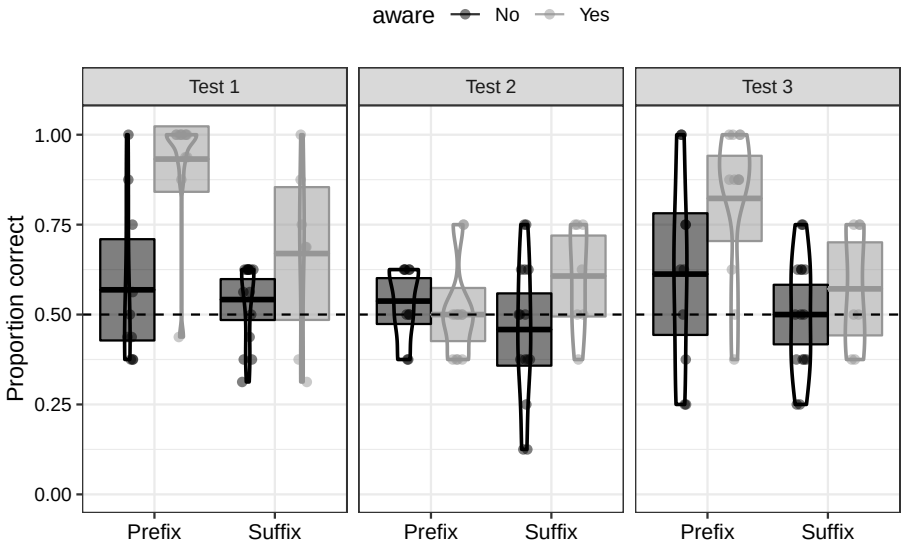


Figure 3.14: Proportion of correct responses on the Semantics and Phonology generalization tests 1 to 3 (left to right). Points show mean scores by participant, and violins show the kernel probability density of the mean scores of participants who did not report explicit awareness (black) and for those who did (grey). Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

both conditions.

As discussed in Experiment 2a, the poor performance on test 2 might be due to the fact that all the useful cues for generalization in this test are related to the phonological cues in the input, which may have been harder for participants to learn. To elucidate this, Experiment 2b included two tests which test the learning of phonological cues in the absence of semantic (visual) cues. Here, performance was low in both conditions, although numerically it was better in the prefix condition than in the suffix condition (contrary to the predicted direction). While there was ambiguous or no evidence for learning in the suffix condition, there was evidence for learning in the prefix condition in test 2, although the average performance was low. Therefore, we conclude that the learning of phonological cues in the experiment was poor. The implications of this finding are discussed in the General Discussion of Study 1(Section 3.5).

Turning to item learning, in Experiment 2b, like Experiment 2a, there was evidence for a prefixing advantage in the Item learning test, which is consistent with our predictions. Experiment 2b included an additional test of item learning, the Semantics and Phonology Trained items test. While the Item learning test tested participants' learning of the association between individual visual referents and trained noun+affix bigrams, in the Trained items test we were interested in participants' ability to match the trained items (nouns and pictures) with the correct affix. On this test too, there was evidence for better performance in the prefix condition compared to the suffix condition. Note, however, that while participants can perform above chance on this test based on their memory of trained items, it is also possible to perform above chance without any item-based learning, but entirely on the basis of learning the discriminating features. The original idea with this test was that, if participants in the prefix condition perform poorly on generalization (as predicted) but well on the trained items test, this would be an indication of strong item-based learning in this condition. However, this is not what we observed, and therefore we cannot rule out the possibility that the performance on generalization and trained items tests reflects the same kind of learning – strong learning of discriminating features in the prefix condition.

In the General Discussion of this study, the results from this experiment are discussed together with Experiment 2a, and implications for the key questions in this thesis are considered.

3.5 General Discussion of Study 1

The aim of the two experiments presented in this chapter was to design and test a paradigm to use in further work testing two predictions of the discriminative learning framework: that suffixing promotes better learning of discriminating features, which in turn results in better generalization, whereas prefixing promotes better item-learning. We conducted two experiments: Experiment 2a was run in the lab, and Experiment 2b was an on-line replication with an extended paradigm. The results of the two experiments were largely consistent. Below we discuss each of the two predictions with respect to each of the two experiments, however, following Dienes (2016), we consider evidence from a larger sample

to be more robust, and so we draw our conclusions on the basis of the joined data.

Starting with item-learning, as predicted, this was better in the prefix condition compared to the suffix condition. This was the case in both experiments as well as with combined data. Experiment 2b also included an additional test of the knowledge of trained items. Participants could answer this test relying on item-based knowledge or based on learning the discriminating features. If performance is based on the discriminating features, we predict no difference between trained and new items, however, we found evidence for stronger performance with trained items, suggesting that participants relied on item-based knowledge, and possibly also on the knowledge of discriminative features (the two are not mutually exclusive). While the performance was stronger in the prefix condition than the suffix condition (consistent with the same effect in the item-learning test), there was also evidence for learning in the suffix condition, whereas the evidence was ambiguous in the equivalent test with novel items (generalization test 3).

Turning to generalization, contrary to the prediction, we did not find evidence for a suffixing advantage. In Experiment 2a, the evidence was ambiguous in tests 1 and 2, whereas test 3 showed evidence in the opposite direction – better performance in the prefix condition than in the suffix condition. In Experiment 2b, there was evidence for a prefixing advantage in tests 1 and 3, but the evidence was ambiguous in test 2, and the same pattern was observed with joined data. Note that we cannot interpret an ambiguous result with respect to test 2, however, it is useful to reflect on why performance on this test was on average poorer than performance on tests 1 and 3 (although this was not explicitly compared statistically). Below we discuss two aspects of these findings: first, overall poorer performance in test 2 than in tests 1 and 3, and second, the prefixing advantage in tests 1 and 3, which is not predicted by our theory.

Recall that test 2 is the one generalization test where the association between discriminating semantic features and affixes was not a useful generalization cue. Instead, the test required knowledge of either (or both) (i) the association between discriminating phonological features of the nouns (the vowel) and the affix or (ii) the association between those phonological features and the semantic features. Experiment 2b included two tests equivalent to tests 2 and 3 but presented without pictures, meaning that learners could respond correctly to both tests only if they had learned (i). Here, there was evidence for learning in the prefix condition, but in the suffix condition, the evidence was either ambiguous (phonology test 2), or there was evidence for no learning (phonology test 1). It is possible that performance on phonology-only tests was poorer than performance on generalization tests 1 and 3 simply because the phonology tests provided fewer useful generalization cues to participants. The design of the phonology-only tests may have made them harder, too, as participants were required to hold unfamiliar labels in working memory without a picture to associate them with. However, if we consider the phonology-only tests together with the Semantics and Phonology test 2, where performance was also poorer that on tests 1 and 3, this raises the possibility that participants found phonology-based cues (in either test) harder to learn or less salient than semantics-based cues. The role of different cues in linguistic generalization has been a topic of extensive debate in the literature, and is a

relevant and interesting question that I return to in the General Discussion of Chapter 4.

Moving to tests 1 and 3, there was evidence for better performance in the suffix condition than in the prefix condition. This finding is not predicted by our theory, and is inconsistent with previous work addressing this question (St Clair et al., 2009; Ramscar et al., 2010; Ramscar, 2013). However, our study differed in several important aspects to the previous work, and this may have resulted in the surprising results. I discuss these aspects in below.

Unlike previous work, our exposure task was self-paced (St Clair et al. (2009) and Ramscar et al. (2010) used timed exposure), and participants were encouraged to repeat out-loud what they heard (although in Experiment 2a the experimenter was not in the cubicle and therefore did not check whether participants did this, and Experiment 2b was web-based and no spoken data were collected). This was done primarily to create a child-friendly paradigm suitable for future use (following Wonnacott, 2011; Wonnacott, Boyd, Thomson, & Goldberg, 2012, who used a similar design with child learners). It is possible that the self-paced training encouraged explicit learning, and that the observed prefixing advantage can be explained by explicit awareness of the associations in the language. To this end, we used a post-experiment questionnaire, in which we asked participants about noticing patterns in the language. In Experiment 2a, visual analysis and summary statistics suggested that it was indeed the case that in both conditions those participants who reported explicit awareness showed high performance, whereas those who reported no awareness performed more poorly or were at chance. The higher average in the prefix condition could therefore be due to more participants reporting explicit awareness in this condition than in the suffix condition. It is possible that strong item-learning observed in the prefix condition (stronger than in the suffix condition) boosted generalization via explicit mechanisms. However, it is perhaps more likely that greater explicit awareness in the prefix condition was due to chance sampling, particularly considering the small sample size in this study.

In Experiment 2b, more participants reported explicit awareness in the prefix condition than in the suffix condition, although proportionally this difference was considerably smaller than in Experiment 2a. While on average participants who reported awareness performed better than participants who did not, this alone did not explain the prefixing advantage in test 1, where excluding participants who reported explicit awareness in both conditions still resulted in better performance in the prefix condition. In test 3, excluding participants who reported awareness resulted in ambiguous evidence (which can be expected in a relatively small sample). Therefore, while it is the case that many participants were explicitly aware of the association between the body shapes and the affixes, this alone does not explain the prefixing advantage we observed. Even with those participants removed, there was still evidence for better performance in the prefix condition in at least one test, meaning that the question still remains – why was generalization better in the prefix condition than in the suffix condition? Before turning to this, I point out again that the results from the questionnaire are interpreted with caution, not in the least because it is possible that the relationship between generalization and answering the questionnaire is reverse – it might be that participants who are generally strong learners and/or motivated to do well

in psychological experiments pay more attention (explicit or implicit) to the experiment, perform well, and are in turn more motivated to fill out the questionnaire diligently.

While the present study may have elicited more explicit learning than previous work (although we cannot be sure of this as no similar assessments of explicit learning were used in previous work), this study differs from previous work with respect to the design of the input set, which in retrospect may not have been most appropriate for studying the effects of prediction error on learning. Specifically, the stimuli in this set of experiments were relatively low-dimensional compared to the stimuli in previous work. (St Clair et al., 2009) used verbal labels where the discriminative features were at the level of phonological features, rather than individual phonemes – the individual sounds varied across words, but words in one category contained consonant clusters, unrounded high vowels, nasals, and stops (e.g., *gwemb*, *dreng*, *klimp*); whereas the words in the other category had no consonant clusters, had rounded low vowels, and fricatives (e.g., *zodge, shufe, foth*). Such cue structure suggests that there were many cues to compete against each other. By comparison, in our study, the discriminative features were two different vowels, and there was otherwise little variation across words, meaning that the labels may not have provided rich cue structure.

With respect to the visual stimuli, it is possible that the key discriminating features in the current study were much more salient than the other features. We used black-and-white hand-drawn images of "aliens", where the body was the largest part of the stimulus, and the varying elements – number of legs, eyes, and the shape of the hands, may have been too low-dimensional for participants to pick up on, especially since the potentially most salient cue was also the most predictive cue to affix usage. In (Ramscar et al., 2010), on the other hand, the most salient features (item body shape and colour) were used inconsistently with respect to the label, and so the learning of the correct discriminative features involved "unlearning" the salient features through prediction error. Therefore, by aligning feature saliency, frequency, and informativeness, as opposed to contrasting it as was done in Ramscar et al. (2010), we may have obscured the effects of learning via prediction error. In our input set, successful generalization did not require "unlearning" other, less salient and less consistent stimuli. Generalization merely required learning two perfect correlations: shape1 $\rightarrow$ affix1, and shape2 $\rightarrow$ affix2. This required learning of the kind involved in label-feature or "prefix" learning (Section 1.4) – computing co-occurrences in the input from positive evidence alone. Although this was not predicted by the simulation presented in Experiment 1, I return to this point in Chapter 5 and demonstrate in a new series of simulations that the prefix condition may be better than the suffix condiyion at learning perfectly correlated cues is the absence of salient *non-discriminating* cues.

To summarize, in Experiments 2a and 2b, there was evidence for the predicted prefixing advantage in item learning, but contrary to prediction, generalization was also better in the prefix condition. I interpret this result primarily with respect to explicit learning as well as the limitations of the input set we designed – while our aim to create the most learnable input set possible was appropriate given our plans to use it with child learners, the input set may not have been rich enough to study the complex effects of cue competition and prediction error on generalization. Therefore in the next chapter, we design an input set in

which generalization requires dissociating uninformative cues – in theory, this was also true in the current experiment, however, the effect was harder to observe as the uninformative cues were relatively less salient, and less frequent, than the informative cue. To that end, we design an artificial language similar to that used in Ramscar et al. (2010). To gain a more complete understanding of the predictions, we also train two computational models on the same artificial language input using the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972).

Chapter 4

Study 2: Experiments 3 – 6

4.1 Introduction

The key premise of this thesis is that generalization occurs via discriminative learning through cue competition. Based on previous work, as well as the performance of two computational models implementing the Rescorla-Wagner learning rule, which captures the basic principles of discriminative learning, we predicted better generalization via cue competition in the suffix condition compared to the prefix condition. Contrary to this prediction, however, Study 1 of this thesis found a prefixing benefit in generalization. One possibility is that clearest benefits of suffixing are observable where appropriate generalization involves dissociating frequent, salient, but nonpredictive features, and thus boosting the key informative features, which was not the case in Study 1. This is why in the experiments in this study, we borrowed the design of Ramscar et al. (2010), in which successful generalization involves "unlearning" frequent, but uninformative cues (which can only occur if there is cue competition), and adapted it for the learning of suffixes and prefixes. The nouns and affixes of the artificial language were the same as the ones used in Study 1, but we used the visual stimuli from Ramscar et al. (2010) called *fribbles*[1], and manipulated the semantic features in manner directly analogous to that study: there were two sub-classes of items associated with each of the two affixes, which we call "low type-frequency" and "high type-frequency". Critically, high type-frequency items made up 75% of the items which co-occurred with the same affix and shared a visual feature (body shape) with the low type-frequency items which co-occurred with the opposite affix. Thus we simulated a learning environment in which appropriate generalization involves unlearning a frequent, salient feature, due to it being uninformative.

The chapter begins with Experiment 3, which presents another series of simulations implementing the Rescorla-Wagner model of learning (Rescorla Wagner, 1972). As in Experiment 1, two computational simulations were trained on input conceptually equivalent to the newly designed artificial languages (see Figure 4.1). Experiments 4-6 were done with humans[2]; participants were trained on the languages described above, but the number

[1]Stimulus images courtesy of Michael J. Tarr, Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, retrieved from https://wiki.cnbc.cmu.edu/Novel$_O$*bjects*

[2]Experiments 4 and 5 of this chapter were written up as a manuscript submitted to the *Journal of*

of exemplars in the training set was manipulated across the experiments. This revealed interesting interactions with generalization, discussed in Section 4.7.6.

Pre-registered hypotheses for Experiments 4 and 5 can be found here: https://rpubs.com/MasaVujovic, and all the data and R analyses scripts in this chapter can be found here: https://osf.io/tqp5a/. Simulation code (Experiment 3) is available here: https://github.com/masavujovic/RescorlaWagner

4.2 Experiment 3 (Computational Model)

4.2.1 Simulations

The same learning model was used as in Experiment 1 (Section 3.2.1). The design of the simulations was similar to Experiment 1, but with an important difference: the training input used in the present simulations incorporated a type-frequency manipulation, such that an uninformative feature occurred 75% of the time with one affix (high-type frequency or HF trials), and 25% of the time with the opposite affix (low type-frequency or LF trials). The discriminative features, on the other hand, were 100% consistent with the affix – these were clusters of features, such that, for example, all LF exemplars occurring with *ma* had three legs, a trunk-like "nose", and an appendage on the top of the "head" (see Figure 4.1). Modelling these features was less straightforward, as they would be entirely novel to the learner. Therefore, we followed Ramscar et al. (2010) and abstractly modelled each of these clusters as a single abstract feature (see Figure 4.1).

Experiment 2 of this thesis showed poor generalization on the basis of phonological features. Therefore, in the current simulations, we focus on the learning of semantic cues only (which were also the only kind of cues to include the critical type-frequency manipulation) and do not model phonological cues.

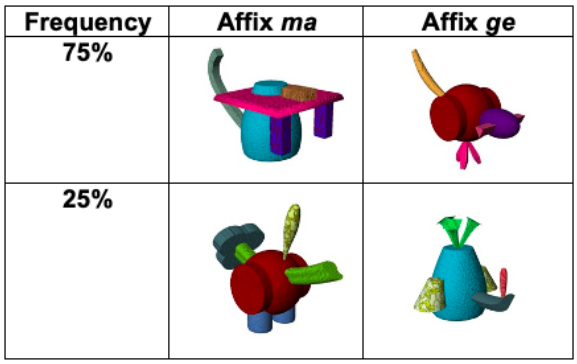
Finally, the evaluation metric used was the same as in Experiment 1 (Section 3.2.3), and simulations were trained for 7000 trials each.

4.2.2 Results

Figure 4.2 shows the associative strength (weights) of each feature for a randomly chosen affix ma in both conditions in a single model run.

Figure 4.3 shows the sums of raw weights between a set of features corresponding to a randomly chosen test item and the correct affix ma and the incorrect affix ge, for HF and LF items. As can be seen in Figure 4.3, the HF test item is associated more strongly with the correct than the incorrect affix in both conditions. The LF test item, on the other hand, is associated more strongly with the correct affix only in the suffix condition; in the prefix condition, the item is associated more strongly with the incorrect affix. When the raw sums of weights were normalized into a probability distribution, the probability of choosing the correct affix was greater in the suffix condition than in the prefix condition for both HF and LF items. This is shown in panels C and D of Figure 4.3, which plot the probability of

Memory and Language. The manuscript was co-authored with Elizabeth Wonnacott and Michael Ramscar, and a pre-print can be accessed here: https://osf.io/f2n9d/.



		Discriminating features				Non-discriminating features	
		Discrim1	Discrim2	Discrim3	Discrim4	Shape1	Shape2
Affix *ma*	75%	1	0	0	0	1	0
	25%	0	1	0	0	0	1
Affix *ge*	75%	0	0	1	0	0	1
	25%	0	0	0	1	1	0

Figure 4.1: The images used as the training set for modelling (top panel), and the same training set represented as a matrix on which the models were trained (bottom panel). (The visual stimuli were created by Michael Tarr's lab at Brown University.)

the network choosing the correct affix (for simplicity the probability of the incorrect affix was not plotted, but this corresponds to 1 – p(correct)).

4.2.3 Discussion

As in Experiment 1, discriminating cues for the opposite affix were assigned negative weights in the suffix condition, and zero weights in the prefix condition. However, unlike Experiment 1, in this experiment, it was only in the suffix condition that informative features for the given affix were discriminated from uninformative features. This is because in Experiment 1 there were no non-discriminating features which occurred more frequently with the opposite affix. In Experiment 3, on the other hand, a non-discriminating feature of LF items (body shape) actually occurred with the opposite affix 75% of the time. As in the prefix condition there were no trials on which a body shape occurred, but was not followed by the affix (simply because the affix always occurred first), the model was unable to "unlearn" body shape as a non-discriminating cue. In this condition, the associative strengths of features merely reflected their frequencies in the input, and therefore, *shape1* was associated more strongly with affix *ma* than *ge* (and *shape2* more strongly with affix *ge* than *ma* – the figures in the Results section only show the weights for one of the affixes for simplicity), despite the fact that *shape1* does not consistently predict the affix *ma.*

On the other hand, in the suffix condition, the discriminating features for the given affix (ma) were associated (1) equally strongly, despite the fact that one of them was three times more frequent than the other; and (2) more strongly than the uninformative features, despite the fact that one of these features (*shape1*) occurred with that affix more frequently than the LF discriminating feature. This is a clear demonstration of cue competition –

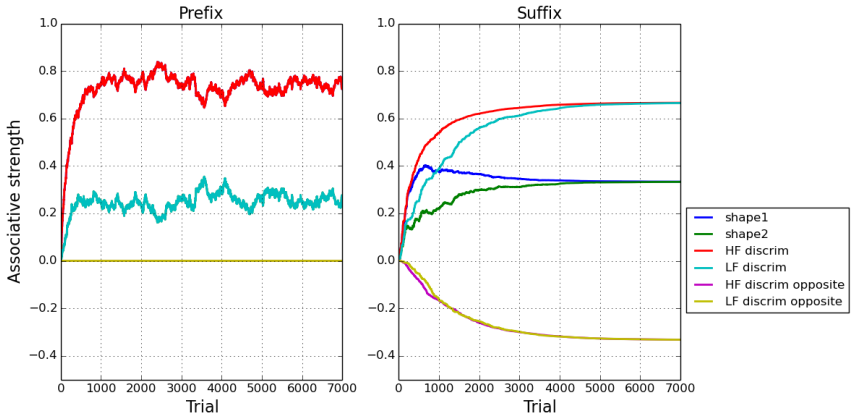


Figure 4.2: Associative strength between affix1 and shape1 (HF non-discriminating feature), shape2 (LF non-discriminating feature), discriminating feature1 (HF discriminating feature), discriminating feature2 (LF discriminating feature), and discriminating features 3 and 4 (HF and LF discriminating features for category 2, respectively) over time. In the prefix condition, non-discriminating features shape1 and 2, represented by dark blue and dark green lines, are learned equally well as the HF and LF discriminating features (red and bright blue lines respectively), and are thus over-plotted.

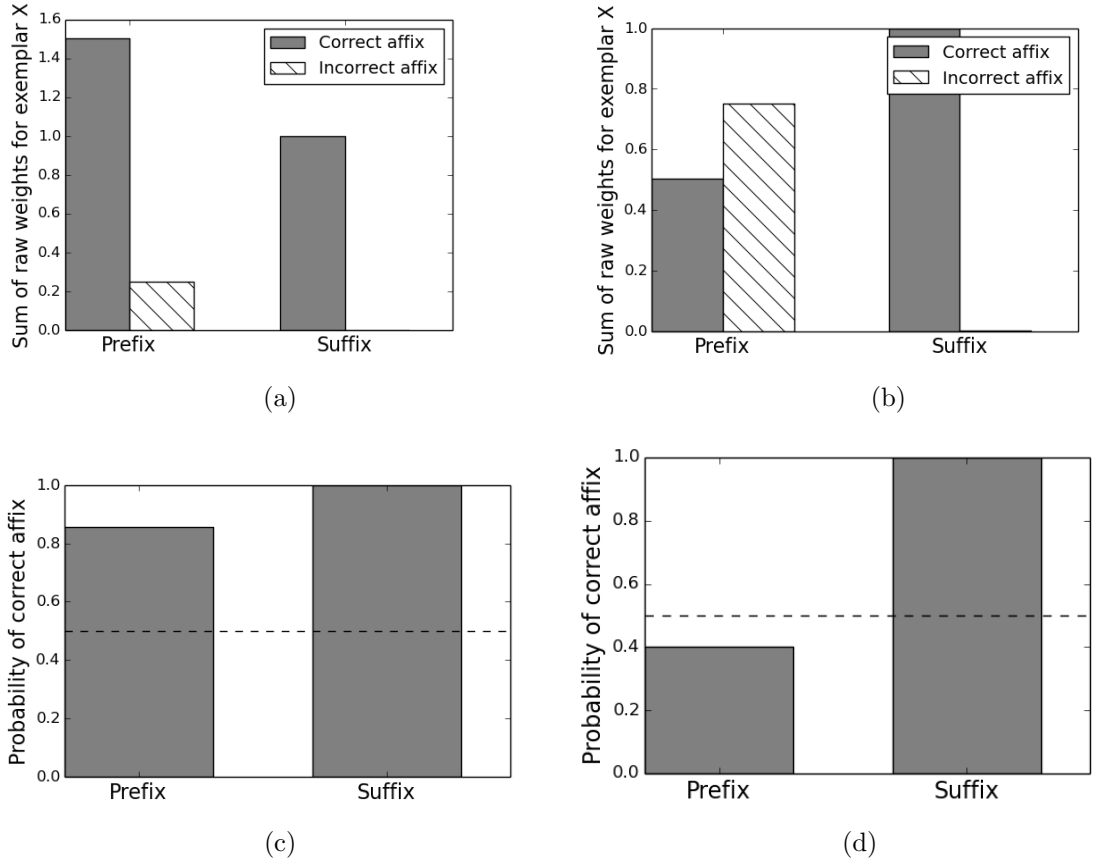


Figure 4.3: The sum of raw associative strengths for the correct affix (grey bar) and the incorrect affix (white striped bar) in each affix condition, for a HF test item (a) and a LF test item (b). The sums for correct affixes normalized into probability of choosing the correct affix out of two options (dashed-line is chance, 0.5) for the HF test item (c) and the LF test item (d).

for example, during the training phrase, those trials in which shape1 (uninformative cue) occurred with the opposite affix decreased the associative strength of that cue for the affix of interest, which in turn increased the associative strength of the other cues present. Similarly, those trials in which the target affix occurred without shape1 (LF trials) decreased the associative strength of shape1 to the benefit of the LF discriminating features. This can be observed in Figure 4.2, where approximately after 500 trials, shape1 starts to lose associative strength to the benefit of the LF discriminating feature (*discriminating feature 2*). Therefore, in the suffix network, *shape1* and *shape2* were associated equally strongly with both affixes (despite the fact that each occurred more frequently with one of the affixes), and less strongly than the discriminating features. In the prefix network, on the other hand, *shape1* and *shape2* were associated equally strongly as the discriminating features (despite the fact that they are non-discriminating), and more strongly with those affixes with which they occurred more frequently, even though they are not fully predictive of those affixes.

Consequences of different associative strengths in the two conditions became critical at test. While for HF items, the prefix network was more likely to choose the correct affix, for LF items, it was more likely to choose the incorrect affix. This occurred because LF items contained an uninformative feature (shape2 in this case) which was more strongly associated with the incorrect affix, due to their more frequent co-occurrence with that affix, and despite its low predictive value (that is, low informativeness). The suffix network, on the other hand, was equally accurate with HF and LF items – both were classified with near perfect accuracy, thus replicating a comparable finding from Ramscar et al. (2010). This pattern of results demonstrates that what is learned in the suffixing condition is the predictive value of different cues, whereas the prefix condition learns mere frequencies of co-occurrences in the input. This may, on the surface, be sufficient when informativeness and frequency are aligned (as is the case with HF items; although note there was a suffixing advantage with these items, too), but has drastic consequences when high-frequency cues are uninformative.

On the basis of the performance of these two simulations (and from Ramscar et al., 2010), we predict that, in the behavioural experiment, learners in the prefix condition will be unable to discriminate between cues based on their predictiveness, rather than frequency, compared to the learners in the suffix condition. This will be evidenced by an affix-by-type-frequency interaction in tests of generalization – specifically, participants in the suffix condition will generalize HF and LF items equally accurately, whereas participants in the prefix condition will be better with HF items compared to LF items. Following Study 1, we also predict a prefixing benefit in item-learning. These predictions are tested in Experiments 4a, 4b, 5, and 6.

4.3 Experiment 4a

4.3.1 Method

4.3.1.1 Participants

Eighty-four participants (42 per condition; prefix: *M age* = 36.46, *SD age* = 12.61, 24 female; suffix: *M age* = 34.38, *SD age* = 11.39; 26 female) were recruited through Prolific Academic and participated on-line. All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for participation.

4.3.1.2 Stimuli

Audio stimuli were the same as in Experiment 2, except that an additional 16 nouns were generated to accommodate a larger training set in this study. For visual stimuli, in each category, there were six high type-frequency (HF) items and two low type-frequency (LF) items (that is, 75% versus 25%). An additional four HF and four LF items per category (16 in total) were reserved for testing. See for a sample training set of pictures and labels.

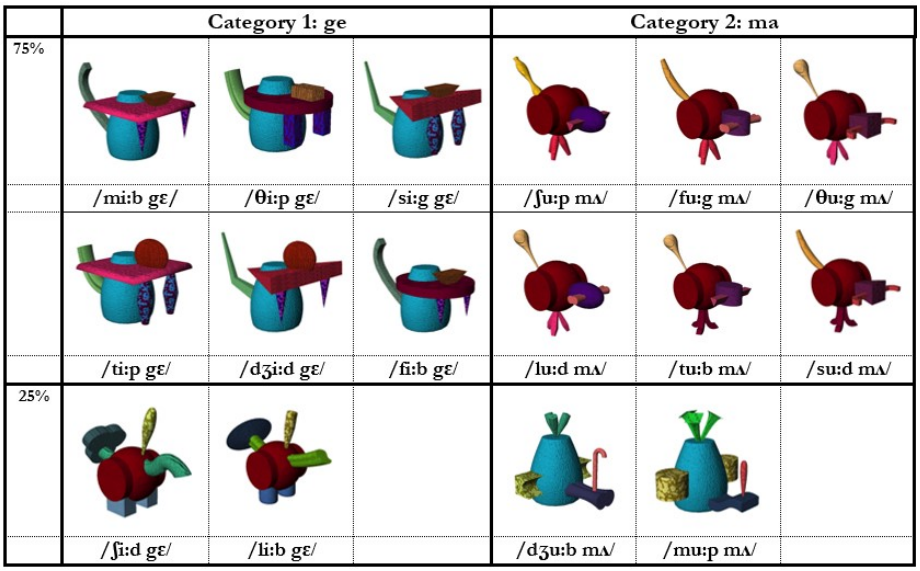


Figure 4.4: Experiment 4a: Sample training set (nouns were assigned to pictures and affix-categories randomly on a participant-by-participant basis.

4.3.1.3 Procedure

Procedure was the same as in Experiment 2b, except that there were more training items as well as more exposure per item. Further key changes concerned the timing of the training and testing trials, as well as the design of the item learning test. Below, I outline the details of the procedure which differed from that experiment – everything else can be assumed to be the same.

Training. The 15-minute training consisted of 4 blocks of 64 trials each (256 trials in total, 16 exposures per item). The fribbles were presented in pseudo-random order, such that low type-frequency items from one category were not followed or preceded by high type-frequency items from the opposite category, in order to minimize explicit learning. The timing of stimulus presentation was the same as in Experiment 2b, but the transition between individual trials was not self-paced, like in Experiments 2a and 2b, but automatic – a blank screen was displayed for 1000ms automatically as soon as the trial ended. Figure 4.5 shows a schematic representation.

Testing. The generalization tests were the same as in Experiment 2b, but the item-learning test was different. Unlike Experiment 2b, however, all the tests were timed, rather than self-paced – following Ramscar et al. (2010), if no response was received after 3500ms, a buzzer sound was played and a new trial would start after 1000ms. Prior to the testing, participants were instructed to respond as quickly as possible as they would have limited time to respond.

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 (noun + affix in the suffix condition, or affix + noun in the prefix condition) that was associated with the target item. The foil item was a different trained item from the same category. HF and LF items were tested against HF and LF foils, respectively. Half of the trained served as targets (chosen randomly), and the remaining half as foils, giving a total of eight trials – six HF and two LF trials. The change in design compared to Experiment 2b was made for two reasons – first, it was considered unfeasible to present all 16 trained items on-screen in a grid, as this would probably overwhelm participants; second, even if a sample of trained items was presented, it was not straightforward to determine whether 3500ms would be enough for participants to respond to a large number of stimuli, risking floor effects.

Semantics and Phonology Tests: Generalization. The same tests were used as in Experiment 2b (except, of course, that images of fribbles were used instead of the aliens). In each test, half of the testing items (eight) were randomly chosen as targets, and half (eight) as foils. Each item appeared only once, giving a total of eight trials (four HF and four LF). Unlike in Experiment 2b, the order of generalization tests 2 and 3 was counterbalanced [3].

Semantics and Phonology Test: Trained items. Same as Experiment 2b, except that there was a total of 16 trials (12 HF and four LF), as each trained item appeared

[3]Half of the participants did test 1, followed by tests 2 and 3, and the other half completed test 3 after test 1, and finished with test 2. This was done in order to look at the possibility that the different performance in the generalization tests we observed in Experiments 2a and 2b was somehow caused by the order of the tasks. Ideally, we would have wanted to counterbalance all three generalization test (e.g., test2 – test 1 – test 3), but it seemed unnatural for participants to switch between tests where they choose from two labels versus two pictures, and back, particularly as we did not wish to draw explicit attention the fact that the test tested different aspects of learning. The phonology generalization tests were also counterbalanced, as in Experiment 2b there was somewhat better performance in test 2, even though the two tests were essentially the same. However, the order of testing was not a significant predictor of accuracy in any of the models, and the order did not interact with any other factor. It did not improve the fit of the models, either, and therefore for simplicity, we will not discuss it further.

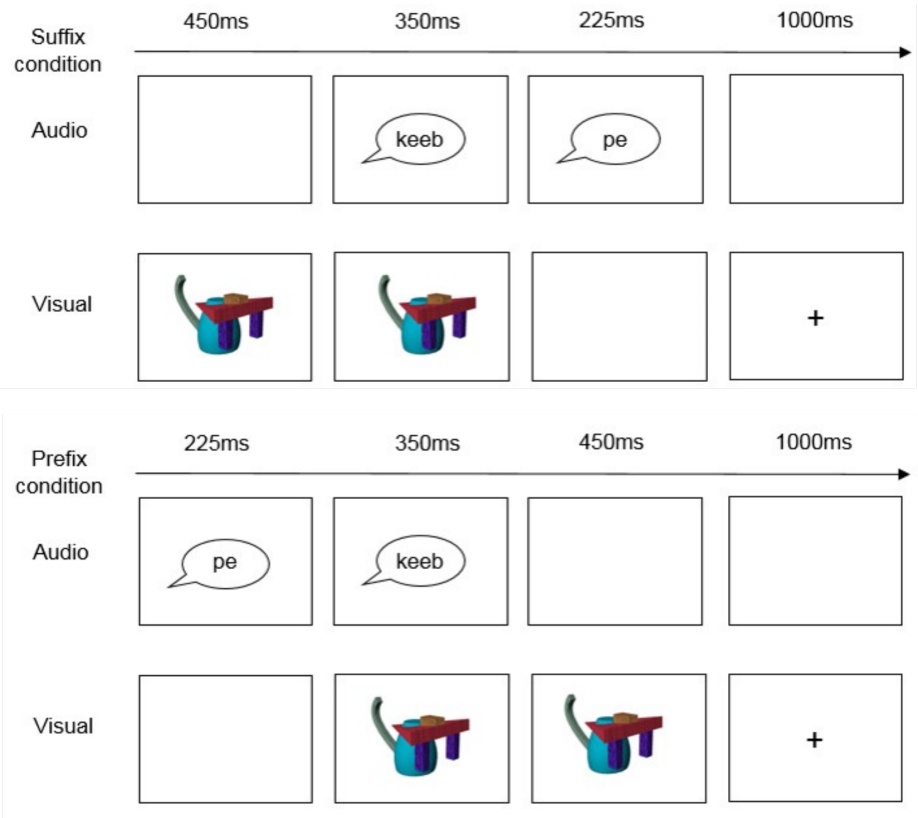


Figure 4.5: Schematic representation of the timing and stimulus display during a single training trial in the suffix condition (top panel) and in the prefix condition (bottom panel).

as the target once. (Note that, due to a technical error, the data from this test was not recorded. We include the description here since participants did take the test).

4.3.2 Results

4.3.2.1 Item learning test

Figure 4.6 shows the data. Participants appear to be at chance in every cell. For the Bayes factors, since no independent data were available to inform the H1, and because overall performance was low, there was no value from within the data to inform the H1 either. However, we computed Bayes Factors for every plausible value of H1 (as outlined in 2.2.1; Table 4.1). This revealed that we would have evidence for the null for plausible values up to and including a mean value as great as: 0.571 for Prefix HF (corresponding to mean performance of 63.9% compared to chance-level), 0.481 for Prefix LF (61.8%), 0.461 for Suffix HF (61.3%), and 0.351 for Suffix LF (58.7%). There are no plausible estimates of H1 for which the current data would provide evidence for that hypothesis (H1).

4.3.2.2 Semantics and Phonology Generalization Test

The data for all three tests are shown in Figure 4.7 and inferential statistics in Table 4.2. We predict an overall advantage for the suffix condition and an interaction with type-frequency. For each of the three tests, there was no overall suffix advantage, but the

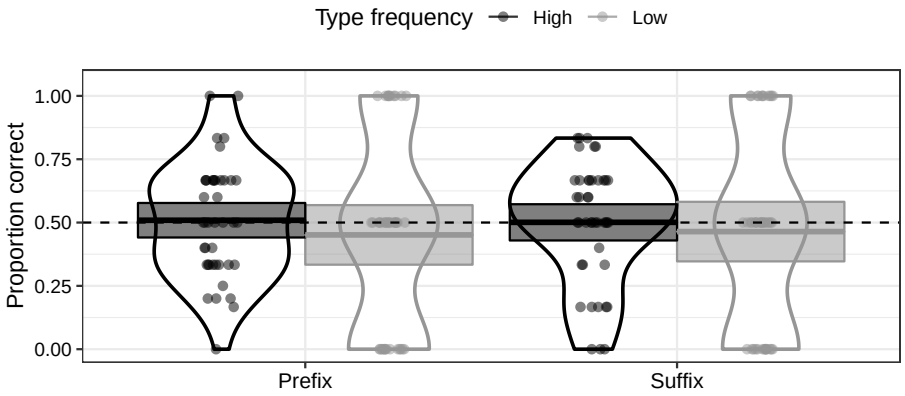


Figure 4.6: Experiment 4a: Proportion of correct responses on the Item learning test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

Table 4.1: Experiment 4a: Item Learning Test Statistics.

Hypothesis	**Contrast in the lmer**	**Mean difference**	**SE**	**Robustness region**	**p**
PrefixHF above chance	Intercept-PrefixHF	0.06	0.14	ambiguous: [0: 0.561], H0: [0.571 : ∞]	.715
PrefixLF above chance	Intercept-PrefixLF	-0.11	0.24	ambiguous: [0: 0.471], H0: [0.481 : ∞]	.636
SuffixHF above chance	Intercept-SuffixHF	0.02	0.14	ambiguous: [0: 0.451], H0: [0.461 : ∞]	.867
SuffixLF above chance	Intercept-SuffixLF	-0.23	0.23	ambiguous: [0: 0.341], H0: [0.351 : ∞]	.335

evidence was ambiguous (Test1: prefix: $M = 59\%$, $SD = 21\%$, suffix: $M = 61\%$, $SD = 18\%$; Test2: prefix: $M = 56\%$, $SD = 18\%$, suffix: $M = 57\%$, $SD = 19\%$; Test3: prefix: $M = 62\%$, $SD = 22\%$, suffix: $M = 62\%$, $SD = 19\%$).

For the interaction, there was evidence that the type frequency manipulation had a stronger effect in the prefix condition than in the suffix condition (indicated by an affix-by-type-frequency interaction) in Tests 1 and 3, whereas in Test 2 the evidence was ambiguous.

Breaking down by affix condition showed the same pattern in tests 1 and 3: evidence of an effect of type frequency in the prefix condition (as indicated by the fact that any plausible value of H1 would yield evidence for the H1), and evidence for no effect of type frequency in the suffix condition. In test 2, there was evidence for an effect of type frequency in the prefix condition for any value of H1 smaller than or equal to 1.241 (corresponding to 27.6% difference between HF and LF items) - given that this value is larger than the one observed in tests 1 and 3 (0.94 and 0.89, respectively), where there was evidence for the effect, we conclude that in test 2 we also see evidence for an effect of type frequency in the prefix condition; in the suffix condition, on the other hand, the evidence is ambiguous.

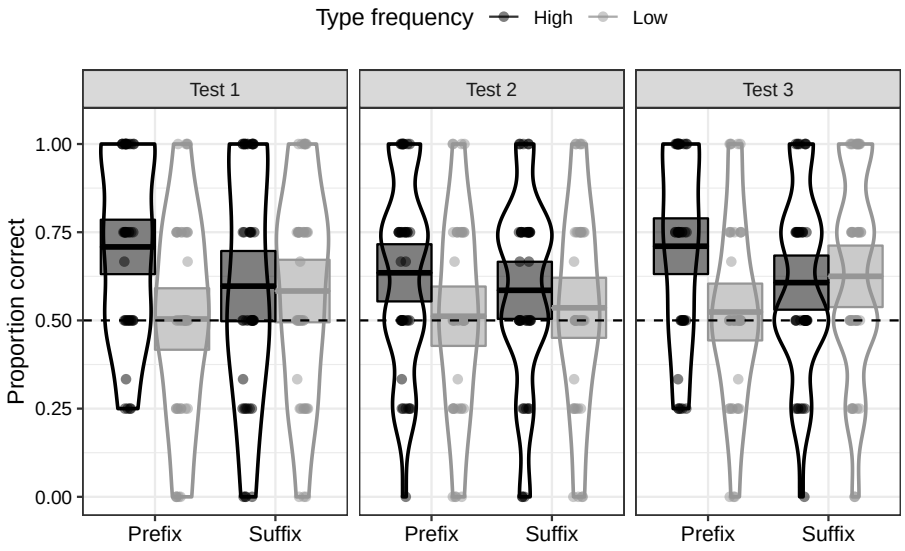


Figure 4.7: Experiment 4a: Performance on the Semantics and Phonology generalization tests 1 to 3. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

Table 4.2: Experiment 4a: Semantics and Phonology Generalization Test Statistics.

Hypothesis	**Contrast in lme**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Test 1							
Suffix better than Prefix	Main effect of affix	0.10	0.20	0.49[1]	0.58	[0 : 0.811]	.595
Stronger type-frequency effect in Prefix	Affix by type-fequency interaction	1.01	0.45	0.54[2]	5.32	[0.291 : 3.311]	.025
Breaking down by affix condition							
Type-frequency in Prefix	Type-fequency in Prefix	0.91	0.35	–	–	H1: [0: >4.591]	.009
Type-frequency in Suffix	Type-fequency in Suffix	-0.07	0.33	0.91[3]	0.29	[0.861: ∞]	.833
Cell-by-cell comparisons to chance							
PrefixHF above chance	Intercept-PrefixHF	1.06	0.24	–	–	H1: [0: >4.591]	<.001
PrefixLF above chance	Intercept-PrefixLF	0.02	0.19	1.06[4]	0.19	[0.551 : ∞]	.929
SuffixHF above chance	Intercept-SuffixHF	0.45	0.23	1.06[4]	2.68	[0 : 1.051]	.048
SuffixLF above chance	Intercept-SuffixLF	0.42	0.20	1.06[4]	3.25	[0 : 1.311]	.033
Test 2							

Hypothesis	Contrast in lme	Mean difference	SE	H1	B	Robustness region	p
Suffix better than Prefix	Main effect of affix	0.08	0.17	0.30[1]	0.70	[0 : 0.501]	.644
Stronger type-frequency effect in Prefix	Affix by type-fequency interaction	0.31	0.37	0.36[2]	1.22	[0 : 2.451]	.407
Breaking down by affix condition							
Type-frequency in Prefix	Type-fequency in Prefix	0.53	0.28	–	–	H1: [0 : 1.241], ambiguous: [1.251: >4.591]	.059
Type-frequency in Suffix	Type-fequency in Suffix	0.22	0.27	0.53[3]	0.90	[0 : 1.991]	.42
Cell-by-cell comparisons to chance							
PrefixHF above chance	Intercept-PrefixHF	0.59	0.18	–	–	H1: [0: >4.591]	.001
PrefixLF above chance	Intercept-PrefixLF	0.08	0.17	0.59[4]	0.41	[0 : 0.611]	.665
SuffixHF above chance	Intercept-SuffixHF	0.36	0.18	0.59[4]	3.50	[0 : 1.301]	.045
SuffixLF above chance	Intercept-SuffixLF	0.15	0.17	0.59[4]	0.65	[0 : 1.461]	.382
Test 3							
Suffix better than Prefix	Main effect of affix	0.08	0.20	0.52[1]	0.50	[0 : 0.821]	.702
Stronger type-frequency effect in Prefix	Affix by type-fequency interaction	0.93	0.34	0.36[2]	10.62	[0 : >4.591]	.005
Breaking down by affix condition							
Type-frequency in Prefix	Type-fequency in Prefix	0.89	0.26	–	–	H1: [0: >4.591]	.001
Type-frequency in Suffix	Type-fequency in Suffix	-0.04	0.25	0.89[3]	0.24	[0.621 : ∞]	.864
Cell-by-cell comparisons to chance							
PrefixHF above chance	Intercept-PrefixHF	0.95	0.19	–	–	H1: [0: >4.591]	<.001
PrefixLF above chance	Intercept-PrefixLF	0.12	0.18	0.95[4]	0.35	[0 : 1.101]	.509
SuffixHF above chance	Intercept-SuffixHF	0.45	0.18	0.95[4]	7.76	[0 : 2.061]	.012
SuffixLF above chance	Intercept-SuffixLF	0.55	0.19	0.95[4]	22.92	[0 : >4.591]	.004

[1]Intercept from the same lme (method B)

[2]Main effect of type-frequency from the same lme (method C)

[3]Type frequency effect in Prefix from the same lme (method A)

[4]Prefix HF intercept from the same lme (method A)

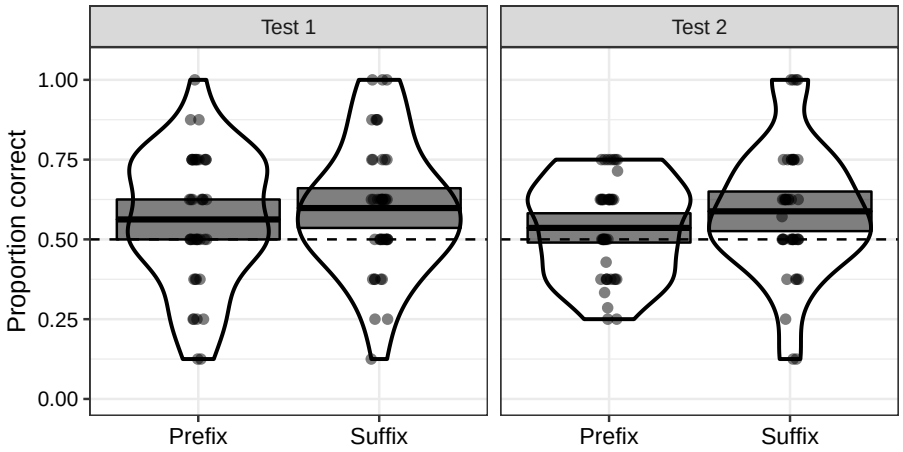


Figure 4.8: Experiment 4a: Performance on Phonology Generalization tests 1 (left) and 2 (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

4.3.2.3 Semantics and Phonology Test: Trained items

No data were collected for this test due to a technical error.

4.3.2.4 Phonology generalization

The data are shown in Figure 4.8 and the inferential statistics in Table 4.3. The same pattern of results was observed in both test types: no evidence of the predicted suffixing advantage in either test, although the evidence was ambiguous; evidence for above-chance performance in the suffix condition (Test 1: M = 56%, SD = 21%; Test 2: M = 54%, SD = 15%), and ambiguous evidence in the prefix condition (Test 1: M = 60%, SD = 21%, Test 2: M = 59%, SD = 21%).

4.3.2.5 Language Awareness Questionnaire (LAQ)

The results from the Language Awareness Questionnaire are summarized in Table 4.4. 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. In an exploratory analysis, we asked two questions about

Table 4.3: Experiment 4a: Phonology Generalization Test Statistics.

Hypothesis	Contrast in the lmer	Mean difference	SE	H1	B	Robustness region	p
Test 1							
Suffix better than Prefix	Main effect affix	0.16	0.19	0.343[1]	0.96	[0 : 1.541]	0.41
Prefix above chance	Intercept-Prefix	0.27	0.13	0.422[2]	3.472	[0.401 : 4.201]	0.05
Suffix above chance	Intercept-Suffix	0.42	0.14	–	–	H1: [0 : >4.591]	0.00
Test 2							
Suffix better than Prefix	Main effect affix	0.24	0.16	0.252[1]	2.114	[0 : 2.701]	0.14
Prefix above chance	Intercept-Prefix	0.13	0.11	0.371[2]	0.955	[0 : 1.171]	0.23
Suffix above chance	Intercept-Suffix	0.37	0.11	–	–	H1: [0 : >4.591]	0.00

[1]Intercept from the same lme (method B)
[2]Suffix intercept from the same lme (method A)

how participants' explicit knowledge might be related to their performance.

First, is participants' ability to generalize with HF items dependent on their being able to describe the high-saliency visual feature, body shape? To look at this, we removed all participants who mentioned this cue in the questionnaire (i.e. 11 prefix and 7 suffix participants) and tested whether the HF cells in the three generalization task remained above chance. In the prefix condition, HF-learning remained above chance in test 1 (β = 0.833, SE = 0.248, p = .001, evidence for H1 [0 : >4.591]) and test 3 (β = 0.77, SE = 0.216, p <.001 evidence for H1 [0 : >4.591]), although the evidence for learning was ambiguous in test 2 (β = 0.346, SE = 0.198, p = .08, [0 : >4.951]). In the suffix condition, HF-learning remained above chance in test 3 (β = 0.555, SE = 0.196, p = .005, BF = 21.231, RR [0 : >4.591]), and was ambiguous in test 1 (β = 0.191, SE = 0.222, p = .391; BF = 0.578, RR [0 : 1.521]) and test 2 (β = 0.214, SE = 0.185, p = .245, BF = 1.345, RR [0: 1.861]).

Second, is participants' generalization over phonological cues dependent on being able to describe the role of vowels in the stem? Here we considered the suffix condition only, since this was the only condition above chance. We removed the one participant who noted this feature, and found that performance remained above chance in each of the two phonology generalization tests (test 1: β = 0.418, SE = 0.135, p = .002, evidence for H1 [0 : >4.591]; test 2: β = 0.397, SE = 0.116, p = .001, evidence for H1 [0 : >4.591]).

4.3.3 Discussion

The original predictions in this thesis were that the suffixing condition is better at generalization than prefix condition. Having seen evidence for the opposite in Study 1, in Study 2 focus is shifted on a more fine-grained hypothesis – that a suffixing advantage is observed in the learning contexts in which cue competition is critical for generalization – that is, where generalization requires the learner to discriminate between frequent, but unpredictive cues,

Table 4.4: Experiment 4a: Language awareness questionnaire response summary

	Number of participants reporting awareness			
	Shape and/or colour of fribbles	Vowel in nouns	None	Other
Prefix	11 (26.19%)	3 (7.14%)	24 (57.14%)	4 (9.25%)
Suffix	7 (16.67%)	1 (2.38%)	30 (71.43%)	4 (9.52%)

and predictive cues. Therefore, we tested the hypothesis that the prefix condition will show greater type-frequency effects in generalization than the suffix condition. We still tested for an overall suffixing advantage (in line with the original prediction), as it was not possible to rule this out based on Study 1, given the high incidence of explicit learning in that study. With respect to item learning, better performance was predicted in the prefix condition compared to the suffix condition. The findings of Experiment 4a regarding generalization and item-learning are discussed in turn.

Starting with the tests of Semantics and Phonology generalization, there was no evidence for an overall benefit of suffixing over prefixing, although the Bayes Factors suggested ambiguous evidence, meaning that we cannot draw conclusions as to this prediction. However, turning to the prediction that participants are affected by the manipulation of type frequency in the prefix condition, there was clear evidence for this in two out of the three tests (tests 1 and 3), as evidenced by an affix-by-type-frequency interaction. Breaking down the interaction, there was evidence for an effect of type frequency in the prefix condition (as indicated by the fact that any plausible value of H1 would yield evidence for the H1), and evidence for no effect of type frequency in the suffix condition. This result is directly analogous to the finding in Ramscar et al. (2010) whereby participants showed better learning of the cues discriminating word meanings in the feature-label condition than the label-feature condition, specifically for low frequency items.

Interestingly, the interaction between type frequency and affix was not seen in test 2, although the evidence was ambiguous. While an ambiguous result cannot be interpreted, it is useful to reflect on why this test may not have provided evidence for the prediction. Recall that test 2 is the one generalization test where the association between discriminating semantic features and affixes was not a useful generalization cue. Instead, the test required knowledge of at least one of the phonology-based associations. However, there is no reason to expect a type frequency interaction for phonology-based cues, given that there is no type frequency manipulation for these cues. In retrospect, therefore, this particular test may not be appropriate for capturing this interaction. In addition, the performance on this test was numerically lower than on tests 1 and 3 in both conditions, which is similar to Experiments 2a and 2b, and suggests an overall difficulty with generalization on the basis of phonological cues (which is further supported by poor performance on the Phonology generalization tests, discussed below).

The interaction between affix and type frequency, observed in two out of three Semantic

Phonology generalization tests, confirms one of the central predictions of the discriminative learning approach. Note that this finding also demonstrates that participants' behaviour is specifically affected by semantic cues, since type frequency was not manipulated for phonology. For the Phonology Generalization test, the suffixing advantage was predicted, following the original prediction. However, although the suffixing condition showed numerically higher performance than the prefix condition, and there was only evidence of learning in the suffix condition, the evidence for learning in the prefix condition and for an a difference between conditions was ambiguous. This means that we cannot draw any conclusions for this hypothesis. As pointed out above, the performance in this test was generally low compared with the equivalent tests where semantic cues were available alongside the phonology (recall that Test 1 was identical to Semantics and Phonology Generalization Test 2 but without the semantics, and that Test 2 was identical to Semantics and Phonology Generalization Test 3), again indicating that semantic cues played a critical role in generalization in this experiment, which is also consistent with findings from Experiment 2.

Finally, one of the key changes made in the design of this study, compared to Experiment 2, was to make the training and testing procedures timed rather than self-paced, to discourage explicit learning. Indeed, results from the informal analysis of the LAQ suggest that, unlike in the earlier experiments, the generalization was not mainly driven by explicit awareness of the patterns in the language. In particular, no participant reported awareness of the role of the discriminating features which were key for generalization with the LF items, despite showing above-chance generalization of these in the suffix condition. Some participants did appear aware of the features relevant to generalization of HF items (i.e. the fribbles shape), as well as the relationship with vowels in the nouns. However, for at least some of the relevant tests, performance remained above chance with those participants excluded, suggesting that these "aware" participants did not drive the patterns of results reported above (for other tests, performance was "ambiguous", possibly due to decreased power as a consequence of removing participants, rather than there being evidence for chance-level performance). Thus it appears that participants were able to generalize even when they could later describe the relevant features over which this generalization occurred. It also appears that the modified paradigm in this experiment encouraged more implicit learning than was the case in Experiment 2, although the questionnaire may not 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 to verbalize them, due to the objects being novel, and/or unwilling. I return to this in the General Discussion of this Chapter (Section 4.7).

Turning to item learning, it is not possible to draw any conclusions about the potential benefits of prefixing, since there was no evidence that participants were above chance in any cell. Since there was no value to base H1 on here, the lowest values for which the null would hold were computed. These values were smaller than the observed mean performance reported elsewhere in this chapter where there was evidence for item learning (albeit using a different test) – while no formal inferences can be made based on this insight, we take it as

an indication of poor item learning in this experiment. This suggests that participants did not have strong learning of the associations between individual fribbles and the nouns; floor effects prevents us from being able to detect any potential effects of affix (i.e. the predicted prefix advantage). This possibility is addressed in a follow up experiment (Experiment 4b) which used a less demanding item learning test: rather than having participants choose between two trained items, participants were presented with one trained and one untrained item. Although this does not test participants' ability to differentiate items from training from each other, it does require them to recognize trained items, which we note they could not do if they were only tracking the features which are relevant for generalization. Experiment 4b also rectified the fact that, due to a programming error, no responses from the Trained items test were collected. Finally, Experiment 4b also serves as a pre-registered replication of the tests of generalization (https://osf.io/6gc4t). The training and generalization tests were identical except that, since the data are relatively noisy (as is common in learning experiments), we dropped some tests and increased the number of test items in others, with the goal of increasing the power without necessarily having to increase the number of participants. Specifically, tests 1 and 2 for Semantics and Phonology Generalization were excluded, retaining test 3 which is where the key effect (the interaction) was strongest, and which specifically tests relationship between the phonological and semantic cues and the affix (rather than testing the relationship between the noun phonology and the semantics) which is what is key to our theoretical prediction for a difference between suffixing and prefixing. Similarly, Phonology Generalization test 1 was excluded, retaining Phonology Generalization test 2, which is a phonology-only equivalent of Semantics and Phonology Generalization test 3.

4.4 Experiment 4b

4.4.1 Method

4.4.1.1 Participants

One hundred and twenty participants (60 per condition) were recruited through Prolific Academic (prefix: *M age* = 34.84, *SD age* = 9.66, 21 female; suffix: *M age* = 35.99, *SD age* = 10.13, 29 female). All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for participation.

4.4.1.2 Stimuli

Same as Experiment 4a.

4.4.1.3 Procedure

Same as Experiment 4a, except that:

In the **Item learning test**, foil items were novel items from the same category, as opposed to using the trained items from the same category. There were 16 trials (12 HF and four LF).

In the **Semantics and Phonology Generalization test**, we only used Test 3. Sixteen novel items (eight HF and eight LF) were used, giving a total of 16 trials.

In the **Phonology Generalization test**, we only used Test 2. The same 16 items from the Semantics and Phonology Generalization test were used, but without the corresponding pictures.

4.4.2 Results

Wherever possible, in analysing the data from this experiment we used values from Experiment 4a to inform the H1. This was not possible for Item Learning test, as there was no evidence for learning in Experiment 4a. Instead, values from Experiment 5a were used to inform the H1, because this was the first experiment where there was evidence for item learning, and where the same test was used as in Experiment 4b [4]. The set of values used to inform H1 were pre-specified in the pre-registered plan.

In the pre-registered analysis plan, we indicated starting with 20 participants in each condition, analysing the data, and continuing with adding 10 participants per condition before inspecting the data at each step, until we have found substantial evidence for the interaction (Bayes Factor greater than 3) in test 3, which is the key aspect which we hoped to replicate. There was substantial evidence for this with 30 participants per condition. However, evidence for item-learning was ambiguous for HF items, and since our resources allowed us to collect more data, we doubled the sample size, resulting in 60 per condition. Therefore, we inspected the data at: 20 participants per condition, 30, and finally 60. Recall that unlike p-values, as a measure of the strength of evidence, Bayes Factors are interpretable after adding more data (Dienes, 2016; J. N. Rouder, 2014). I nevertheless continue to also report p-values since these are more familiar to the reader, but note that they should be interpreted with caution since they are not adjusted for optional stopping (and are not interpreted in the thesis).

4.4.2.1 Item learning

The data can be viewed in Figure 4.9 and the statistics in Table 4.5. We predicted better performance in the prefix condition than the suffix condition. This was not the case, with substantial evidence for the null (prefix: M = 56.18%, SD = 14.26%; suffix: M = 56.76%, SD = 14.82%). Cell-by-cell comparisons to chance showed the same pattern in both affix conditions: evidence for no learning of HF items and evidence for learning of LF items.

Although not part of the original predictions, we therefore tested the hypothesis that overall, LF items were better identified than HF items and found evidence that this was the case (HF: M = 51.57%, SD = 19.09%; LF: M = 71.58%, SD = 19.09%).

[4] note that Experiment 5a was conducted prior to 4b but here were report them in this order for clarity of exposition

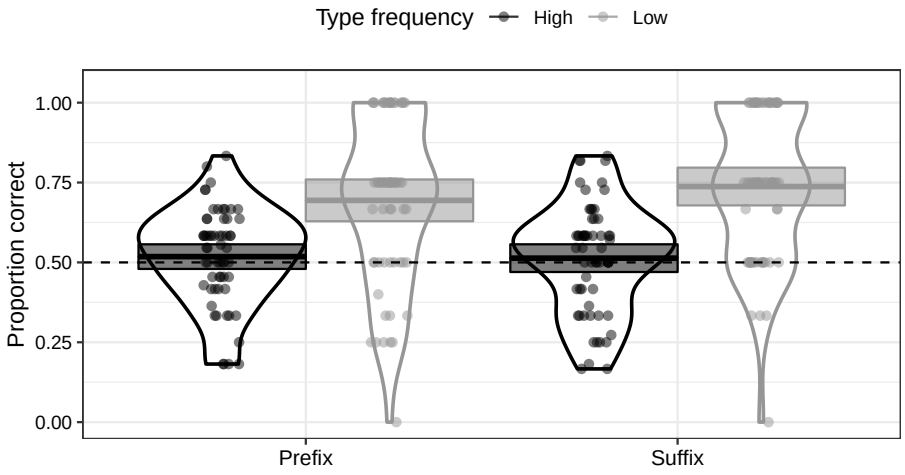


Figure 4.9: Experiment 4b: Proportion of correct responses on the Item Learning test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

4.4.2.2 Semantics and Phonology test: Generalization

The data are shown in Figure 4.10 and inferential statistics are in Table 4.6. As in Experiment 4a, there was no evidence for overall higher performance in the suffix condition than in the prefix condition, however unlike in Experiment 4a, here there was substantial evidence for the null (suffix: M = 61.1%, SD = 19.21%; prefix: M = 56.61%, SD = 16.8%), and this was the case also for combined data.

Turning to the interaction we again saw that the type frequency manipulation had a stronger effect in the prefix condition than in the suffix condition, with evidence for the interaction between affix and type frequency. Breaking down by affix condition showed that HF items were learned better than LF items in the prefix condition (as in Experiment 4a), whereas in the suffix condition, the evidence comparing HF and LF items was ambiguous (unlike in Experiment 4a, where we had evidence for the null) and this was also the case even when combining data across the experiments.

4.4.2.3 Semantics and Phonology Trained items test

The data are shown in Figure 4.11 and inferential statistics in Table 4.7. (Note that since the figure shows proportions, the differing number of test trials (see section 2.1.3) for HF and LF means that there are fewer possible scores per participant with the latter test type). To see whether participants relied on the same cues in this test as in the generalization task, we looked for the same affix-by-type-frequency interaction, but did not find it, with ambiguous evidence. We then looked at whether participants overall performed better in this test than in the generalization test, which we would predict if they used item-level cues as well as the more general ones. This was found to be the case (trained: M = 63.79%, SD = 20.97%; new: M = 58.86%, SD = 18.11%).

Table 4.5: Experiment 4b: Item Learning Test Statistics.

Hypothesis	Contrast in the lme	Mean difference	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect of affix	-0.05	0.11	0.27[1]	0.28	[0.221 : ∞]	.663
LF better than HF	Main effect of TF	0.91	0.12	0.56[2]	6.96×10^{10}	[0: >4.591]	<.001
Cell-by-cell comparisons to chance							
Prefix HF above chance	Prefix HF Intercept	0.08	0.08	0.90[3]	0.24	[0.641 : ∞]	.347
Prefix LF above chance	Prefix LF Intercept	0.84	0.15	0.90[3]	6.1×10^{5}	[0: >4.591]	<.001
Suffix HF above chance	Suffix HF Intercept	0.06	0.08	0.90[3]	0.18	[0.471 : ∞]	.494
Suffix LF above chance	Suffix LF Intercept	1.10	0.16	0.90[3]	1.45×10^{9}	[0: >4.591]	<.001

[1]Intercept from the same lme (method B)
[2]Same effect from Experiment 5a (method A)
[3]Prefix LF intercept from Experiment 5a (method A)

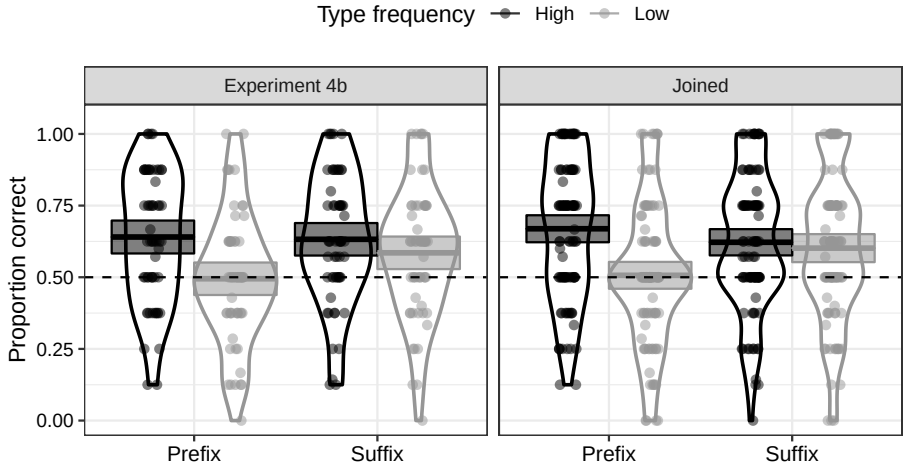


Figure 4.10: Proportion of correct responses on the Semantics and Phonology generalization test in Experiment 4b (left) and with combined data (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line is chance-level performance

Table 4.6: Experiment 4b: Semantics and Phonology Generalization Test Statistics.

Hypothesis	Contrast in lme	Data	Mean difference	SE	H1	B	Robustness region	p
Suffix better than Prefix	Main effect of affix	4b	-0.19	0.15	0.52[1]	0.136	[0 : > 4.591]	.21
		Joined	-0.12	0.12	0.449[2]	0.147	[0 : > 4.591]	.345
Stronger type frequency effect in Prefix	Affix by type-frequency interaction	4b	0.45	0.22	0.932[3]	3.4	[0 : 1.091]	.039
		Joined	0.59	0.18	0.42[4]	60.71	[0: >4.591]	.001
Breaking down by affix condition								
Type frequency Prefix	TF effect in Prefix	4b	0.63	0.16[5]	0.894	834	[0: >4.591]	<.001
		Joined	0.69	0.13	–	–	H1: [0: >4.591]	<.001
Type frequency Suffix	TF effect in Suffix	4b	0.16	0.17	0.894[5]	0.468	[0 : 1.261]	.345
		Joined	0.10	0.14	0.692[6]	0.377	[0 : 0.761]	.487
Cell-by-cell comparisons to chance								
Prefix HF above chance	Prefix HF Intercept	4b	0.65	0.14	0.954[7]	9439	[0 : >4.591]	<.001
		Joined	0.75	0.12	–	–	H1: [0: >4.591]	<.001
Prefix LF above chance	Prefix LF Intercept	4b	-0.02	0.12	0.954[7]	0.119	[0.321 : ∞]	.892
		Joined	0.03	0.10	0.747[8]	0.179	[0.291 : ∞]	.747
Suffix HF above chance	Suffix HF Intercept	4b	0.61	0.14	0.954[7]	2947	[0: >4.591]	<.001
		Joined	0.57	0.11	0.747[8]	85942	[0: >4.591]	<.001
Suffix LF above chance	Suffix LF Intercept	4b	0.40	0.13	0.954[7]	32.281	[0: >4.591]	.002
		Joined	0.45	0.11	0.747[8]	1648	[0: >4.591]	<.001

[1]Intercept from lme with Experiment 4a data (method A)
[2]Intercept from same lme (method B)
[3]Affix by type-frequency interaction from lme with Experiment 4a (method A)
[4]Main effect of type-frequency from same lme (method C)
[5]Effect of type frequency in Prefix from lme with Experiment 4a data (method A)
[6]Effect of type frequency in Prefix from same lme (method A)
[7]Prefix HF intercept from lme with Experiment 4a data (method A)
[7]Prefix HF intercept from same lme (method B)

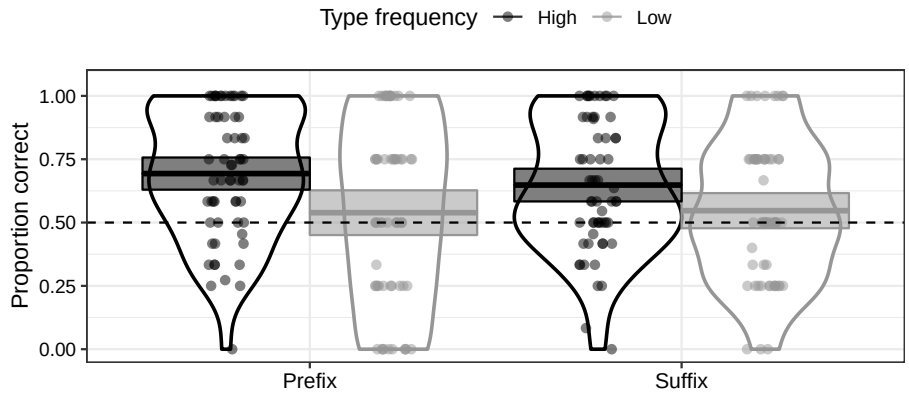


Figure 4.11: Proportion of correct responses on the Semantics and Phonology Trained items test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance

Table 4.7: Experiment 4a: Semantics and Phonology Trained Items Test Statistics.

Hypothesis	Contrast in lme	Mean difference	SE	H1	B	Robustness region	p
Suffix better than Prefix	Prefix better than Suffix	0.19	0.23	0.52[1]	0.85	[0.351 :]	0.395
Stronger TF effect in Prefix	Affix by TF interaction	0.31	0.37	0.93[2]	0.79	[0.591 : ∞]	0.405
Cell-by-cell comparisons to chance							
Prefix HF above chance	Prefix HF vs chance	1.12	0.20	0.95[?]	913828	[0: >4.591]	<.001
Prefix LF above chance	Prefix LF vs chance	0.20	0.19	0.95[?]	0.55	[0 : 1.611]	0.308
Suffix HF above chance	Suffix HF vs chance	0.84	0.20	0.95[?]	2119	[0 : >4.591]	<.001
Suffix LF above chance	Suffix LF vs chance	0.23	0.19	0.95[?]	0.72	[0: 2.151]	0.218
Comparison with generalization test							
Trained items better than generalization	Item novelty effect	-0.23	0.09	0.11[4]	0.19	0 : 1.841]	.008

[1]Intercept from generalization in Experiment 4a (method A)
[2]Same effect in generalization in Experiment 4a (method A)
[3]Prefix HF generalization Intercept from Experiment 4a (method A)
[4]Same effect from Experiment 5a (method A)

4.4.2.4 Phonology Generalization Test

The data are shown in Figure 4.12 and the inferential statistics are in Table 4.8. The results for this test were identical to those of Experiment 1a: the evidence for the predicted better learning in the suffix condition was ambiguous (suffix: M = 54.79%, SD = 13.59%; prefix: M = 53.53%, SD = 14.57%). However, there was strong evidence for above-chance performance in the suffix condition, whereas the evidence for the prefix condition was ambiguous. This was the case for BFs based both on Experiment 1b alone, and for the combined data from Experiment 1a and Experiment 1b.

4.4.2.5 Language Awareness Questionnaire

The same coding process was used as described in Experiment 2a. Results are summarized in Table 4.9. Here, although we again had no participants who described the discriminating features which were relevant to correct generalization of LF items, there were four participants in each condition who reported noticing that there were "exception" items (e.g., "X aliens went with Y, but I noticed exceptions"; "All X aliens went with Y, except some which went with the opposite one"). This time, we therefore asked three exploratory questions about how participants' explicit knowledge might be related to their performance.

First, is participants' ability to generalize with low frequency nouns in the suffix condition, driven by those participants who noticed the exceptions? To look at this, we removed the four participants who mentioned "exceptions" in questionnaire and check whether the suffix condition remained above chance with LF items, which we found to be the case (β =

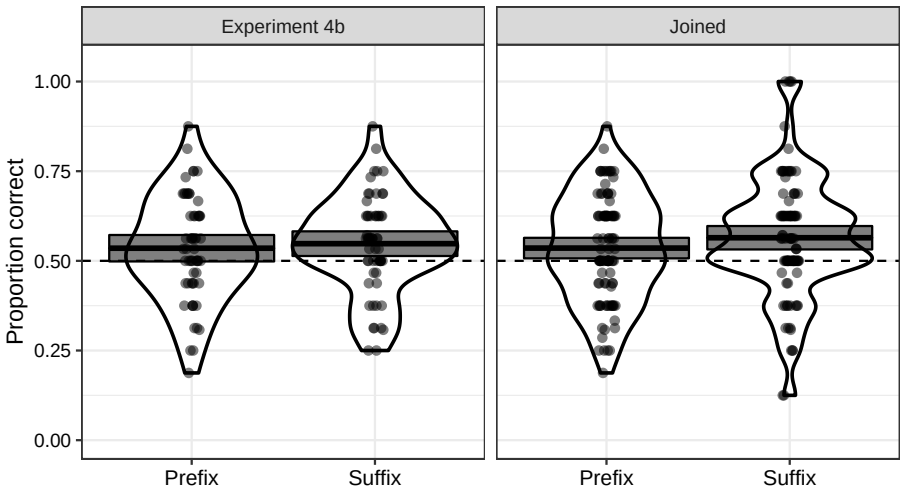


Figure 4.12: Experiment 4b: Proportion of correct responses on the Phonology generalization test in Experiment 4b (left) and with combined data (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance

Table 4.8: Experiment 4b: Phonology Generalization Test Statistics.

Hypothesis	**Contrast in lme**	**Data**	**Mean diff.**	**SE**	**H1**	**B**	**Robustness region**	**p**
Suffix better than Prefix	Main effect affix	4b	0.05	0.10	0.252	0.576	[0 : 0.471]	.609
		Joined	0.09	0.09	0.197	1.114	[0 : 0.801]	.277
Breaking down by affix condition								
Prefix above chance	Prefix Intercept	4b	0.15	0.07	0.371	2.802	[0.401 : 3.401]	.039
		Joined	0.15	0.06	0.244	8.275	[0 : 0.811]	.014
Suffix above chance	Suffix Intercept	4b	0.20	0.07	0.371	15.344	[0: 2.211]	.006
		Joined	0.24	0.06	–	–	[0 : >4.591]	<.001

[1]Intercept from lme with Experiment 4a data (method A)
[2]Intercept from same lme (method B)
[3]Suffix intercept from lme with Experiment 4a (method A)
[4]Suffix intercept from same lme (method B)

Table 4.9: Experiment 4b: Language awareness questionnaire response summary

	Number of participants reporting awareness				
	Noticed exceptions	Shape and/or colour of fribbles	Vowel in 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 (1.67%)	7 (11.67%)	40 (66.67%)

0.345, SE = 0.129, p = .007, BF = 8.783, RR [0 : 3.041]). As a further check, we checked whether the key type frequency interaction still holds after removing participants in both conditions who reported noticing exceptions, and found that this was the case (β = 0.446, SE = 0.218, p = .041, BF = 3.283, RR [0 : 1.031]).

Second, is participants' ability to generalize with high frequency nouns dependent on their being able to describe the feature on which this generalization depends? This was again found not to be the case for both conditions, with performance remaining above chance once those participants were removed (suffix: β = 0.578, SE = 0.152, p <.001, BF = 358.151, RR [0 : >.951], prefix: β = 0.673, SE = 0.15, <.001, BF = 5664, RR [0 : >4.591]).

Third, is participants' performance generalization over phonological cues dependent on being able to describe the role of vowels in the stem? Again we look at this only with suffix participants, since only this group were above chance - they continue to perform above chance even once the participant who explicitly noticed the vowel cue has been removed (β = 0.19, SE = 0.073, p = .01, BF = 10.021, RR [0 : 1.411]).

4.4.3 Discussion

In Experiment 4b, the key finding from Experiment 4a was replicated, in which the Semantics and Phonology Generalization test showed that cue competition plays a critical role in linguistic generalization: the prefix condition showed a stronger effect of type frequency than the suffix condition, due to, we believe, greater prediction error via cue competition in that condition. Combining data across the experiments showed even stronger evidence for this effect. Moreover, LF items were not learned in the prefix condition (with substantial evidence for the null) whereas they were learned in the suffix condition. Contrary to the original predictions, however, there was again no evidence for overall suffixing advantage in generalization, but this time there was substantial evidence for the null, which also held when we combined the data sets. This finding is contrary to previous work which found a benefit of suffixing without any frequency manipulation (e.g., Hupp et al., 2009; St Clair et al., 2009), and to Study 1 which found a prefixing advantage -– taken together, these patterns of results suggest that the effects of affix order may be more nuanced than has been considered in the literature. I return to this point in the General Discussion of this

chapter (Section 4.7).

As in Experiment 4a, the post-experiment questionnaire suggested little awareness of the key features necessary for generalization in any of the conditions. Again, even when participants who explicitly reported noticing a feature relevant for generalization were removed, the performance of the group remains above chance. One difference in this study compared with 4a was that some participants noticed that there were "exception" items (four participants in each condition), but again removing them did not change the performance of participants in the suffix condition (i.e. the group who showed learning of these items in our main analyses), who remained above-chance. Additionally, there was still substantial evidence for the stronger effect of type frequency in the prefix condition than in the suffix condition (interaction) after removing participants who reported awareness of the "exception" items from both conditions.

Turning to item learning, we predicted better learning in the prefix condition compared to the suffix condition. Recall that in Experiment 4a there was a floor effect, as indicated by evidence for chance-level performance. In Experiment 4b, a different item-learning test showed evidence for learning in both conditions, but, critically, contrary to the prediction, there was no prefixing benefit, with substantial evidence for the null. However, in both affix conditions, LF items were learned better than HF items, with cell-by-cell comparisons to chance showing substantial evidence for the null for HF items in both conditions, a finding which we had not predicted. Note that this difference cannot be understood in terms of the frequency of individual fribbles, since high and low type frequency fribbles were each seen the same number of times in training. What this finding instead suggests is that participants are only able to recognize fribbles when they come from the category with the fewer number of members. A likely explanation is that discriminating between individual items from the same category is perceptually difficult, and that this was harder for HF-type items than LF-items, because there were more items from the same category to be discriminated (meaning greater uncertainty levels). I return to this in the General Discussion of the Chapter (Section 4.7).

In addition to the Item Learning test, in this experiment, data were collected in a second test of trained items. This test had the same structure as the Semantics and Phonology Generalization test, but using fribbles which had occurred during training. We expected that participants should perform overall better on this test than on the generalization test, due to the fact that, in addition to the features relevant in generalization, they can also use idiosyncratic features of the nouns and their referents when determining affix usage. There was substantial evidence for this in the data. We also tested for an interaction between affix and type frequency, which was absent in this test (with ambiguous evidence) and the type frequency manipulation was evident for both groups, specifically, performance on trained LF-items in the suffix condition appeared to be worse here compared with the equivalent novel items in the generalization test. This finding is difficult to interpret in the context of the theory. If anything, there are more features in the trained items that could help determine the correct affix (i.e. the idiosyncratic features of the individual fribbles, which are uniquely associated with that affix). One thing to note is that since this test occurred

after the generalization tests, exposure to novel items in that test may have affected how participants responded to the low-frequency informative features (the trained items test was presented after the generalization test to all participants), perhaps due to overlearning of the more salient features (body shape and colour) which could lead to worse performance on LF items, particularly in the suffix condition. Finally, note that while this test was similar to the generalization test, it was not identical. In the generalization test, different nouns were used for the target and the foil items (e.g., *foop ge* vs *kood ma*), whereas in the trained items test, using two different trained nouns was not possible as the purpose was to test participants' learning of the item-affix mappings, rather than the noun-picture mappings. While there is no theoretical reason for this difference between the tests alone to cause poorer performance on the LF items in the suffix condition, the differences in the set-up may mean that the performance on the two tests should be compared with caution.

To summarize, Experiment 4b provided further evidence for the critical role of cue competition in linguistic generalization. Across Experiments 4a and 4b, there was evidence that low-frequency discriminating features were only learned in the suffix condition, and not in the prefix condition. With respect to item-learning, while it is encouraging that the new test used in Experiment 4b elicited item-learning, this was only true for LF-items, and, contrary to prediction, we did not see the predicted benefit of prefixing over suffixing. One possibility is that stronger item-learning is required to observe an effect of affix. In Experiment 5, this possibility is addressed by reducing the vocabulary size by 50%. Rather than training participants on six HF and two LF items per category, in Experiment 5, participants will be trained on three HF and one LF item per category (thus preserving the 3:1 ratio of items). Two versions of this experiment were run (Experiment 5a and 5b), with identical training, but with tests matching those in Experiments 4a and 4b, respectively, except for the item learning test, where we used the version from Experiment 4b in both 5a and 5b (recall that the item learning test used in Experiment 4a showed a floor effect).

4.5 Experiment 5

4.5.1 Method

4.5.1.1 Participants

Two hundred and twenty seven participants were recruited through Prolific Academic. Of these, 110 (56 in prefix condition; M_{age} = 37.96, SD_{age} = 11, 37 female; 54 in the suffix condition; M_{age} = 37.68, SD_{age} = 12.5, 25 female) were recruited for Experiment 5a, and 120 (60 per condition; prefix: M_{age} = 32.38, SD_{age} = 9.86, 32 female; suffix: M_{age} = 34.97, SD_{age} = 10.56, 30 female) were recruited for Experiment 5b. 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 participating.

For Experiment 5a, sample size was determined via a power simulation based on the interaction we observed in Experiment 4a (see https://rpubs.com/MasaVujovic/discrimlearning2).

Data were inspected only once, after the full sample of 110 participants had been collected. For Experiment 5b, we used an optional stopping procedure as outlined for Experiment 4b, and inspected the data at: 20 participants per condition, 30, and finally 60. As in Experiment 4b, there was d evidence for the key interaction with 30 participants, but resources allowed further testing, thus we doubled the sample.

4.5.1.2 Stimuli

These were drawn from the set as in Experiment 4, however, there were fewer training audio and visual items[5]. In the current experiment, there were 8 training items in total: four per category, three HF, and one LF item per category. See Figure 4.13 for a sample training set of pictures and labels.

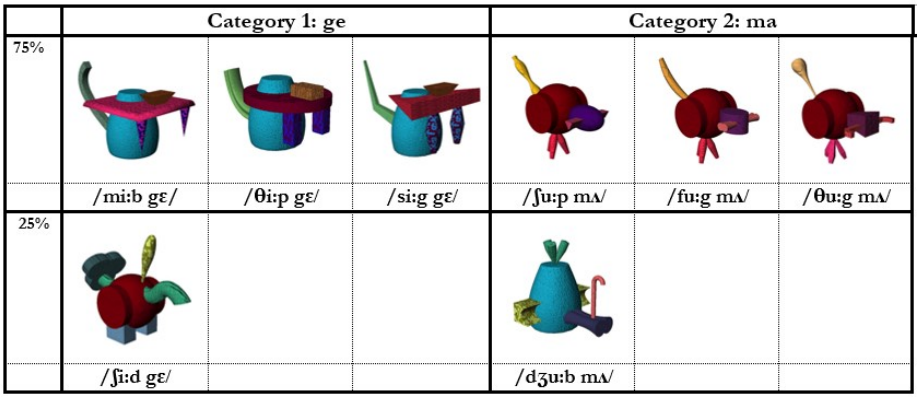


Figure 4.13: Experiment 5: Sample training set. Note that within each category, nouns were assigned to pictures randomly on a participant-by-participant basis.

4.5.1.3 Procedure

This was identical to Experiments 4a and 4b, except that in training, there were 32 exposures per item (same total number of training trials as in Experiments 4a and 4b, respectively), and that in both versions of the experiment, we used the Item learning test from Experiment 4b.

4.5.2 Results

4.5.2.1 Item learning test

The data are shown in Figure 4.14 and the statistics in Table 4.10. In both versions of the experiment, and pooling across the two, evidence for an overall prefix advantage was ambiguous (Experiment 5a: prefix: M = 56.34%, SD = 19.29%; suffix: M = 56.73%, SD = 22.77%; Experiment 5b: prefix: M = 65.32%, SD = 22.01%; suffix: M = 66.79%, SD = 18.62%). In both versions, and pooling across the two, there was evidence that LF items were recognised more accurately than HF items (Experiment 5a: HF: M = 53.41%, SD =

[5]Given that version 5a was a replication of Experiment 4a (and version 4b was a replication of Experiment 5b), version 4a also included a control category. As in Experiment 4a, there was no relationship between participants' performance on the control category trials and their performance on other tasks.

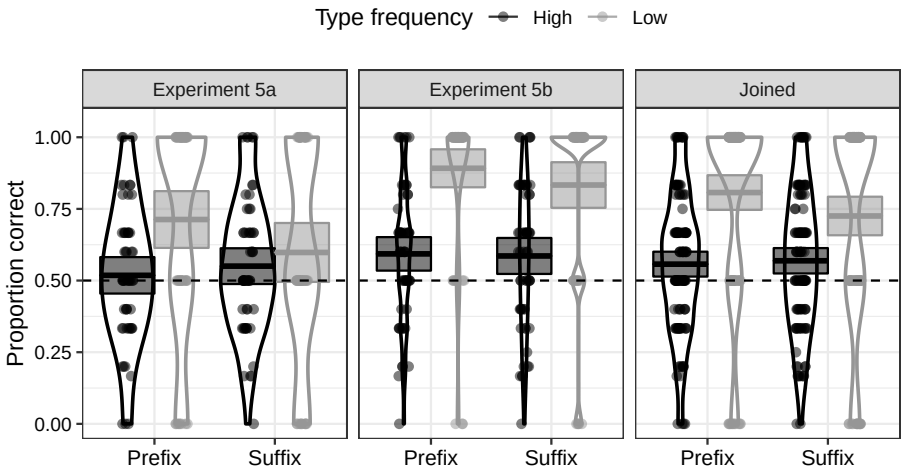


Figure 4.14: Experiment 5: Proportion of correct responses on the Item-learning test in Experiments 5a (left), 5b (middle) and with combined data (left). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance

29.38%; LF: M = 65.71%, SD = 29.89%; Experiment 5b: HF: M = 58.94%, SD = 23.71%; LF: M = 86.25%, SD = 23.71%).

As expected, item learning was better in Experiment 5 compared to Experiment 4b (recall that Experiment 4a used a different item learning test). Robustness regions suggested that the data would provide evidence for any predicted difference up to size 1.931 in log odds space – equivalent to predicting an increase from 56.5% (average performance in Experiment 4b) up to any value smaller than 90% in Experiment 5.

4.5.2.2 Semantics and Phonology Generalization test

The data are shown in Figure 4.15 and the inferential statistics in Table 4.11.

In Experiment 5a, there was evidence for no overall advantage of suffixing compared to prefixing (prefix: M = 57.26%, SD = 20.34%; suffix: M = 61.13%, SD = 18.38%). However, this was ambiguous in Experiment 5b and the combined evidence was ambiguous (prefix: M = 63.6%, SD = 17.06%; suffix: M = 58.8 SD = 15.92%). In Experiment 5a, the evidence that the type-frequency manipulation had a stronger effect in the prefix than in the suffix condition was ambiguous, however in Experiment 5b the evidence was substantial and the evidence from the combined experiments was even stronger.

Breaking down by affix condition, there was strong evidence for an effect of type frequency in the prefix condition in both Experiments 5a and 5b (and in the combined data), but for the suffix condition the evidence was ambiguous in Experiment 5a and there was evidence for no effect of type frequency in Experiment 5b and evidence for no effect of type frequency in the combined data.

In Experiment 5a, cell-by-cell comparisons to chance showed evidence for learning in every cell except Prefix LF where there was evidence for no learning (as in Experiment 4a

Table 4.10: Experiment 5: Item Learning Test Statistics.

Hypothesis	Contrast in lme	Data	Mean difference	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect affix	5a	0.12	0.16	0.274	0.927	[0 : 0.971]	.46
		5b	0.16	0.23	1.048	0.397	[0 : 1.261]	.498
		Joined	0.13	0.14	0.595	0.591	[0 : 1.081]	.322
LF better than HF	Main effect type frequency	5a	0.56	0.20	0.274	14.76	[0: >4.591]	.005
		5b	2.62	0.94	1.048	11.104	[0.421: >4.591]	.005
		Joined	1.30	0.23	0.595	5.1×10^{5}	[0: >4.591]	<.001
Cell-by-cell comparisons to chance								
Prefix HF above chance	PrefixHF Intercept	5a	0.12	0.12	3.369	0.094	[0.961 : ∞]	.333
		5b	0.41	0.13	0.974	42.423	[0: >4.591]	.001
		Joined	0.27	0.09	1.825	10.626	[0: >4.591]	.002
Prefix LF above chance	PrefixLF Intercept	5a	0.97	0.27	3.369	117	[0: >4.591]	<.001
		5b	3.37	1.59	0.974	2.742	[0: 1.081]	.034
		Joined	1.82	0.30	–	–	H1: [0: >4.591]	<.001
Suffix HF above chance	SuffixHF Intercept	5a	0.15	0.12	3.369	0.135	[1.381 : ∞]	.223
		5b	0.37	0.13	0.974	15.067	[0: >4.591]	.004
		Joined	0.27	0.09	1.825	8.195	[0: >4.591]	.003
Suffix LF above chance	SuffixLF Intercept	5a	0.40	0.24	3.369	0.498	[0: >4.591]	.106
		5b	2.85	1.33	0.974	3.253	[0.891: >4.591]	.032
		Joined	1.30	0.26	1.825	9.5×10^{4}	[0: >4.591]	<.001
Comparing to Experiment 4b								
Prefix LF better in Exp5 than Exp4b	Main effect of experiment	Joined	-0.23	0.08	–	–	H1: [0 : 1.931], ambig: [1.941 : >4.591]	.003

[1]Intercept from same lme (method B)
[2]Prefix LF intercept from Experiment 5b (method A)
[3]Prefix LF intercept from Experiment 5a (method A)
[4]Prefix LF intercept from same lme (method B)
[5]Intercept from same lme (method B)

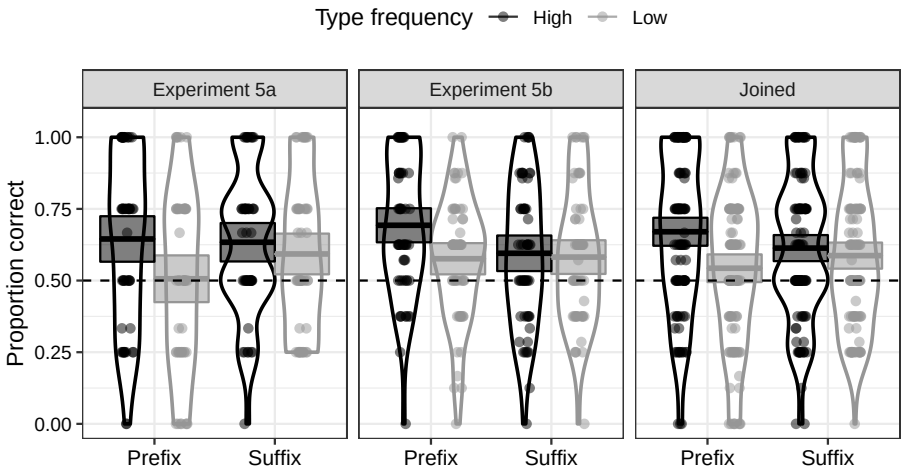


Figure 4.15: Experiment 5: Performace on the Semantics and Phonology Generalization test in Expriment 5a (left), Experiment 5b (middle), and combined data (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance

and 4b). However in Experiment 5b there was evidence for learning in every cell including Prefix-LF and in the combined data evidence for learning in this cell was ambiguous.

4.5.2.3 Semantics and Phonology Trained Items test

The data are shown in Figure 4.16 and inferential statistics in Table 4.12. Overall, performance was better on this test compared to the Semantics and Phonology Generalization test, indicating better performance on trained than on novel items in each version of the experiment (Experiment 5a: trained items: M = 66.96%, SD = 20.78%, new items: M = 59.18%, SD = 12.96%; Experiment 5b: trained items: M = 68.74%, SD = 13.22%, new items: M = 61.2%, SD = 13.22%), as well as for combined data.

As in Experiment 4, the evidence that the type-frequency manipulation had a stronger effect in the prefix than in the suffix condition was ambiguous for both versions of the experiment and the combined data, too.

4.5.2.4 Phonology Generalization test

The data are shown in Figure 4.17 and inferential statistics in Table 4.13. The evidence for the predicted better learning in the suffix condition was ambiguous in both versions of the experiment (Experiment 5a: suffix: M = 53.77%, SD = 18.34%; prefix: M = 54.4%, SD = 20.19%; Experiment 5b: suffix: M = 53.77%, SD = 15.39%; prefix: M = 52.76%, SD = 11%), as well as with combined data. In the prefix condition, the evidence for above-chance performance was ambiguous in both versions of the experiment, but combined data showed evidence for learning. In the suffix condition, the evidence for learning was ambiguous in Experiment 5a, but there was evidence for learning in Experiment 5b and with combined data.

Table 4.11: Experiment 5: Semantics and Phonology Generalization Test Statistics.

Hypothesis	Contrast in lme	Data	Mean difference	SE	H1	B	Robustness region	p
Suffix better than Prefix	Main effect affix	5a	-0.16	0.17	0.529[1]	0.162	[0.241 : ∞]	.333
		5b	0.26	0.14	0.529[1]	2.263	[0.401 : 4.01]	.071
		Joined	0.10	0.11	0.529[1]	0.491	[0 : 0.801]	.361
Greater TF effect in Prefix	Affix by TF interaction	5a	0.47	0.34	0.905[2]	1.413	[0 : 4.591]	.178
		5b	0.56	0.27	0.905	3.835	[0.211 : 1.281]	.041
		Joined	0.51	0.21	0.905	6.928	[0 : 2.411]	.016
Breaking down by affix condition								
TF effect in Prefix	TF effect	5a	0.67	0.25	0.894[3]	13.915	[0: >4.591]	.008
		5b	0.46	0.22	0.894[3]	3.394	[0 : 1.051]	.041
		Joined	0.56	0.17	0.894[3]	75.852	[0 : >4.591]	.001
TF effect in Suffix	TF in Suffix	5a	0.20	0.26	0.894[3]	0.574	[0 : 1.611]	.439
		5b	-0.02	0.20	0.894[3]	0.201	[0.521 : ∞]	0.903
		Joined	0.06	0.16	0.894[3]	0.238	[0.631 : ∞]	0.723
Cell-by-cell comparisons to chance								
Prefix HF above chance	PrefixHF Intercept	5a	0.65	0.17	0.953[4]	317	[0: >4.591]	<.001
		5b	0.98	0.16	0.953[4]	6.87×10^{7}	[0: >4.591]	<.001
		Joined	0.86	0.12	0.953[4]	6.88×10^{10}	[0: >4.591]	<.001
Prefix LF above chance	PrefixLF Intercept	5a	0.02	0.17	0.953[4]	0.188	[0.521 : ∞]	.898
		5b	0.33	0.13	0.953[4]	7.437	[0 : 2.501]	.009
		Joined	0.22	0.10	0.953[4]	2.03	[0.621 : 4.591]	.031
Suffix HF above chance	SuffixHF Intercept	5a	0.58	0.17	0.953[4]	76.4	[0: >4.591]	.001
		5b	0.45	0.15	0.953[4]	25.3	[0: >4.591]	.003
		Joined	0.50	0.11	0.953[4]	3119	[0: >4.591]	<.001
Suffix LF above chance	SuffixLF Intercept	5a	0.41	0.17	–	6.165	[0 : 2.131]	0.015
		5b	0.35	0.13	0.953[4]	11.586	[0 : 3.941]	0.005
		Joined	0.37	0.10	0.953[4]	163	[0: >4.591]	<.001
Comparing Prefix LF learning to Experiment 4								
Prefix LF better in Exp5 than 4b	Effect of experiment in PrefixLF	Joined	0.19	0.15	0.125[5]	1.801	[0 : 1.721]	0.213

[1]Intercept from Experiment 4a (method A)
[2]Affix-by-type-frequency interaction from Experiment 4a (method A)
[3]Type Frequency effect in Prefix from Experiment 4a (method A)
[4]Prefix HF Intercept from Experiment 4a (method A)
[5]Intercept from the same lme (method B)

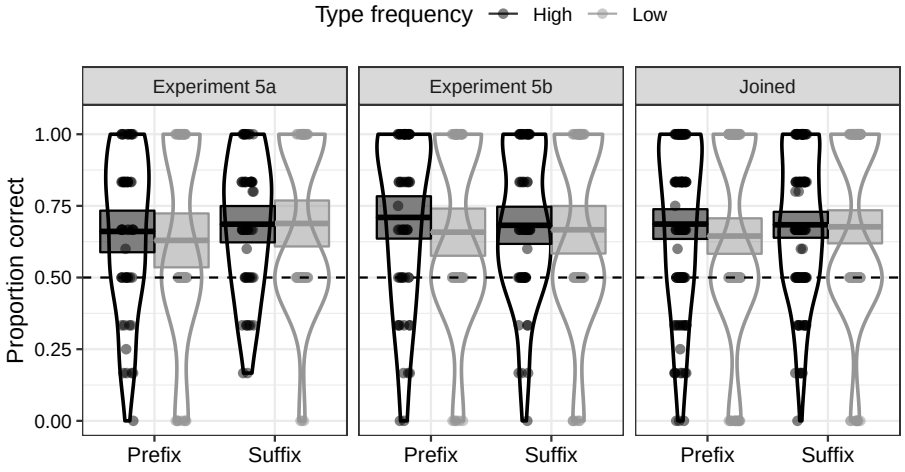


Figure 4.16: Experiment 5: Proportion of correct responses on the Semantics and Phonology trained items test in Experiments 5a (left), 5b (middle) and with combined data (left). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance

Table 4.12: Experiment 5: Semantics and Phonology Trained Items Test Statistics.

Hypothesis	**Contrast in lme**	**Data**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Suffix better than Prefix	Main effect affix	5a	-0.16	0.20	0.52[1]	0.23	[0.331 : ∞]	.44
		5b	0.15	0.25	0.52[1]	0.72	[0 : 1.261]	.535
		Joined	-0.01	0.16	0.52[1]	0.29	[0.441 : ∞]	.977
Greater TF effect in Prefix	Affix by type-freq interaction	5a	0.16	0.38	0.93[2]	0.54	[0 : 1.611]	.67
		5b	0.20	0.37	0.93[2]	0.58	[0 : 1.751]	.586
		Joined	0.17	0.26	0.93[2]	0.49	[0 : 1.391]	.52
Cell-by-cell comparisons to chance								
Prefix HF above chance	Prefix HF Intercept	5a	0.80	0.18	0.95[3]	5019	[0 : >4.591]	<.001
		5b	1.19	0.22	0.95	6.59×10^5	[0 : >4.591]	<.001
		Joined	0.99	0.14	0.95	9.19×10^9	[0 : >4.591]	<.001
Prefix LF above chance	Prefix LF Intercept	5a	0.53	0.20	0.95	12.33	[0 : >4.591]	.008
		5b	0.70	0.20	0.95[3]	117	[0 : >4.591]	0.001
		Joined	0.62	0.14	0.95[3]	3018	[0 : >4.591]	<.001
Suffix HF above chance	Suffix HF Intercept	5a	0.92	0.18	0.95[3]	58469	[0 : >4.591]	<.001
		5b	0.99	0.21	0.95[3]	15261	[0 : >4.591]	<.001
		Joined	0.95	0.14	0.95[3]	2×10^9	[0 : >4.591]	<.001
Suffix LF above chance	Suffix LF Intercept	5a	0.81	0.21	0.95[3]	462.97	[0 : >4.591]	<.001
		5b	0.70	0.20	0.95[3]	120.28	[0 : >4.591]	.001
		Joined	0.75	0.15	0.95[3]	1.16×10^5	[0 : >4.591]	<.001
Comparison to generalization test 3								
Trained items better than generalization	Main effect item novelty	5a	0.11	0.04	0.23[4]	16.60	[0 : 1.431]	.005
		5b	0.40	0.11	0.23	149	[0: >4.591]	<.001
		Joined	0.58	0.13	0.23	1170	[0: >4.591]	<.001

[1]Intercept from generalization lme with Experiment 4a data (method A)
[2]Same effect from generalization lme with Experiment 4a data (method A)
[3]Prefix HF intercept from generalization lme with Experiment 4a data (method A)
[4]Same effect from Experiment 4b (method A)

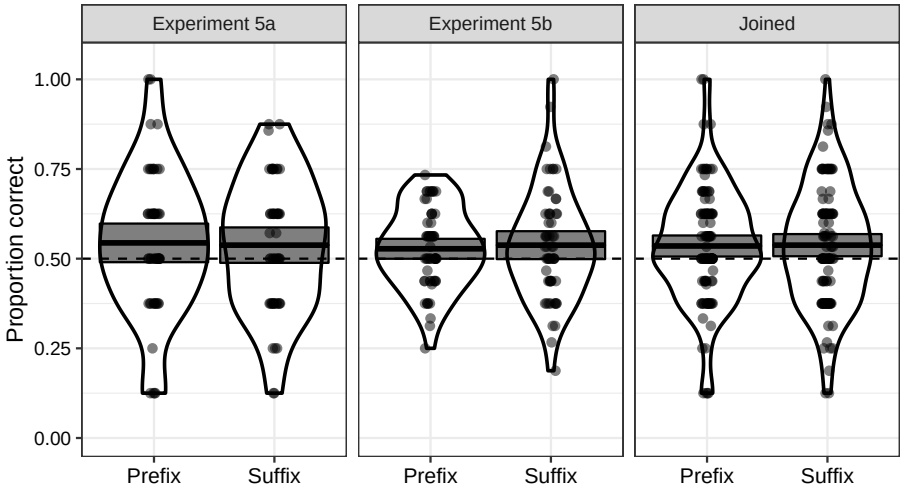


Figure 4.17: Proportion of correct responses on the Item-learning test in Experiments 5a (left), 5b (middle) and with combined data (left). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance

Table 4.13: Experiment 5: Phonology Generalization Test Statistics.

Hypothesis	**Contrast in lme**	**Data**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Suffix better than Prefix	Main effect of affix	4a	-0.03	0.15	0.25[1]	0.46	[0 : 0.361]	.85
		5b	0.03	0.10	0.25[1]	0.48	[0 : 0.371]	.73
		Joined	0.01	0.08	0.25[1]	0.35	[0 : 0.261]	.87
Breaking down by affix condition								
Prefix above chance	Prefix Intercept	5a	0.18	0.11	0.37[2]	2.01	[0 : 2.601]	.09
		5b	0.11	0.07	0.37[2]	1.29	[0 : 1.541]	.10
		Joined	0.14	0.06	0.37[2]	4.30	[0 : 0.551]	.02
Suffix above chance	Suffix Intercept	5a	0.15	0.11	0.37[2]	1.25	[0 : 1.561]	.16
		5b	0.15	0.07	0.37[2]	3.22	[0 : 0.401]	.03
		Joined	0.15	0.06	0.37[2]	7.37	[0 : 1.031]	.01

[1]Intercept from same lme with Experiment 4a data (method A)
[2]Suffix Intercept from lme with Experiment 4a data (method A)

Table 4.14: Experiment 5: Language awareness questionnaire response summary

	Number of participants reporting awareness					
	Noticed exceptions	Shape and/or colour of fribbles	Vowel in 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%)

4.5.2.5 Language Awareness Questionnaire

The same coding was used as described for Experiment 1a. Results for both versions of the experiment combined are summarized in Table 4.14. Again we asked three questions.

First, is participant's ability to generalize with low frequency nouns in the suffix condition driven by those participants who noticed the exceptions? To look at this, we removed the two participants who mentioned "exceptions" in questionnaire and check whether the suffix condition remained above chance with LF items, which we found to be the case (β = 0.347, SE = 0.101, p = .001, BF = 70.699, RR [0 : >4.591]). As a further check, we assessed whether the key type frequency interaction still held with all participants who reported noticing exceptions removed (eight from prefix, two from suffix) and found that this was the case (β = 0.473, SE = 0.22, p = .032, BF = 4.123, RR [0 : 1.351]).

Second, we asked whether participants' ability to generalize with high frequency nouns depended on their being able to describe the feature on which this generalization depends. This was again found not to be the case, with performance in both suffix and prefix conditions remaining above chance once the relevant participants (26 from prefix, 24 from suffix) were removed (prefix HF: β = 0.642, SE = 0.173, p <.001, BF = 276.806, RR [0 : >4.951]; suffix HF: β = 0.629, SE = 0.166, p <.001, BF = 359.365, RR [0 : >4.591]).

Finally, we asked whether participants' performance in generalization over phonological cues depended on being able to describe the role of vowels in the stem. This was again found not to be the case in the suffix condition (β = 0.13, SE = 0.058, p = .026, BF = 3.472, RR [0 : 0.441]), however, in the prefix condition, removing the two participants resulted in ambiguous evidence for above-chance performance (β = 0.124, SE = 0.058, p = .032, BF = 2.772, RR [0.351 : 3.361]).

4.5.3 Discussion

Experiments 5a and 5b were identical to Experiments 4a and 4b, respectively, except that participants were exposed to only half of the training set: eight items in total, three HF and one LF item per affix. We predicted that training participants on fewer items would lead to better item learning, which might in turn unmask a prefixing advantage. In addition to item learning, we were interested in whether the type frequency effect, for which there was evidence in Experiments 4a and 4b, persists with a smaller training set.

Reducing the number of trained items led to overall better item learning. Notably,

in the current experiment there was evidence of learning both LF and HF-type items (unlike in Experiment 4b, where only LF items were learned). While numerically LF items were learned better than HF items, unlike in Experiment 4b, there was no evidence of an overall benefit of LF items, possibly because both types of items were learned well in this experiment. Crucially, however, despite this stronger overall learning we did not find evidence for the overall prefixing benefit which was originally predicted, although the data here were ambiguous (unlike in Experiment 4b where there was evidence for the null). I return to this in the General Discussion of this chapter (Section 4.7).

Turning to generalization, as in Experiments 4a and 4b, a greater difference between HF and LF items in the prefix condition than in the suffix condition was predicted, as evidenced by an interaction between affix and type-frequency. Looking at the two versions of the experiment separately, although numerically in the predicted direction in both experiments, the interaction was ambiguous in Experiment 5a, but there was substantial evidence for the interaction in Experiment 5b, and when the data were combined. Given that Bayes Factors are a measure of the strength evidence, following Dienes (2016), I consider evidence from a larger sample to be more robust, and conclude that for test 3, Experiment 5 (like Experiments 4a and 4b), provided support for the interaction. However, in contrast to Experiment 4, there was no clear evidence that this interaction stemmed from participants in the prefix condition being at chance LF items. In fact, while this appeared to be the case in 5a in 5b we see substantial evidence for above chance learning. Presumably, these contradictory results are due to chance sampling, and therefore conclusions are made from the combined data, which suggest ambiguous evidence. Similarly, comparing across Experiments 4 and 5, there was no evidence that Prefix-LF items were better learned in Experiment 5 (the evidence being ambiguous for all sensible predicted differences).

The Phonology generalization test showed a similar pattern of results as Experiment 4: evidence for learning in the suffix condition, somewhat weaker evidence in the prefix condition (although here there was evidence for learning with combined data, whereas in Experiment 4 it was ambiguous) and ambiguous difference between conditions. In the General Discussion of this chapter, I consider the relatively poor learning of the phonological cues, which has been seen in all experiments so far.

Finally, in the trained items test, as in Experiment 4, there was evidence that participants did better in this test than in the semantics phonology generalization test, suggesting that they used item specific knowledge, as well as the features which were relevant in generalization. Again as in Experiment 4, the interaction between affix and type frequency, which was seen in generalization, was ambiguous. However, the pattern of results looks somewhat different across the two experiments. Whereas in Experiment 4b the performance with LF items in the suffix condition had dropped, so that there was no longer evidence for a difference between conditions (in terms of a type-frequency effect), in Experiment 5 there was strong evidence for learning of both HF and LF items in both affix conditions, with comparable means. This suggests that participants in this experiment may have based their performance on item-specific cues more so than in Experiment 4b, possibly due to there being fewer items in Experiment 5.

The findings from the LAQ revealed that, as in Experiment 4, most participants did not report explicit awareness of any patterns in the language. Critically, as in Experiment 4, there was evidence of generalization for both HF and LF semantic features, and phonological features, even when participants who explicitly reported noticing these features were removed. This suggests that learning was largely implicit, although it is important to be cautious in interpreting this type of self-report, a point I return to in the General Discussion (Section 4.7).

In sum, although Experiment 5 did not show evidence of a prefixing advantage for item-learning, the stronger effect of type-frequency on generalization in the prefix condition compared to the suffix condition persisted even with a significantly smaller input set. This is in contrast with the work suggesting that greater input variability leads to better linguistic generalization (Perry, Samuelson, Malloy, & Schiffer, 2010; Potter & Saffran, 2017), and to what would be expected if we think of generalization occurring following a critical mass of item-learning (Marchman & Bates, 1994). To further understand how the relationship between item-learning and generalization is affected by variability, in Experiment 6 participants were exposed to double the size of the original set: 32 items in total, 12 HF and four LF items per category (thus preserving the 3:1 ratio of items). The same predictions were tested as in the other experiments in this chapter.

4.6 Experiment 6

4.6.1 Method

4.6.1.1 Participants

One hundred and twenty participants (60 per condition) were recruited through Prolific Academic (prefix: Mage = 33.31, SDage = 10.1, 30 female; suffix: Mage = 33.59, SDage = 10.4, 31 female). All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for participation.

4.6.1.2 Stimuli

Same as Experiment 4, except that there were more training audio and visual items. In the current experiment, there were 32 training items in total: 16 per category, 12 high type-frequency, and four low type-frequency item per category. See Figure 4.18 for a sample training set of pictures and labels.

4.6.1.3 Procedure

Same as Experiment 4, except that in training, there were eight exposures per item (same total number of training items as in Experiment 4).

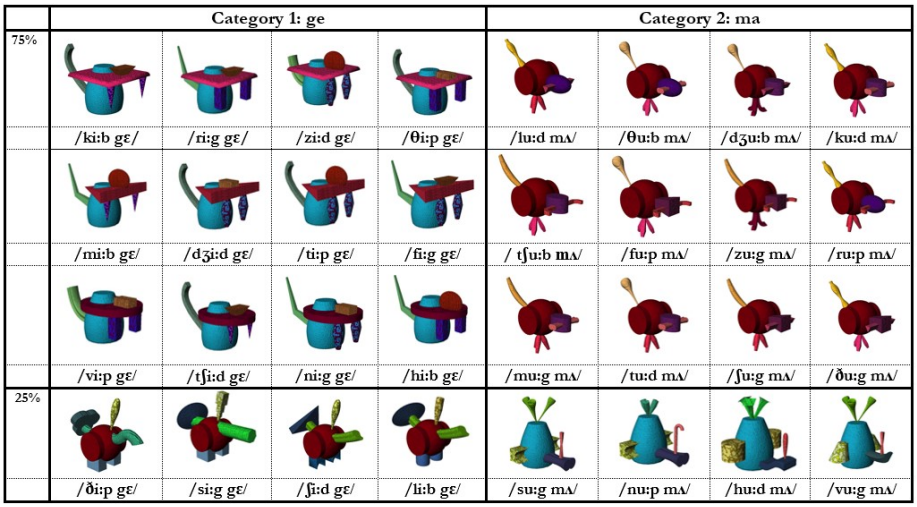


Figure 4.18: Experiment 6: Sample training set. Note that within each category, nouns were assigned to pictures randomly on a participant-by-participant basis.

4.6.2 Results

As per the optional stopping procedure, the data were analysed at 20 participants per condition, 30, and finally 40, where there was evidence for the null for one of our key hypotheses. However, given that our resources allowed it, and in line with the sampling in Experiments 4b and 5b, data from another 20 participants per condition were collected and therefore the data were expected at 60 per condition, too.

4.6.2.1 Item learning

The data are shown in Figure 4.19 and inferential statistics in Table 4.15. The evidence for an overall prefix benefit was ambiguous (prefix: M = 53.92%, SD = 13.79%; suffix: M = 54.24%, SD = 14.18%), and, in contrast to the previous experiments, the evidence for the overall LF benefit was ambiguous, too (LF: M = 55.06%, SD = 16.88%; HF: M = 52.97%, SD = 16.88%), possibly due to ambiguous evidence for learning in every cell except Suffix-HF, where there was evidence for no learning. An exploratory comparison across experiments revealed that overall item learning was better in Experiment 1b compared to Experiment 3. This is evidenced by the fact that any value of H1 smaller than 3.781 would give evidence for the H1, corresponding to predicting chance-level performance in Experiment 3 and 97.7% accuracy in Experiment 1b (as this is greater than a value we would plausibly use, we conclude that there is evidence for better learning in Experiment 1b). There was also evidence for better learning in Experiment 2 compared to Experiment 3, as evidenced by the fact that any plausible value of H1 would give evidence for this difference.

4.6.2.2 Semantics and Phonology test: Generalization

The data are shown in Figure 4.20 and inferential statistics in Table 4.16. As in the previous experiments, there was no overall suffixing benefit (suffix: M = 62.65%, SD = 17.47%; prefix: M = 61.03%, SD = 16.41%) with substantial evidence for the null. Unlike the

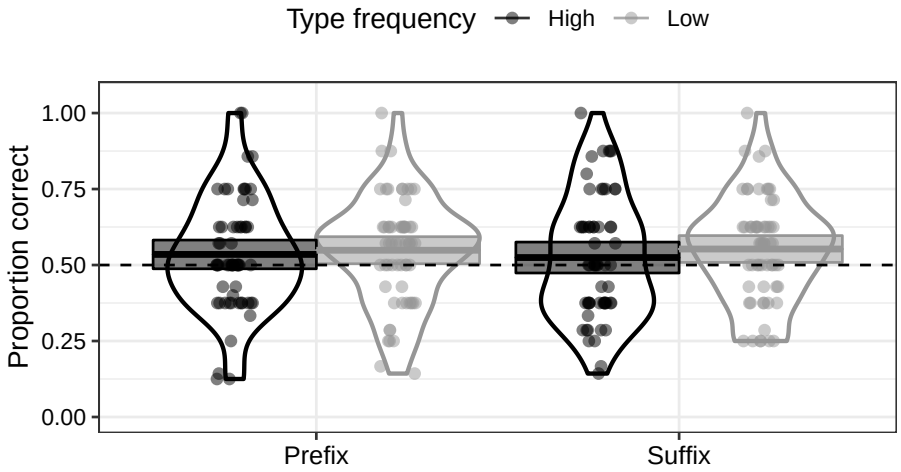


Figure 4.19: Experiment 6: Proportion of correct responses on the Item-learning test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

previous two experiments, there was substantial evidence for no an affix by type frequency interaction, and there was evidence for a benefit of HF over LF in both affix conditions, with evidence for above-chance performance on both HF and LF items.

4.6.2.3 Semantics and Phonology test: Trained items

The data are shown in Figure 4.21 and inferential statistics in Table 4.17. As in Experiment 5, the evidence for a stronger effect of type frequency in the prefix condition compared to the suffix condition was ambiguous (unlike Experiment 4b, where there was evidence for the null).

A further analysis compared the performance on this test with the generalization test, and this showed ambiguous evidence for a benefit of trained items (trained: M = 60.944%, SD = 18.714; new: M = 61.839%, SD = 16.898). An exploratory analysis compared the benefit of trained items across experiments. There was strong evidence for a greater benefit of trained items in Experiment 5 compared to Experiment 6 (indicated by strong evidence for the effect for any plausible value of H1).

Compared to Experiment 4b, we would have found evidence for the null (no difference in the trained-item benefit between Experiments 4b and 3) for any value of H1 greater than 1.741, corresponding to predicting that there would be a 1% benefit of trained items in one experiment and a benefit of a least 5.5% in the other.

4.6.2.4 Phonology Generalization

Data are shown in Figure 4.22 and inferential statistics in Table 4.18. There was evidence for no benefit of suffixing over prefixing (suffix: M = 50.11%, SD = 12.64; prefix: M = 52.39%, SD = 13.56). However, there was evidence for no learning in the suffix condition,

Table 4.15: Experiment 6: Item Learning Test Statistics.

Hypothesis	Contrast in lme	Mean difference	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect of affix	0.01	0.10	0.27[1]	0.36	[0 : 0.291]	.92
LF better than HF items	Main effect of type frequency	0.08	0.09	0.27[1]	0.74	[0 : 0.661]	.38
Cell-by-cell comparisons to chance							
Prefix HF above chance	PrefixHF Intercept	0.12	0.10	0.97[2]	0.40	[0 : 1.161]	.20
Prefix LF above chance	PrefixLF Intercept	0.18	0.09	0.97[2]	1.17	[0.331 : 3.511]	.06
Suffix HF above chance	SuffixHF Intercept	0.09	0.10	0.97[2]	0.25	[0.711: ∞]	.36
Suffix LF above chance	SuffixLF Intercept	0.20	0.09	0.97[2]	1.63	[0.491: >4.591]	.04
Comparison of overall item-learning with Experiments 4b and 5							
Exp4b better than Exp6	Exp4b vs Exp6 contrast	0.23	0.08	–	–	H1: [0 : 3.771], ambig: [3.781 : >4.591]	.003
Exp5 better than Exp6	Exp5 vs Exp6 contrast	0.45	0.08	–	–	H1: [0 : >4.591]	<.001

[1]Intercept from Experiment 5a (method A)
[2]Prefix LF intercept from Experiment 5a (method A)

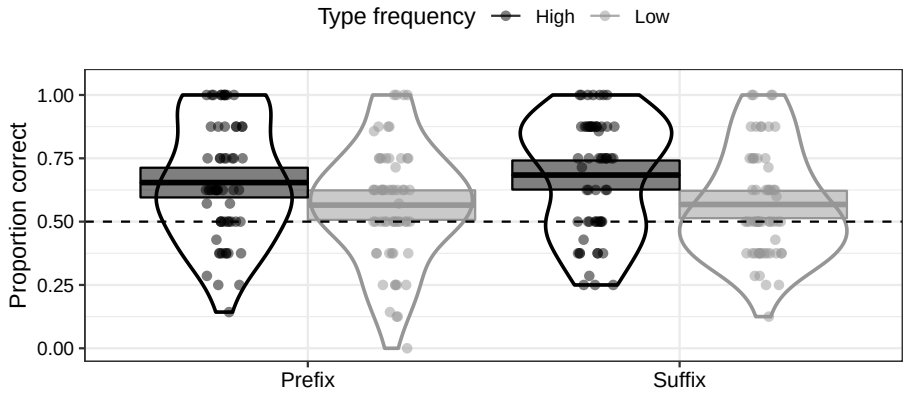


Figure 4.20: Experiment 6: Performance on the Semantics and Phonology Generalization test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

Table 4.16: Experiment 6: Semantics and Phonology Generalization Test Statistics.

Hypothesis	**Contrast in lme**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Suffix better than Prefix	Main effect of affix	-0.08	0.15	0.52[1]	0.19	[0.271 : ∞]	.569
Stronger type frequency effect in Prefix	Affix by TF interaction	-0.14	0.26	0.93[2]	0.18	[0.481 : ∞]	.594
Breaking down by affix condition							
Type frequency effect in Prefix	TF in Prefix	0.46	0.18	0.80[3]	9.02	[0 : 3.071]	.011
Type frequency effect in Suffix	TF in Suffix	0.60	0.18	0.80[3]	66.22	[0: >4.591]	.001
Cell-by-cell comparisons to chance							
Prefix HF above chance	PrefixHF Intercept	0.74	0.15	0.95[4]	44846	[0: >4.591]	<.001
Prefix LF above chance	PrefixLF Intercept	0.28	0.12	0.95[4]	3.30	[0: 1.051]	.022
Suffix HF above chance	SuffixHF Intercept	0.90	0.15	0.95[4]	5.77×10^{6}	[0: >4.591]	<.001
Suffix LF above chance	SuffixLF Intercept	0.30	0.12	0.95[4]	4.46	[0: 1.461]	.016

[1]Intercept from Experiment 4a (method A)
[2]Same effect from Experiment 4a (method A)
[3]Effect of type frequency in Prefix in Experiment 4a (method A)
[4]Prefix HF intercept from Experiment 4a (method A)

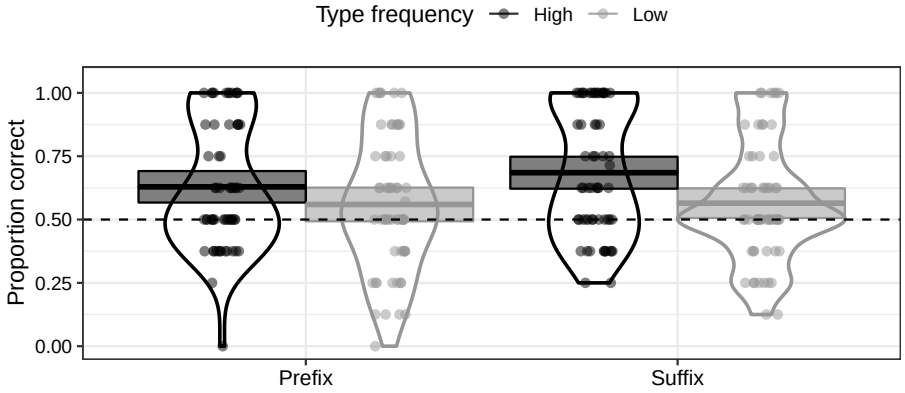


Figure 4.21: Experiment 6: Performance on the Semantics and Phonology Trained Items test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

and ambiguous evidence in the prefix condition.

4.6.2.5 Language Awareness Questionnaire

Results are summarized in Table 4.19. Given that there was no evidence for learning of phonological cues, our analyses focused on the awareness of semantic cues.

First, is participant's ability to generalize with LF items in both conditions driven by those participants who noticed the exceptions? To look at this, we removed the participants who mentioned "exceptions" in questionnaire and check whether the two conditions remained above chance with LF items. This was found to be the case in the prefix condition ($\beta = 0.295$, SE = 0.131, p = .024, BF = 3.237, RR [0.09 : 1.03]). In the suffix condition, however, the evidence for above-chance performance with LF items became ambiguous once the three participants who reported eplixict awareness of exceptions were excluded ($\beta = 0.246$, SE = 0.127, p = .052, BF = 2.993, RR [0.45 : 4.591]).

Second, we asked whether participants' ability to generalize with HF items depended on their being able to describe the feature on which this generalization depends. This was found not to be the case, with performance in both suffix and prefix conditions remaining above chance once the relevant participants (10 from prefix, 14 from suffix) were removed (prefix HF: $\beta = 586$, SE = 152, p <.001, BF = 504, RR [0.05 : >4.951]; suffix HF: $\beta = 0.803$, SE = 0.163, p <.001, BF = 44456, RR [0.04 : >4.591]).

4.6.3 Discussion

Experiment 6 was identical to Experiments 4a, 4b, and 5, except that participants were exposed to a larger input set: 24 items in total, 12 HF and four LF items per category. An unexpected finding in this study is the evidence for an advantage of HF items over LF items in both affix conditions in the Semantics and Phonology Generalization test. Recall that in previous experiments we only found this HF-advantage in the prefix condition. In

Table 4.17: Experiment 6: Semantics and Phonology Trained Items Test Statistics.

Hypothesis	Contrast in lme	Mean difference	SE	H1	B	Robustness region	p
Suffix better than Prefix	Prefix vs Suffix	-0.17	0.18	0.56	0.17	[0 : 1.401]	.323
Stronger effect of type frequency in Prefix	Affix-by-type-frequency interaction	-0.30	0.30	0.93	0.16	[0 : 2.451]	.312
Cell-by-cell comparisons to chance							
Prefix HF above chance	Prefix HF vs chance	0.66	0.17	0.95	381.69	[0 : >4.591]	<.001
Prefix LF above chance	Prefix LF vs chance	0.29	0.15	0.95	1.82	[0.511 : >4.591]	.055
Suffix HF above chance	Suffix HF vs chance	0.99	0.18	0.95	655802	[0 : >4.591]	<.001
Suffix LF above chance	Suffix LF vs chance	0.31	0.15	0.95	2.58	[0.801 : >4.591]	.036
Comparing to generalization							
Trained items better than new	Main effect item novelty	0.01	0.08	0.23	0.35	[0 :0.231]	.938
Comparison of the benefit of trained items with Experiments 4b and 5							
Exp4b greater than Exp6	Item-novelty by experiment interaction	0.17	0.13	–	–	ambig: [0 : 1.74], H0: [1.75: ∞]	.175
Exp5 greater than Exp6	Item-novelty by experiment interaction	0.35	0.12	–	–	H1: [0 : >4.591]	.003

[1]Intercept from generalization Experiment 4a (method A)
[2]Same effect from generalization Experiment 4a (method A)
[3]Prefix HF intercept from generalization Experiment 4a (method A)
[4]Same effect from Experiment 5a (method A)

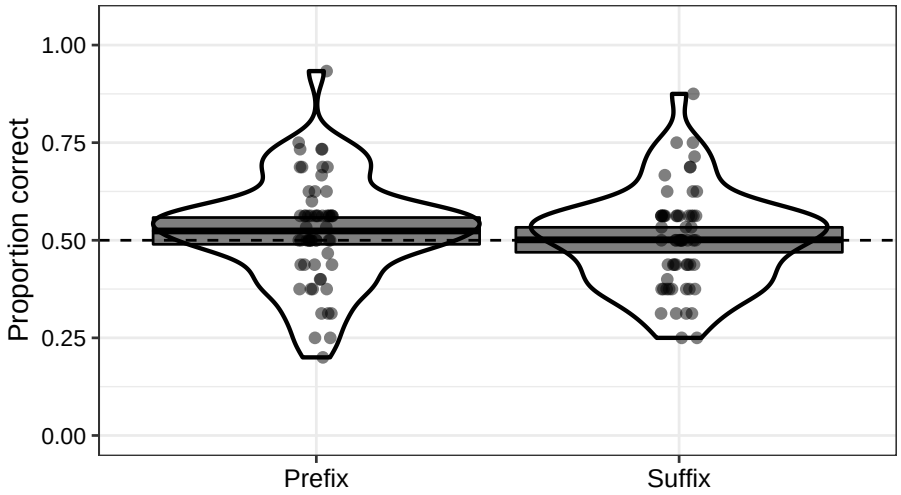


Figure 4.22: Experiment 6: Performance on the Phonology Generalization test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance.

Table 4.18: Experiment 6: Phonology Generalization Test Statistics.

Hypothesis	**Contrast in lme**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Suffix better than Prefix	Suffix vs Prefix	-0.09	0.10	0.25[1]	0.20	$[0 : \infty]$	.34
Prefix above chance	Prefix Intercept	0.09	0.07	0.37[2]	0.82	$[0 : 0.951]$	.16
Suffix above chance	Suffix Intercept	0.00	0.07	0.37[2]	0.18	$[0 : \infty]$	.97

[1]Intercept from Experiment 4a Phonology test 1 (method A)
[2]Suffix Intercept from Experiment 4a Phonology test 1(method A)

Table 4.19: Experiment 6: Language awareness questionnaire response summary

	Number of participants reporting awareness					
	Noticed exceptions	Shape and/or colour of fribbles	Vowel in nouns	Vowel and shape	Other	None
Prefix	6 (10%)	10 (16.6%)	2 (3.33%)	0	6 (10%)	36 (60%)
Suffix	3 (5%)	14 (23.3%)	2 (3.33%)	1 (1.67%)	6 (10%)	34 (56.6%)

Experiment 6, both conditions were above chance with LF items (unlike Experiment 4 where the prefix condition was at chance with these items), but performance was better with HF items in both conditions. Note that numerically the suffix condition was poorer with LF items in this experiment than in previous experiments, and there is some indication that the above-chance performance was driven by three participants who reported noticing "exceptions" to the pattern, suggesting that overall learning of LF discriminating features was poor in the suffix condition (in the prefix condition, removing six participants who reported explicit awareness did not affect overall above-chance learning of LF features). I return to this intriguing finding in the General Discussion of this chapter.

Item-learning was poorer in this experiment compared to the previous ones. Performance on the item-learning test was generally low (although the evidence was ambiguous in every cell except Suffix HF), and there was evidence that it was weaker than in Experiments 4b and 5 (we did not compare with Experiment 4a because a different item learning test was used there). In addition, there was strong evidence that the benefit of trained items compared to novel items in the Semantisc and Phonology tests was weaker in Experiment 6 than in Experiments 5 and 4b. While the evidence for the difference between trained and novel items was ambiguous in Experiment 6, the fact that it was weaker than in the previous experiments suggests that performance on the trianed items may have been based more on the abstraction of features than on item-based idiosyncrasies. I revisit this point in the General Disucssion of this chapter.

4.7 General Discussion of Study 2

The experiments presented in this chapter followed up on a possibility raised in Study 1, that the predicted suffixing benefit in generalization can be observed most clearly when learning critically depends on cue competition – specifically, when successful generalization depended on "unlearning" frequent, but unpredictive cues, and reinforcing predictive cues. The Rescorla-Wagner learning model (Rescorla & Wagner, 1972), implemented in Experiment 3 of this chapter, demonstrated that only in the suffix condition can this "unlearning" occur, due to the fact that cue competition only happens in this condition. This finding was tested with human learners in Experiments 4-6. Although there was no overall benefit of suffixing over prefixing in generalization in any experiment, there was evidence for the difference in "unlearning" frequent but uninformative cues between suffixes and prefixes in Experiments 4 and 5. This was evidenced by the prefix condition performing better with high type-frequency items (where the most frequent cue was also the predictive, discriminating cue) compared to low type-frequency items (in which the most frequent cue was unpredictive), and with the suffix condition performing equally well with both types of items. Intriguingly, however, this effect was absent (with evidence for the null) in Experiment 6, in which participants were trained on a large set of items.

The second result was with regard to item-learning, where we predicted a benefit of prefixing over suffixing (based on Arnon & Ramscar, 2012; Dye et al., 2017, 2018; Ramscar, 2013), but did not see this result in either experiment, with evidence for the null in Experiment 4, and ambiguous evidence in Experiments 5 and 6. Each of these findings is discussed in turn below.

4.7.1 Generalization

Our theory predicts that generalization is affected by type-frequency in the prefix condition only, and not in the suffix condition. In the test with the largest sample (i.e. test 3, where samples can be joined across experiments), there was substantial support for this prediction in Experiments 4 and 5: the suffix condition generalized HF and LF items equally well, whereas the prefix condition were better at generalizing HF items than LF items. This result can be explained in terms of different cue structure in the two conditions: 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. Specifically, greater cue competition in the suffix condition meant that an uninformative but highly frequent cue (body shape) lost predictive value to more informative cues. Therefore, when it came to generalizing LF items, participants in the prefix condition were conflicted between a highly frequent, uninformative cue (body shape) and informative cues, resulting in poorer generalization of LF items compared to HF items. Participants in the suffix condition, on the other hand, had "unlearned" body shape as an uninformative cue and were able to correctly generalize on the basis of informative cues. This effect was clearest in Experiment 4 and somewhat weaker in Experiment 5. In Experiment 6, however, the effect was absent altogether – both conditions showed a type-frequency effect. This is an

intriguing finding which we did not predict. However, in Chapter 5, I re-visit the model presented in Experiment 3 and show that the effect of type-frequency in the suffix condition is predicted early on in learning, and discuss why this may be applicable to Experiment 6.

In the experiments reported in this chapter, there was no overall suffixing advantage in generalization in either of the experiments, with evidence for no advantage in Experiments 4 and 6, and ambiguous evidence in Experiment 5. Overall better generalization (i.e., better learning of discriminating features) in the suffix condition is predicted by previous work (Ramscar et al., 2010; St Clair et al., 2009), and was demonstrated in the computational model in Experiment 3. In Experiments 4 and 5, however, a suffixing advantage was observed for LF items only, and even though this was not tested statistically, generalization of HF items was numerically higher in the prefix condition. There may be multiple reasons for this pattern of results. Recall that our theoretical approach defines generalization as the dissociation of uninformative cues – both shared across items (body shape), *as well as* idiosyncratic cues associated with individual items. Our paradigm is well designed for tapping into the learning, that is, "unlearning" of body shape (by manipulating its frequency), which is particularly important for LF items. However, the paradigm might not be sensitive to other aspects of learning where suffixing should theoretically be an advantage, such as the dissociation of idiosyncratic cues. Testing the dissociation of idiosyncratic cues might be particularly important for seeing a suffix advantage for HF items, where not only is it not necessary to dissociate body shape as a cue, but where body shape can actually be used as a cue for generalization.

The pattern of results in the prefix condition is indeed suggestive of generalization on the basis of body shape – high performance on HF items and chance-level performance on LF items. In addition, the single cue in question – body shape – is also likely to be more perceptually salient and familiar as a cue to word meaning than other cues in our stimuli, which may have biased participants towards even stronger generalization on the basis of this cue in the prefix condition (in the suffix condition, this bias would have been "overridden" by negative evidence). However, this is not captured by the theory or in the computational simulations, which model a naïve learner; adult human learners, on the other hand, bring learning biases and years of experience with language to the experimental task.

To summarize, the differences between conditions in task difficulty and the effects of potential learning biases might mean that in the suffix condition more exposure would have resulted in even greater dissociation of uninformative cues and reinforcement of informative cues, and in turn stronger generalization of *both* HF and LF items compared to the prefix condition (note that, by comparison, simulations in Experiment 3 were trained to asymptote in both conditions). Finally, an overall suffixing advantage is also predicted by previous work discussed in the Introduction (Ramscar et al., 2010; St Clair et al., 2009). However, the learning task in the current study was more complex – not only did participants learn to associate the two affixes with correct informative features, but also to associate individual fribbles with individual "nouns", while encoding simultaneous features in two modalities (visual and auditory). This, combined with the effects of relatively short exposure discussed above, might explain why there was no overall suffix advantage in generalization – I explore

this in further simulations in Chapter 5.

The fact that there was no overall benefit of suffixing in the Semantics and Phonology Generalization test is consistent with the lack of differences in the Phonology test. Here, since each vowel occurred consistently with one affix, learning did not require "unlearning" a salient but uninformative feature, although again theoretically speaking, suffixing should have provided the opportunity to dissociate less relevant phonetic features and thus boost generalization over the informative vowel feature. The current study, however, is inconclusive as to this issue, because the evidence for a main effect of affix was ambiguous (rather than supporting the null) in all three experiments, possibly because the learning of phonology was overall poor across the experiments. This is contrary to St Clair et al. (2009) who found above-chance generalization in both affix conditions, with suffix condition being significantly better than prefix. However, that study involved phonological cues only (participants were not presented with pictures), whereas in our study, the input contained redundant cues: learners could make correct generalizations on the basis of semantic cues alone or phonological cues alone. It is possible that semantic cues were more salient than phonological cues in our study, and were learned first, which may have in turn blocked the learning of phonology (see Kamin, 1968, for a discussion of blocking effects). 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 cues which were more salient. 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 the paradigm presented in this chapter, it is not possible to draw strong conclusions about the learning of phonology and semantics since the manipulation of these cues is different (there was no comparable type-frequency manipulation with phonological cues). Nevertheless, it is notable that phonological learning in this work is relatively weak, despite the vowel being theoretically the most consistent cue to affix usage. Again, it is worth recalling that our theoretical approach 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 different role played by these different information sources is an important area for future research.

A final question is the extent to which the mechanisms of generalization in this paradigm are implicit. Analysing the responses on the language awareness questionnaire, majority of participants did not report noticing any of the features relevant to generalization, and generalization remained even when those participants who did notice the relevant features were removed. While this method is imperfect (as discussed throughout this work), note that in other artificial language work questionnaires have been useful in revealing that learning is largely explicit. Brown et al. (2018) taught learners artificial languages somewhat similar to those presented here, with semantic cues to affix use (animals co-occurred with one affix, vehicles with another). When treated as a group, participants' performance in generalization test with novel nouns was above chance. However, the questionnaire data revealed this was driven by a small group of participants who had explicitly noticed the

semantic condition: these "explicit" participants were near ceiling in their responses, while others remained at chance. The experiments in this chapter did not show this pattern and thus, I tentatively conclude that learning was largely implicit, and thus more analogous to naturalistic first language learning. In addition, it is unclear how explicit learning methods could lead to the benefit of suffixing for learning of informative features that was observed in Experiments 4 and 5.

4.7.2 Item learning

The studies in this Chapter included two tests of participants' knowledge about the specific items on which they had been trained. The Item Learning test tested participants' recall of trained fribbles. Contrary to the prediction, there was no evidence for a prefixing advantage, which was predicted on theoretical grounds (Dye et al., 2017, 2018) and on the basis of Arnon and Ramscar (2012). I note that the conditions which are compared in this chapter and Arnon and Ramscar (2012) are different – they compared learning to identical input where the ordering of that input was changed (participants were either exposed to articles + nouns before hearing nouns in isolation, or the reverse); the current work compared learning of different exposure sets, one in which nouns are preceded by affixes (akin to their articles), another in which nouns are followed by affixes. Nevertheless, Arnon and Ramscar (2012) reported better learning of noun-object pairings (as well as article-noun pairs which was in fact the key focus of that study) and explain this in terms of a theory in which prefixes (like articles) aid learning and processing by reducing the entropy of upcoming nouns (in line with the approach in Dye et al., 2017, 2018). Given this, we predicted better item-learning under prefixing compared to suffixing, since in the latter condition the affix is not encountered until after the noun, and therefore the entropy for that noun cannot be smoothed over the affix-noun pair. However there was no evidence for this effect in the current study.

One possibility is that the nature of our stimuli and the test are not suitable for capturing this effect. In Arnon and Ramscar (2012), item-learning was tested by showing participants a trained picture and playing two trained labels, one in which the noun matched the picture, and one in which the noun was incorrect. In Experiment 4a, 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 4a showed a floor effect, which is why in Experiment 4b (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 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. In addition to this, while the test required participants to distinguish between different items within the same category and frequency type, it is possible that within-category distinctions in the input set 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 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. Study 3 (Chapter 6) investigates item-learning in an input set where individual items are more distinguishable: specifically, phonological cues were removed by using maximally discriminable artificial labels (for example: *tombat, deecha, paylig*, as opposed to: *foop, moob, kood*), and visual items with more distinctive within-category variation are used.

Although it was not part of the initial predictions, a notable finding in the item test was that, where there was learning, it was better for LF items than for HF items, despite the fact that individual items occurred equally frequently in the input (the frequency manipulation is at the type level). This finding can be explained in terms of different levels of uncertainty in the different categories. Specifically, individual LF-items involved lower entropy (less uncertainty) than with HF-items, a consequence of there being fewer LF-items. During training, in Experiment 4b, once participants discriminated which category and which type of item they were presented with, for HF-type, there were six possible items to distinguish between, whereas for LF-type, there were only two possible items to distinguish. This may have made it harder to memorize the individual HF items, and at test, participants were unable to distinguish them from novel items. In Experiment 5, participants were able to memorize individual items from groups of three (HF), however, this was still poorer than LF items, where there was only one item per affix (in Experiment 6 we did not find evidence for item-learning in any cell). It is also possible that the learning of LF items was qualitatively different across the two experiments in which learning was observed. Given that in Experiment 4b there were two LF items, participants may have learned them as a "category" on their own, whereas in Experiment 5, there was only one LF item per affix, and these items may have been learned as an "exception to the rule". Therefore learning two "categories" per affix in Experiment 4b may have been harder than learning a "category" and an exception.

4.7.3 Relationship between item-based learning and generalization

An additional test of trained items was included, which had the same design as the generalization test, but was done with trained items instead. The prediction for this test was less clear, given that participants could complete it based on the association between informative features and affixes, in which case a similar pattern to the generalization test is predicted, or on the basis of item knowledge, in which case we might expect better learning under prefixing. However, the results from this test did not fit either of these patterns. Critically, in all experiments in this chapter, there was no evidence for the greater effect of type-frequency in the prefix condition compared to the suffix condition with trained items. While the evidence for (no) interaction was ambiguous in all experiments, meaning that this result must be taken with caution to avoid over-interpreting, it is still useful to reflect on why this may have been the case. I (tentatively) argue that the reason for this might be different across the three experiments.

The explanation is simplest in Experiment 5, where performance was above-chance and relatively flat across the two levels of type frequency in the two conditions. Here it seemed that participants were able to use knowledge about which specific noun-fribble pairs co-occur with each affix to aid performance, thus boosting performance with the LF items in the prefix condition compared with the generalization test. This pattern of results is indicative of what Tomasello (2000) describes as item-based learning in natural language, where children use a particular construction (in our case, an affix) with certain items from the input, but are unable to generalize it to novel items.

In contrast, in Experiment 6, performance on the trained items test was comparable to that of the generalization test, in that performance was on average higher on HF than LF items in both conditions (although the evidence for the interaction was ambiguous with trained items but there was evidence for the null with novel items). Again, being cautious of over-interpreting ambiguous results, informally, performance on the trained and novel items tests was more similar to each other in Experiment 6 than in Experiment 5 – this is interesting given that the input set in Experiment 6 was three times the size of the set in Experiment 5. This intuition is further corroborated by evidence for a stronger effect of item-novelty (better performance with trained compared to novel items) in Experiment 5 than in Experiment 6.

Therefore, as the set size increases, learning seems to move away from individual items. However, in Experiment 4b, performance in the prefix condition was above-chance in generalization but ambiguous in the trained items test (performance in the prefix condition remained similar to that in the generalization test); similarly, in Experiment 6, performance with LF items was ambiguous with trained items, but above-chance with novel items in both conditions. This is hard to explain in terms of our theory, though in the Section 4.4.3 I suggested various aspects of our experimental set up that may have interfered with performance in this test.

Whatever the explanation, I note that the lower performance with trained LF items in Experiments 4b (suffix condition) and 6 than with novel items, as well as strong item-learning of LF items in the prefix condition and subsequent failure to generalize these items (Experiment 4b, 5b), suggest a nuanced relationship between item-learning and generalization, which is affected by information structure and cue competition, as well as the saliency and complexity of the generalization, and the items, in question.

Study 2 therefore suggests that what has traditionally been described as abstract generalization can be understood as discrimination between informative or uncertainty-reducing cues and uninformative cues, that is, cues which do not reduce uncertainty; critically, this process cannot happen without cue competition and prediction error from negative evidence. What is the advantage of re-thinking generalization in this way – what does this contribute in comparison to existing alternatives?

Within the nativist/generativist framework, Pinker and Prince (1988) discussed the so-called U-shaped curve in the proportion of children's correct use of verb past tense in English. Famously, children go through a stage of producing correct past tense forms (e.g., *went*) to a stage of overregularizing (*goed*), eventually converging on adult-like production of

regular and irregular forms. Pinker and Prince (1988) described the shift from remembering individual forms to generalizing (albeit incorrectly at first) as "the process of coming to recognize that two forms constitute the present and past tense variants of the same verb" (p. 44), but offer no testable model of this process of recognition (see also Marcus et al., 1992).

Connectionist models, on the other hand, showed that generalization is affected by the number of items in the input, and Marchman and Bates (1994) demonstrated this empirically. They found that the child verb vocabulary size predicts the child's use of irregular past tense. From this, they propose that generalization occurs after a "critical mass" of item-learning, and suggest that item-learning and abstraction are on a continuum underpinned by the same learning mechanism – but it remains unclear what this mechanism is.

In Study 2, we have seen indications that, as the number of items in the input set (vocabulary size) grows, learning becomes less item-based. While consistent with the critical mass hypothesis, this finding can be described in terms of a set of independently defined principles of discriminative learning. Specifically, what is traditionally thought of as abstract generalization becomes (in discriminative learning terms) the association between informative cues and the relevant forms, and dissociation of uninformative cues strong enough to generate the correct form to some criterion. By the same token, what Marchman and Bates (1994) and others view as item-based learning, can be viewed as the degree to which less informative, idiosyncratic cues are associated with the outcome. This process may be precisely modelled in terms of cue competition and prediction error. Note, however, that while Study 2 showed indications of moving from item-based learning to generalization as more input is accrued, the study was not designed with the intention to test hypotheses regarding this process per se. Future work should therefore formulate more precise predictions about the interaction between item-based learning and generalization (by modelling the process using the Rescorla-Wagner model), and test them with human learners.

Chapter 5

Unexpected findings from Studies 1 and 2: re-visiting the models

5.1 Introduction

There were two findings in Studies 1 and 2 with respect to generalization which were opposite to what was predicted by the discriminative learning theory and the computational models reported in Experiments 1 and 3: the prefixing advantage in generalization in Experiment 2, and the type-frequency effect in generalization in the suffix condition in Experiment 6. In the remainder of this chapter, I address these findings by revisiting the models used to generate the predictions for these experiments. It is important to note that I do not claim that this work "proves" the unexpected findings were in fact somehow predicted all along – they were not. Instead, the goal of this chapter is to identify conditions under which the discriminative learning model performs in a way similar to what was observed with humans (but was not predicted). Future work should then test these predictions in new experiments with humans. Therefore, this chapter represents a step in the scientific method where new hypotheses are generated following new empirical observations.

5.2 Prefixing advantage in generalization (Experiment 2) and numerical prefixing advantage with HF items (Experiments 4-5)

In Experiment 2, the prefix condition performed better than the suffix condition in generalization. I identified several aspects of our paradigm which may have caused the prefix advantage in generalization (Section 3.5), and these aspects were modified in Experiments 4-6, in which there was no overall prefixing advantage. However, even though this was not tested statistically, in Experiments 4 and 5 the prefix condition was numerically better than the suffix condition at generalization with HF items. This pattern of results therefore raises the possibility that there may indeed be contexts in which prefixing may facilitate better generalization compared to suffixing. Below I revisit the simulations in Experiments 1 and 3 to investigate the possibility that the prefix condition may indeed be better at

generalization than the suffix condition under certain conditions.

Observing the raw weights in Experiment 1 (Figure 3.3) shows that the difference between the weights for the correct affix and the incorrect affix is grater in the prefix condition, whereas in Experiment 3 this is true for HF items (Figure 4.2). Our normalization metric, the Luce's choice axiom (Luce, 1959), however, did not capture this potential prefixing advantage. The choice axiom 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. In our terms, this means that the probability of an item – the sum of associative weights – for one affix is related to the probability of that item with the other affix. Viewed in this way, our metric rewards greater discrimination between the two options, that is, stronger "unlearning" of the opposite affix. As the raw weights for the opposite affix have been consistently higher in the prefix condition (Experiments 1 and 3) than in the suffix condition (that is, the proportional difference between the probability of the correct affix vs incorrect affix was smaller in the prefix condition compared to the suffix), this resulted in lower chance of choosing the correct affix in that condition, as per the choice axiom. However, it is possible that "unlearning" the opposite category is less critical in Experiments 2 and for HF items in Experiments 4-5, where learning from positive evidence alone may be sufficient. Specifically, for HF items, the most frequent cues (body shape and colour of the fribbles) are also fully predictive of affix occurrence, and it is only with LF items that predictive value is contrasted with frequency. Therefore, a metric that primarily captures the learning of the correct affix (rather than the "unlearning" of the opposite affix) may be more appropriate in this case. One such metric is the softmax function (used extensively for normalization in machine learning). The softmax is related to the choice axiom, but rather than computing over raw weights, this function is computed over the exponentials of the raw weights:

$$S(w_i) = \frac{e^{w_i}}{\sum_j e^{w_j}} \tag{5.1}$$

The exponential transformation means that the higher sum of weights will be transformed into an exponentially larger value than the lower sum of weights – in other words, the function "rewards" the more strongly associated affix, rather than rewarding the greater difference between the two affixes (which is the case with the choice axiom). Therefore, if the correct affix is also the affix with the greater sum of weights, the softmax should give better performance in the condition with the greatest raw sum of weights for the correct affix. Specifically, if we observe the raw weights in Experiments 1 and 3, we see greater raw sum for the correct affix in the prefix condition than in the suffix condition (overall in Experiment 1, and for HF items in Experiment 3) – the exponential transformation will further augment this difference and result in a prefix advantage. To explore this possibility, I re-tested the original models from Experiment 1 (this was the original model, and not the one reported in the previous section with saliency manipulations) and Experiment 3 using the softmax instead of Luce's choice axiom. Figure 5.1 plots these results alongside the

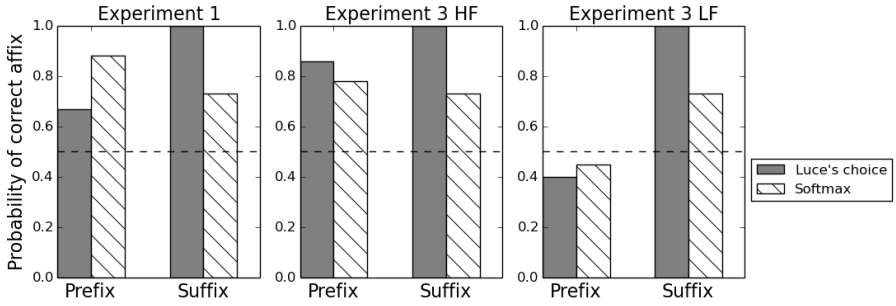


Figure 5.1: The probability of choosing the correct affix: sums of raw weights for the correct affix in each condition, normalized by Luce's choice axiom (grey bars) or by softmax function (white striped) in Experiments 1 (modelling human Experiment 2) and 3 (modelling human Experiments 4 and 5) for high-frequency (HF) and low-frequency (LF) learning. The dashed line is chance (0.5).

original results for comparison, and shows that the results of the softmax are more consistent with the findings observed in this thesis: a strong prefixing benefit in Experiment 2, a numerical prefixing benefit for HF items in Experiments 4-5, and a clear suffixing benefit for LF items in Experiments 4-5.

5.2.1 Discussion

The analyses presented so far in this chapter suggest that the strong prefixing advantage in Experiment 2, and a numerical prefixing advantage for HF items in Experiments 4 and 5 can be understood in terms of the importance of learning from negative evidence. Specifically, I argued that positive evidence alone may be sufficient for learning the correct generalizations in Study 1 and with HF items in Study 2. However, this difference in the extent to which different designs necessitate negative evidence may be obscured somewhat by the evaluation metric we used, the Luce's choice axiom (following Ramscar et al., 2010), which rewards the "unlearning" of the opposite affix – the lower the weights for the opposite affix, the greater the probability of choosing the correct affix. With this in mind, I considered an alternative metric – the softmax – which primarily captures the learning of the correct affix. This metric was more consistent with the findings from the experiments with humans: with the softmax, better performance in the prefix condition is predicted in Experiment 2 and with HF items in Experiments 4-5, whereas better performance is predicted in the suffix condition with LF items. This is remarkably in line with was observed with human learners, and suggests that the differences between the two conditions which were observed across the two studies can be re-interpreted as a function of the extent to which the learning task critically depends on negative evidence from prediction error. This was most clearly the case with LF items, less so with HF items (Study 2), and possibly not at all in Experiment 2 (Study 1) – as we move along this continuum, the difference between the conditions moves from a suffixing advantage to a prefixing advantage.

It is important to note that the neither evaluation metric is inherent to the Rescorla-Wagner model. In the original paper, Rescorla and Wagner (1972) analyse raw associative strenghts 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. With respect to this study, the Luce's choice axiom was chosen for consistency with previous work (Ramscar et al., 2010), but there was no theoretical reason to prefer it over the softmax, or the other way round. The simulations presented so far, however, demonstrate that, the choice of metric can be informed by our understanding of the learning task, that is, of the exact learning mechanisms that the study design evokes. Specifically, when learning critically depends on negative evidence (from prediction error and cue competition), such as in Experiment 3 for LF items, both metrics make the same prediction qualitatively: above-chance performance in the suffix condition and below-chance performance in the prefix condition. However, when learning is less critically dependent on "unlearning" the uninformative cues, which can only occur with negative evidence, the two metrics make different predictions: Luce's choice, which captures learning from negative evidence, predicts a suffixing advantage, whereas the softmax predicts better performance in the prefix condition. In cognitive modelling, deciding what metric to use is informed by the modeller's assumptions about the cognitive mechanisms that underlie how humans respond in the same task. For the work reported in this thesis, however, there were no pre-existing measures of learning which could be used, and for which there is consensus about what they are measuring and how robustly. A lot of the work in this thesis involved fine-tuning a training and testing paradigm for a novel theoretical framework, and therefore it was not possible to always have a strong theoretical understanding of what exactly participants are doing in each task to incorporate into the modelling. The simulations presented in this chapter are a valuable contribution to this issue. Finally, it is important to note that while it may be the case that prefix learning may be better than suffix learning in the context of learning from positive evidence alone, this is not explicitly predicted by the discriminative learning framework or the Rescorla-Wagner model. Therefore, in addition to determining through further experimentation whether prefix learning is indeed better for learning from positive evidence, future work must also provide a theoretical explanation as to *why* this may be the case.

5.3 Type-frequency effect in generalization in the suffix condition (Experiment 6)

Another finding from Study 2 concerns the type-frequency effect in generalization in the suffix condition in Experiment 6, whereas our theory predicts a type frequency effect in the prefix condition alone.

To investigate this, I re-visited the trajectory of learning in the model in Experiment 3. It can be seen in Figure 4.2, that in the early stages of learning, up until the 1000^{th} trial, the HF non-discriminating feature *shape1* has higher associative strength than the LF discriminating feature in both conditions. In the sufffix condition, the LF discriminating feature "catches up" with *shape1* around the 1000^{th} trial and takes over, eventually gaining more associative strength than *shape1* (recall that the critical difference between the conditions is that in the prefix condition, the LF discriminating features never "catch up" with the HF non-discriminating features, simply because of lower frequency). This means that,

were we to test the model in earlier stages of learning, a type-frequency effect – betteer learning of HF items than LF items – would be observed in the suffix condition, too.

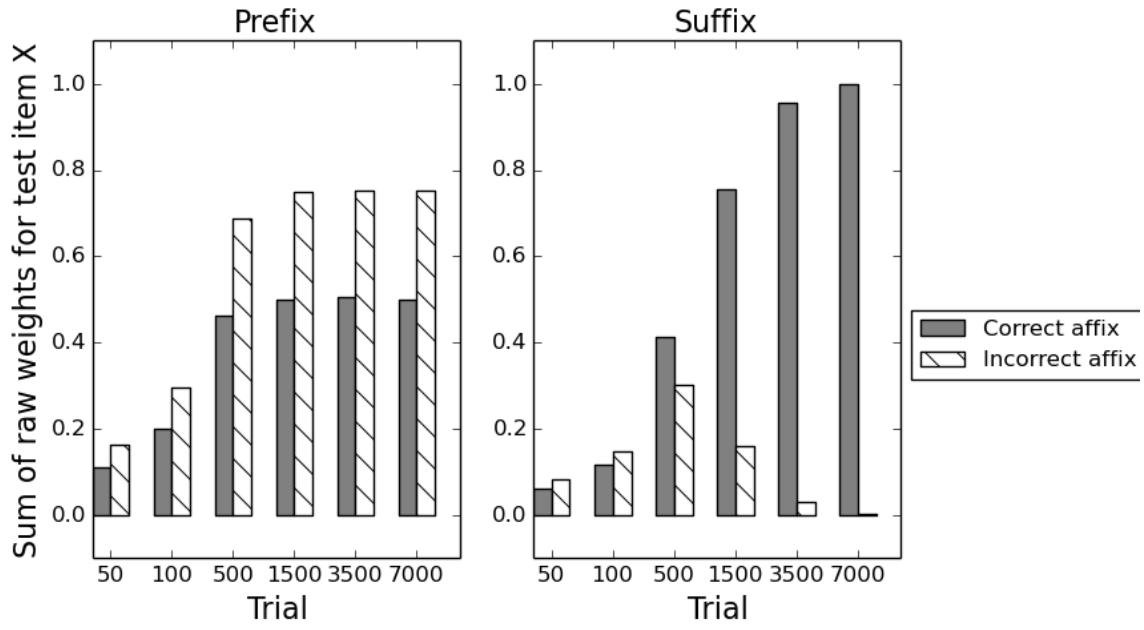
Therefore I tested the model reported in Experiment 3 at various stages of learning – Figure 5.2 shows the raw sum of weights for the same LF item (as the one used in Experiment 3) over time. In the prefix condition, the item is more strongly associated with the *incorrect* affix throughout learning. Critically, however, this also happens in the suffix condition in the early stages of learning, meaning that the suffix network was more likely to select the incorrect affix for the LF item in the early stages of learning, as I demonstrated with the 100^{th} trial (Figure 5.2). This suggests that in the earlier stages of learning, a type-frequency effect is observed in both conditions.

5.3.1 Discussion

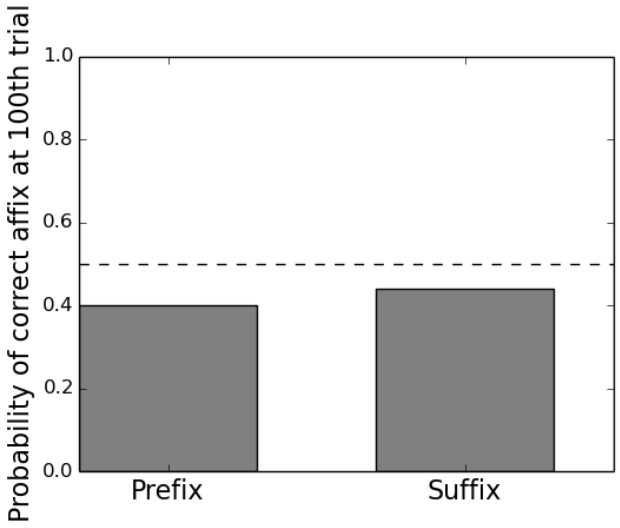
Re-examining the model from Experiment 3 suggested that, depending on where in the learning trajectory learners are tested, a type frequency effect may be observed in the suffix condition, too. Why could it be that this effect was observed in Experiment 6, and not in Experiments 4 and 5? One possibility is that Experiment 6 was harder than the other two experiments and thus had a slower learning trajectory. Specifically, while Experiment 6 contained the largest training set of all comparable experiments (32 items, compared to 16 in Experiment 4 and eight in Experiment 5), participants were given the same number of training trials as in the other two experiments. It is possible, therefore, that the more complex input set in Experiment 6 required longer exposure, and that learning in the suffix condition had not "reached asymptote" (to the extent that this applies to humans) at the point when learners were tested in Experiment 6. From this it is reasonable to predict that the type-frequency effect observed in Experiment 6 should disappear with more exposure.

While a re-analysis of the model presented here suggests that the type-frequency effect in the suffix condition was observed due to insufficient exposure, another experiment would need to be conducted to test this explanation. One possibility would be to use a training task with more exposure trials – a difficulty with this option is that such a large number of trials might overwhelm participants, and that they would stop attending to additional trials due to fatigue and/or boredom. An alternative would involve multiple training sessions, where the exposure phase is spread over two days, for example. Further exploration of the effect of the amount of exposure on generalization in the two affix conditions would not only be highly informative for the current work (where set size, but not the amount of exposure to the set, was manipulated), but to wider questions about generalization in the literature. I return to this in the General Discussion of the thesis (Chapter 7).

Note, however, that while there was a type-frequency effect in both conditions in Experiment 6, both conditions were still above chance with LF features (though this was not the case in the suffix condition once particiapnts who reported explicit awareness were removed). And while in the simulations this effect goes away in the suffix condition, it remains in the prefix condition, where the model is below-chance. In the experiment with humans, we did see a type-frequency effect, but both frequency-types were learned with above chance accuracy. This suggests that some other factor, which is not captured by



(a)



(b)

Figure 5.2: (a) Raw weights of the feature *shape1* for the correct affix *ma* (grey bars) and the incorrect affix (white stirped) in both conditions at different stages of learning. (b) Raw weights for the correct affixes normalized into probability of choosing the correct affix out of two options as per Luce's choice axiom at the 100^{th} trial (dashed-line is chance, 0.5).

these simulations, resulted in good LF learning in the prefix condition. One possibility is that in Experiment 6, there were sufficient LF items (four per affix) for the prefix condition to show stronger learning of the LF discriminating feature, that is, for LF items to emerge as a category. The simulations in this chapter modelled the learning of abstract features, rather than individual items, and cannot capture the effects of larger training sets directly. We have seen evidence across the behavioural experiments that this may affect learning in complex ways, and future work should aim to develop models of these effects.

To summarize, in this chapter I revisited previous modelling work following insights from experiments with human learners, particularly in order to better understand those findings which were inconsistent with the original predictions. The advantage of working with a precise model is that, when empirical findings do not confirm the hypothesis, the researcher can re-visit the model and generate new hypotheses. In this chapter, I demonstrated that the extent to which the learning task critically depends on learning from negative evidence (and the extent to which the evaluation metric captures this), has an effect on the direction of the effects of order in the Rescorla-Wagner model. This work also demonstrated that, as the model learns, the directions of order effects may change, and this is something that is critical to consider when using the model to formulate hypotheses about human learning. The next step in developing our theory would be to test the insights generated in this chapter with human learners (as the simulations presented here are used to generate novel hypotheses, rather than to ”explain away” unexpected findings in a post-hoc manner). The next chapter, Chapter 6, addresses another aspect of our theory where there were mixed results in Studies 1 and 2 – the predicted prefixing advantage in item-learning.

Chapter 6

Study 3: Experiments 7 – 9

6.1 Introduction

This chapter contains a series of experiments testing the hypothesis that prefixing enables better item-learning than suffixing. This prediction comes from the work of Dye et al. (2017, 2018), who demonstrated that prenominal articles and adjectives (similar to prefixes) in German and English reduce the entropy (uncertainty) of the upcoming noun. For example, when a speaker of French hears the feminine article la, only a subset of French nouns can follow – that is, only feminine nouns can occur, which significantly reduces the number of potential candidates, and therefore the entropy of the noun. This makes the noun easier to process, which, we predict, should make it easier for participants to learn the "meaning" of the noun (its corresponding picture) in the prefix condition compared to the suffixing condition. While the item learning advantage in the prefix condition was observed in Study 1, it was not observed in any of the experiments in Study 2. However, item-learning in Study 2 was generally low, and I suggested this may be due to the properties of the artificial language. Specifically, the language used in Study 2 was designed to encourage and boost the learning of predictive cues to affix occurrence across individual items, and was possibly not suitable for strong item-learning.

In this study, therefore, we design a different artificial language, in which there are no consistent cues to affix occurrence, and individual items are more perceptually distinguishable (see Section 5.2.1.2 for details) – this should encourage item-learning. With sufficiently learnable stimuli, we expect to observe the effects of reduced entropy on noun processing. The specific details of the design of each experiment are given below, but the overarching rationale is the following: participants in both conditions view a picture on-screen, after which a label is played – in the prefix condition, this is a prefix followed by a noun; in the suffix condition, this is a noun followed by a suffix. The number of individual nouns is varied across the experiments, however, in all experiments, there are two affixes, such that half of the nouns only ever occur with one affix, and the other half occur with the other.

Earlier in this thesis, we saw that in the prefix condition, information is distributed more evenly across the utterance than in the suffix condition – specifically, by the prefix smoothing the entropy over the affix-noun pair. From this, we predict that the noun will be easier to process in the prefix condition, which in turn will make it easier for participants

to learn the mapping with the co-occurring picture – as evidenced by better performance on the noun-picture test in this condition. While this is the key prediction in this chapter, two additional hypotheses are tested. In the suffix condition, once the noun is known to the participant, only one affix can occur - therefore the forward transitional probability of the affix given the noun is 1 (in the prefix condition, on the other hand, multiple nouns can occur following the prefix). This raises the possibility that participants in the suffix condition will be better at learning noun-suffix mappings (but that this will not be helpful in the learning of noun-picture mappings), and we test this hypothesis, too. The learning of the noun-affix pairs is tested by presenting participants with a label in which a trained noun either occurs with the correct affix or with the opposite affix, and participants are asked to indicate whether the label is correct (more detail in Section 6.2.1.3). This test raises an intriguing possibility – if the learning of the noun-picture pairs in the prefix condition is facilitated by the affix, it may be possible that participants in this condition perform more poorly in the noun-picture test with those nouns which occurred with the incorrect affix in the noun-affix test. To test this, we repeat the noun-picture test once more after the noun-affix test. While we note that the power to detect this effect depends on the number of trained items (as only half of the items occur with the incorrect affix), testing this hypothesis is methodologically relevant, too – in artificial language learning experiments, "learning" and "testing" are often separated, and there is little acknowledgement of the fact that learning also occurs during testing, and that test-trials may interfere with what is learned during exposure (Siegelman et al., 2017). Therefore, this effect would not only further corroborate our theory, but it would also speak to broader methodological concerns.

Note that the predictions made above presume a perfect situation in which participants learn the probability distribution perfectly and generate their predictions according to the distribution. The observed effect is likely going to be weaker and noisier – I return to this in the General Discussion.

6.1.1 Outline of the chapter

The chapter begins with a pilot experiment (Experiment 7a). Given that this paradigm is novel in the context of this thesis, and there was no sufficiently similar previous work on which to estimate the number of trained items and exposures per item, we were reluctant to commit to testing participants until there was evidence for/against the key hypothesis (i.e. in case the input set is not learnable to a sufficient degree with the current amount of exposure). Therefore, we inspected the data at 20 participants per condition. This showed above-chance learning in both conditions, but ambiguous evidence for the key hypothesis. Given that we had no pre-registered stopping rule, we did not continue adding data to this sample. Instead, we pre-registered a replication experiment (Experiment 7b) with a stopping rule and H1 values derived from Experiment 7a. In a follow-up experiment, we modified the paradigm used in Experiments 7a and 7b (namely using fewer trained items), and here the same approach was used – a pilot experiment with 20 participants per condition (Experiment 8a), followed by a pre-registered replication (Experiment 8b). Finally, in Experiment 9, the role of prefixes in uncertainty reduction is addressed using a

paradigm which allowed us to track participants' learning during the exposure phase – the cross-situational learning paradigm.

Pre-registered hypotheses for Experiments 7 and 8 can be found here: https://osf.io/gbzmc/ and all the data and R analyses scripts in this chapter can be found here: https://osf.io/tqp5a/

6.2 Experiment 7a

6.2.1 Method

6.2.1.1 Participants

Fourty participants (20 per condition) were recruited through Prolific Academic. All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for their participation.

6.2.1.2 Stimuli

Audio stimuli consisted of 16 two-syllable nouns and four affixes. Nouns were chosen from previous similar studies, and the affixes were the same ones used in previous work in this thesis. One Category A affix and one Category B affix was chosen randomly on a participant-by-participant basis, from two Category A affixes (pe and ge) and two Category B affixes (da and ma). The audio stimuli were synthesized using the MBROLA speech synthesizer with a male British English voice. Visual stimuli were pictures of novel objects (Horst & Hout, 2016). Nouns were assigned to pictures and affixes randomly on a participant-by-participant basis. See Figure 6.1 for a sample training set of pictures and labels.

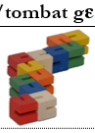
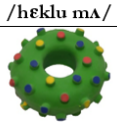
Category 1: gɛ		Category 2: mʌ	
			
/vitord gɛ/	/tʃi:lagɛ/	/panjol mʌ/	/ɛtkot mʌ/
			
/peɪlig gɛ/	/slagum gɛ/	/di:tʃə mʌ/	/wazɪl mʌ/
			
/ku:mo gɛ/	/tombat gɛ/	/slindot mʌ/	/hɛklu mʌ/
			
/mi:pə gɛ/	/hɜ:dip gɛ/	/polip mʌ/	/pɪkru mʌ/

Figure 6.1: Sample training set. Images were taken from NOUN Database (Horst & Hout, 2016)

6.2.1.3 Procedure

Training. The 15-minute training consisted of 2 blocks of 80 trials (160 trials in total, 10 exposures per item) presented in random order. In both conditions, the context image (a man pointing) and the target image (the novel item) appeared on-screen for 1000ms. After this, the images remained on-screen, and the carrier phrase, *Os ferpel en*, was played. In the prefix condition, this was followed by the affix and the noun, and in the suffix condition, the noun was played first, and then the affix. Finally, the two pictures remained on-screen for an additional 450ms. This was followed by a blank screen displayed for 1000ms, after which the next trial was played. Importantly, the duration of a single trial, and the amount of time that the picture was on-screen were identical in both conditions, the only difference was the order in which nouns and affixes were played (Figure 6.2). The training procedure was designed to closely match Arnon and Ramscar (2012).

Noun-picture test. Participants' item-knowledge (recall of noun-picture pairs) was tested by showing participants all training items in a grid and playing a label from training (carrier phrase + affix + noun in the prefix condition or carrier phrase + noun + affix in the suffix condition). Participants' task was to click on the item they thought matched the label. The noun-picture test was repeated at the end of the experiment. The pictures were arranged in a grid in the same order as first time, but the order of individual trials was shuffled. Each trained item was tested twice (once in each round of the test), giving a total of 32 trials.

Noun-affix test. Participants' knowledge of the association between individual nouns and affixes was tested in a grammaticality judgment test. Participants were played a label and asked to click the "correct" button if they thought the label was correct, and the "incorrect" button if they thought the label was incorrect. On half of the trials, the noun was played with the correct affix, and on the other half of the trials, the noun was played with the *opposite* affix (assignment to correct/incorrect was done randomly on a participant-by-participant basis). Each trained noun was tested once, giving a total of 16 trials.

While the training was timed, participants were allowed to take as long as they wished to complete the testing.

6.2.2 Results

6.2.2.1 Noun-picture test

The data are shown in Figure 6.3 and the statistics are in Table 6.1. We predicted that the prefix condition would be better than the suffix condition. The evidence for this prediction was ambiguous (prefix: $M = 43.44\%$, $SD = 38.94\%$, suffix: $M = 36.09\%$, $SD = 32.59\%$).

The effect of incorrect noun-affix bigrams on performance in the second noun-picture test. A secondary hypothesis was that exposure to the affix test would differently affect performance on the repeated noun-picture test for the prefix and suffix conditions. Specifically, we expect that in the prefix condition there would be a strong effect of hearing the word with the *wrong* affix, so that performance would be worse with

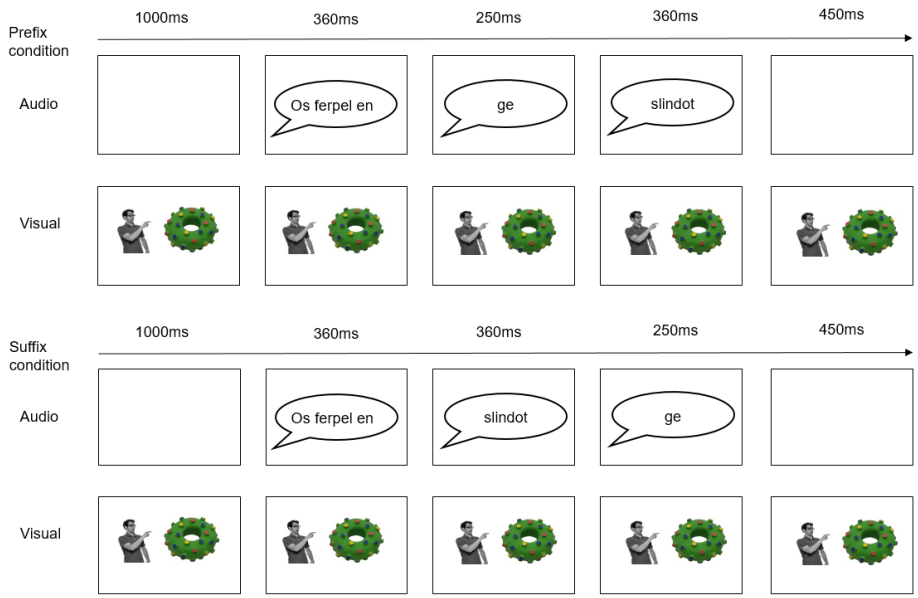


Figure 6.2: Experiment 6a: Schematic representation of a single training trial in the prefix (top) and the suffix condition (bottom). Note: labels were only auditory, and were not presented orthographically.

nouns which had been used in the "incorrect" test items in the noun-affix test. This effect is expected to be less strong, or absent, in the suffix condition. We found that this was the case with substantial evidence for the affix-by-correct-bigram interaction. We followed up this analysis by testing for an effect of correct-bigram in each affix condition, and found evidence for no interaction in the suffix condition (correct: $M = 31.87\%$, $SD = 34.05\%$; incorrect: $M = 37.5\%$, $SD = 32.44\%$), whereas in the prefix condition, the evidence was ambiguous (correct: $M = 45\%$, $SD = 42.03\%$; incorrect: $M = 40.62\%$, $SD = 40.73\%$). Note that in the prefix condition, there was no suitable value to inform the H1, and so the result is interpreted using the robustness region analysis, which suggested ambiguous evidence for the majority of the plausible values tested.

The effect of noun-affix learning on noun-picture learning. We explored whether learning the noun-affix mappings had an effect on the performance in the noun-picture test. Recall that our prediction is that noun-picture learning is facilitated by the prefix. This is because upon hearing the prefix, only a half of the nouns are likely, meaning that the entropy of the noun is reduced; this eases the processing of the noun. However, this facilitatory effect depends on participants learning the conditional probability of nouns given the prefix. Therefore, we might see that those participants who are better at the noun-affix test will be better at the noun-picture test in the prefix condition more so than in the suffix condition. This would be evidenced by a noun-affix-learning by affix condition interaction.

To look at this, participants were grouped based on whether they scored above-chance on the noun-affix test (indication of learning), and this was used as a binary (learning/no-learning) between-participants predictor of accuracy on the noun-picture test. In both conditions, half of participants showed noun-affix learning. On average, participants who showed noun-affix learning performed better on the noun-picture test than those who did not show noun-affix learning both in the prefix condition (learning: $M = 51.56\%$, $SD =$

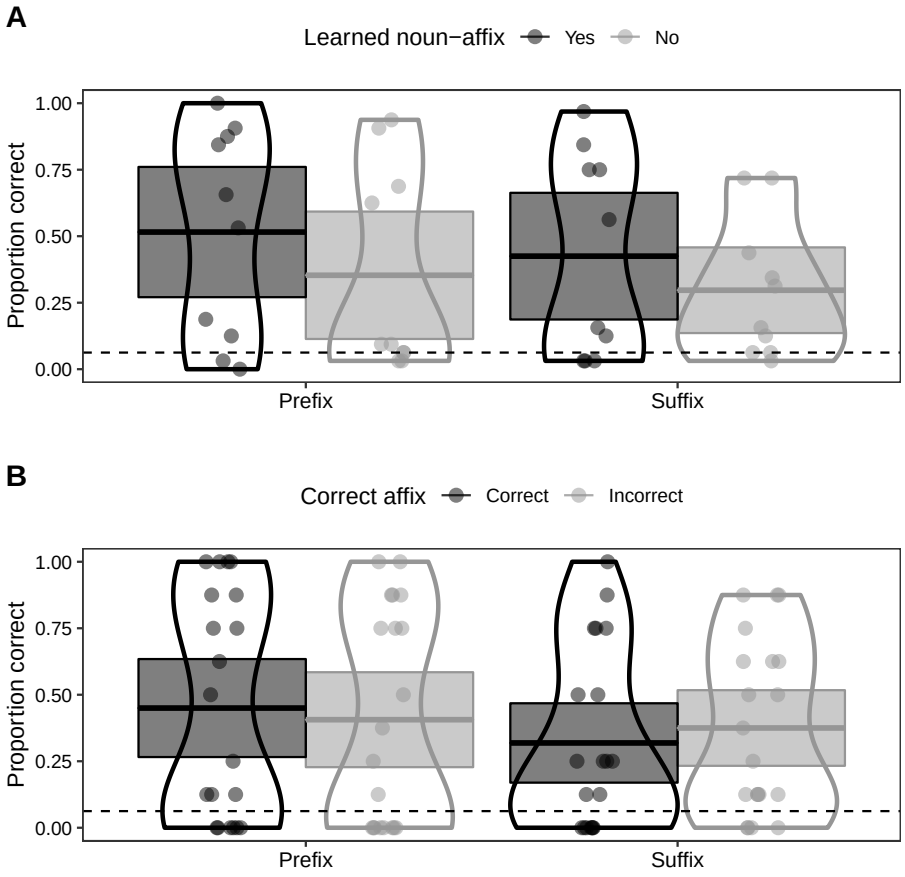


Figure 6.3: Experiment 7a: Panel A: average accuracy on the two noun-picture tests (combined) of participants who showed noun-affix learning (black) and those who did not (grey). Panel B: average accuracy on the second noun-picture test for those nouns which occurred with the correct affix in the noun-affix test (black) and those which occurred with the incorrect affix (grey). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance (0.0625 or 1/16).

39.51%; no learning: $M = 35.31\%$, $SD = 38.64\%$) and in the suffix condition (learning: $M = 42.5\%$, $SD = 38.42\%$; no learning: $M = 29.69\%$, $SD = 25.99\%$), however, the evidence for an main effect of noun-affix learning was ambiguous, as was evidence for a by-affix interaction.

6.2.2.2 Noun-affix test

The data are shown in Figure 6.4 and the inferential statistics in Table 6.2. The evidence for the predicted benefit of suffixing was ambiguous. Fitting separate intercepts for each affix condition showed substantial evidence for above-chance performance in the suffix condition ($M = 56.25\%$, $SD = 11.65\%$), whereas in the prefix condition the evidence was ambiguous ($M = 53.44\%$, $SD = 13.22\%$).

Table 6.1: Experiment 7a: Noun-Picture Test Statistics.

Hypothesis	Contrast in lme	Mean difference	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect of affix	0.41	0.72	1.97[1]	0.55	[0 : 3.49]	.57
Learning better than no learning	Main effect of bigram learning	0.85	0.72	1.97[1]	1.11	[0 : >7.3]	.234
Greater bigram learning effect in Prefix	Affix by bigram learning int.	0.46	1.44	0.98[2]	0.96	[0: 5.47]	.749
Greater bigram test effect in Prefix	Affix by correct bigram int.	0.87	0.47	0.87[2]	3.46	[0.4 : 1.44]	.057
Breaking down the affix-by-correct-bigram interaction by affix-condition							
Correct bigram better than incorrect	Correct bigram effect in Prefix	0.43	0.33	–	–	ambig: [0: 4.1], H0: [4.12 : ∞]	.196
	Correct bigram effect in Suffix	-0.37	0.30	0.33[3]	0.37	[0.35 : ∞]	.155

[1]Intercept from the same lme (method B)
[2]Twice the intercept estimate from the same lme (method C)
[3]Correct-bigram effect in Prefix condition (method A)

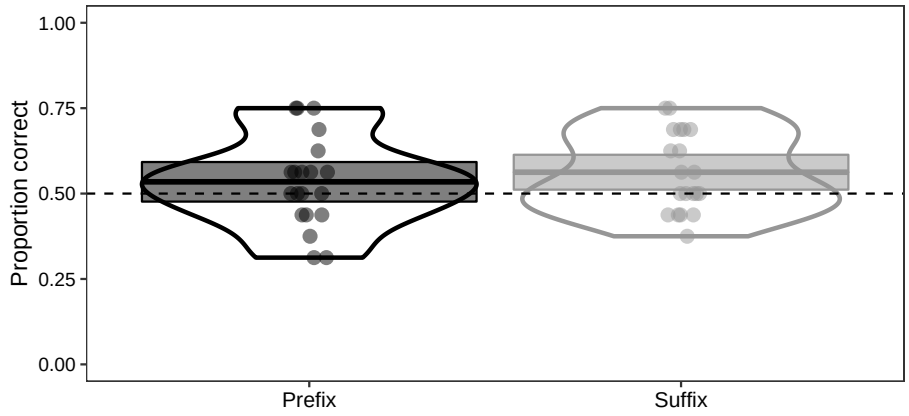


Figure 6.4: Experiment 7a: Proportion of correct responses on the Noun-affix test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line is chance-level performance,

Table 6.2: Experiment 7a: Noun-Affix Test Statistics.

Hypothesis	Contrast in lme	Mean difference	SE	H1	B	Robustness region	p
Suffix better than Prefix	Main effect of affix	0.11	0.16	0.20[1]	1.05	[0 : 0.91]	.47
Prefix above chance	Prefix Intercept	0.14	0.11	0.25[2]	1.34	[0 : 1.27]	.22
Suffix above chance	Suffix Intercept	0.25	0.11	0.14[3]	5.31	[0 : 0.83]	.03

[1]Intercept from the same lme (method B)
[2]Suffix Intercept (method A)
[3]Prefix Intercept (method A)

6.2.3 Discussion

Experiment 7a showed ambiguous evidence for the key hypothesis – better noun-picture learning in the prefix condition compared to the suffix condition. Recall that participants completed the noun-picture test twice – once immediately after training, and once again after the noun-affix test. In the noun-affix test, half of the trained nouns were presented with the wrong affix. We hypothesised that, in the second noun-picture test, participants in the prefix condition would perform more poorly with the nouns which appeared with the wrong affix in the noun-affix test; participants in the suffix condition, on the other hand, should not be affected by this, as they rely on the affix to identify the corresponding picture less than participants in the prefix condition do. There was substantial evidence for this interaction. Follow-up analyses showed that the evidence for better performance with the nouns heard with the correct affix than with nouns heard with the incorrect affix was ambiguous in the prefix condition; in the suffix condition, there was evidence for no difference between correct and incorrect nouns. We interpret this pattern of results as moderate evidence for the interaction. The interpretation would be stronger had there been evidence for an effect in the prefix condition; however, the statistical power of this analysis is weaker due to less data (the analysis of the interaction combines data from both conditions).

Between the two noun-picture tests, participants completed a test designed to tap their learning of the noun-affix bigrams. Here, nouns and affixes from training were played, such that on half of the trials, the noun occurred with the wrong affix. Participants were asked to indicate via button press whether the "sound" was correct or incorrect. We predicted better learning in the suffix condition, but found ambiguous evidence. Analysing the two affix conditions separately showed evidence for above-chance performance in the suffix condition, whereas in the prefix condition this evidence was ambiguous.

To summarize, the ambiguous result with respect to a prefixing benefit in noun-picture learning means that we cannot draw conclusions as to our key hypothesis. Given this, it is less clear why there was substantial evidence for the correct-bigram-by-affix interaction. Therefore this pattern of results requires replicating, and this is what we do in Experiment 7b.

6.3 Experiment 7b

6.3.1 Method

6.3.1.1 Participants

One-hundred thirty-six participants (69 in the Prefix and 67 in the Suffix condition) were recruited.

Note that our pre-registered plan was to start with 20 participants per condition, inspect the data, and continue testing 10 participants per condition until there was evidence for the hypothesis or for the null (https://osf.io/m7jvc). The evidence for the key hypothesis (prefixing advantage in noun-picture learning) was still ambiguous at 168 participants

(planned sample was 170). Combining the data with Experiment 7a, however, showed evidence for the null, therefore, given the time and resource constraints, we stopped testing at this point.

6.3.1.2 Stimuli and Procedure

Same as Experiment 7a.

6.3.2 Results

6.3.2.1 Noun-picture test

The data are shown in Figure 6.5 and the statistics are in Table 6.3. The evidence for the predicted benefit of prefixing in item-learning ambiguous, with the means being in the opposite direction (suffix: $M = 46.46\%$, $SD = 34.14\%$; prefix: $M = 45.39\%$, $SD = 34.62\%$). For combined data, however, there was evidence for the null.

The effect of incorrect noun-affix bigrams on performance in the second noun-picture test. As to the effect of incorrect bigrams on performance on the second noun-picture test, with Experiment 7b data alone, there was evidence for no difference in the effect of incorrect bigrams in the prefix condition compared to the suffix condition (prefix correct: $M = 45.47\%$, $SD = 35.35\%$); prefix incorrect: $M = 43.66\%$, $SD = 37.6\%$; suffix correct: $M = 46.83\%$, $SD = 36.3\%$); suffix incorrect: $M = 46.46\%$, $SD = 35.81\%$). This was followed up by testing for an effect of correct-affix in each affix condition, and we found ambiguous evidence both in the suffix condition (correct: $M = 46.83\%$, $SD = 36.3\%$; incorrect: $M = 46.46\%$, $SD = 35.81\%$), and this was the case with joined data, too. In the prefix condition (correct: $M = 45.47\%$, $SD = 35.35\%$; incorrect: $M = 43.66\%$, $SD = 37.6\%$), there was no clear exact value to base the H1 on, however, we believe that a plausible value falls in the range of values for which the evidence is ambiguous (same for joined data).

The effect of noun-affix learning on noun-picture learning. In Experiment 7b, in the prefix condition, 59% of participants showed noun-affix learning, whereas in the suffix condition 69% of participants showed learning. In the suffix condition, on average, participants who showed noun-affix learning performed better on the noun-picture test than those participants who did not show noun-affix learning (learning: $M = 49.81\%$, $SD = 32.46\%$; no learning: $M = 43.2\%$, $SD = 35.87\%$). In the prefix condition the means were in the opposite direction – – participants who showed no learning of noun-affix bigrams ware on average better at noun-picture learning than those participants who learned the bigrams (learning: $M = 41.7\%$, $SD = 35.64\%$; no learning: $M = 48.06\%$, $SD = 34.06\%$). There was evidence for no affix-by-bigram-learning interaction (for Experiment 7b as well as for combined data). There was also evidence that overall, participants who showed noun-affix learning did not perform better than those who *did not* show learning (both for Experiment 7b and for combined data).

6.3.2.2 Noun-affix test

The data are shown in Figure 6.6 and inferential statistics in Table 6.4.

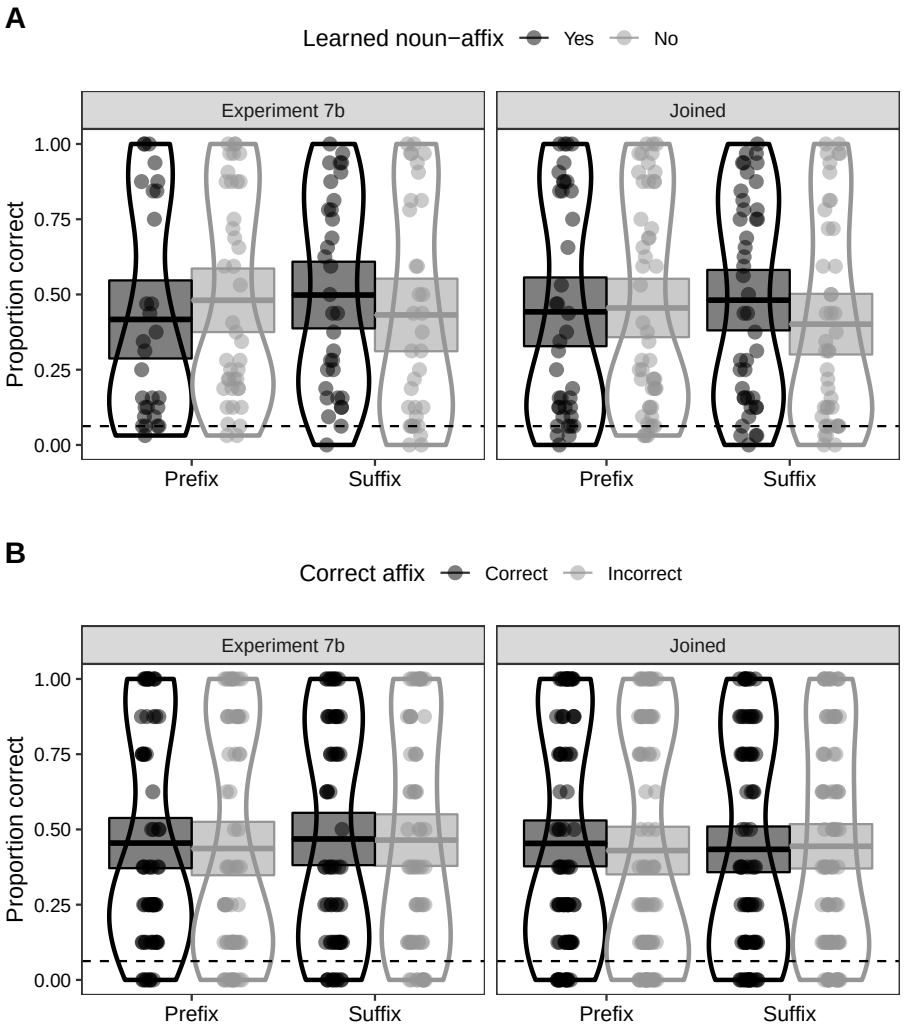


Figure 6.5: Experiment 7b: Panel A: average accuracy on the two noun-picture tests (combined) of participants who showed noun-affix learning (black) and those who did not (grey). Panel B: average accuracy on the second noun-picture test for those nouns which occurred with the correct affix in the noun-affix test (black) and those which occurred with the incorrect affix (grey). Data are from Experiment 7b (left) or combined with Experiment 7a (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance (0.0625 or 1/16).

Table 6.3: Experiment 7b: Noun-Picture Test Statistics.

Hypothesis	Contrast in lme	Data	Mean diff	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect affix	7b	0.04	0.39	0.40[1]	0.74	[0 : 1.21]	.927
		Joined	0.14	0.34	2.43[2]	0.20	[1.44 : ∞]	.786
Learning better than no learning	Main effect bigram learning	7b	-0.01	0.39	2.56[2]	0.14	[2.03: ∞]	.649
		Joined	0.18	0.34	2.43[2]	0.23	[0 : 3.28]	.315
Greater bigram lear. effect in Prefix	Affix by bigram learn. int	7b	-0.81	0.77	1.28[3]	0.28	[0 : 1.39]	.368
		Joined	-0.55	0.69	1.21[3]	0.31	[0: 1.37]	.517
Greater bigram test effect in Prefix	Affix by correct bigram int	7b	0.11	0.23	0.87[4]	0.38	[1.29 : ∞]	.591
		Joined	0.27	0.20	1.18[3]	0.71	[0 : 3.64]	.147
Breaking down the affix-by-correct-bigram interaction by affix condition								
Correct bigram better than incorrect	Correct bigram effect in Prefix	7b	0.12	0.16	–	–	ambig: [0 : 0.94], H0: [0.95:]	.578
		Joined	0.19	0.14	–	–	ambig: [0 : 1.92], H0: [1.93:∞]	.262
	Correct bigram effect Suffix	7b	0.02	0.16	0.12[5]	0.84	[0 : 0.45]	.842
		Joined	-0.07	0.14	0.19[5]	0.43	[0 : 0.23]	.353

[1]Based on an estimate from 7a (method A)
[2]Intercept from the same lme (method B)
[3]Twice the intercept from the same lme (method C)
[4]Same effect from Experiment 7a (method A)
[5]Effect of correct-bigram in the Prefix condition (method A)

There was evidence for no predicted benefit of suffixing with Experiment 6b data alone, but this was ambiguous with combined data. However, unlike in Experiment 6a, there was now evidence for learning both in the prefix condition ($M = 53.37\%$, $SD = 10.7\%$) and in the suffix condition ($M = 54.56\%$, $SD = 12.04\%$), and this was also true for combined data.

6.3.3 Discussion

Experiment 7b was a pre-registered replication of Experiment 7a. The same hypotheses were tested on the data from Experiment 7b, as well as for data from Experiments 7a and 7b combined.

Table 6.4: Experiment 7b: Noun-Affix Test Statistics.

Hypothesis	Contrast in lme	Data	Mean difference	SE	H1	B	Robustness region	p
Suffix better than Prefix	Main effect affix	7b	0.03	0.09[1]	1.33	0.09	[0.49: ∞]	.588
		Joined	0.05	0.07	0.16[2]	0.71	[0 : 0.57]	.399
Cell by cell comparison to chance								
Prefix above chance	Prefix Intercept	7b	0.13	0.06	0.16[3]	6.14	[0: 0.19]	.063
		Joined	0.14	0.05	0.18[3]	11.07	[0: 0.45]	.026
Suffix above chance	Suffix intercept	7b	0.16	0.06	0.13[4]	14.19	[0: 1.43]	.009
		Joined	0.18	0.05	0.14[4]	94.10	[0: >4.59]	.001

[1]Based on an estimate from 7a (method A)
[2]Intercept from the same lme (method B)
[3]Suffix Intercept from same lme (method A)
[4]Prefix Intercept from same lme (method A)

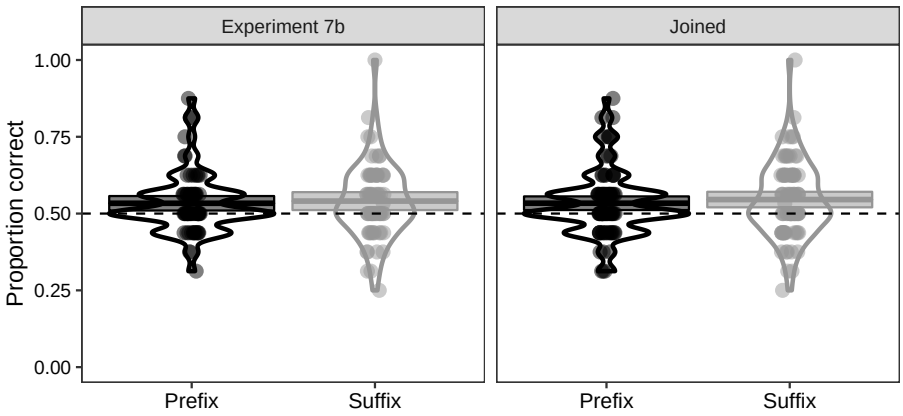


Figure 6.6: Performance on the Noun-affix test in Experiment 7b (left), and combined with Experiment 7a (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. Dashed line is chance-level performance.

Starting with the key hypothesis – a prefix advantage in noun-picture learning – as in Experiment 7a, the evidence was ambiguous in Experiment 7b. However, combined data suggested evidence for no advantage. We make inferences on the basis of the combined data set, as it is more robust (more data means more evidence, and Bayes Factors are a measure of evidence), and conclude that Experiment 7 shows that learning the prefix-noun mappings did not help participants learn the noun-picture mappings better than participants who were exposed to suffixes.

With respect to the secondary hypothesis, whereby noun-picture learning is affected more strongly by exposure to incorrect noun-affix bigrams in the prefix condition than in the suffix condition (as evidenced by an affix-by-correct-bigram interaction), we found evidence for the interaction in Experiment 7a, but this did not replicate in Experiment 7b, where the evidence was ambiguous, and the same was true for combined data. Breaking down by affix condition also showed ambiguous evidence for an effect of noun-affix test on the performance on the noun-picture test in both conditions. In addition, across the two experiments, there was evidence that there was no difference in noun-picture learning between those participants who showed noun-affix learning and those who did not, and evidence that this difference was not different between the two affix conditions.

Turning to the noun-affix test, contrary to prediction, we found evidence for no suffixing advantage with Experiment 7b data alone, however, combined data suggested ambiguous evidence. Comparing each condition to chance-level performance showed evidence for above-chance performance in both conditions, both for Experiment 7b data alone, and for combined data.

To summarize, Experiments 7a and 7b together showed that whether an affix is heard before or after a noun does not affect the learning of the association between that noun and a corresponding picture. This is contrary to our predictions that higher entropy of the noun in the suffix condition compared to the prefix condition would make the noun harder to process, and this more cognitively taxing processing would in turn make it harder to

remember which noun occurred with which picture in the suffixing condition. However, there was no evidence of this across the two experiments, and overall no clear evidence that the learning of the noun-affix bigrams affects the learning of the noun-picture mappings. It is important to note though, when we do see noun-affix learning, it is rather low – no greater than 55% on average. Given that there was no strong noun-affix learning, it is then a little less surprising that there was no evidence that this learning affected (that is, facilitated – in the prefix condition) noun-picture learning. It is possible that, in order to see the predicted effect of smoothing the entropy, the learning of the affix-noun bigrams must be stronger in the first place.

The issue of low noun-affix learning is addressed in Experiment 8a in which the existing paradigm is modified to boost affix-noun learning: a two-day procedure is designed in which, on Day 1, participants are only exposed to noun-affix bigrams without corresponding pictures. On Day 2, participants repeat the exposure from Day 1, and then receive another round of exposure this time with pictures included. Therefore, we expect that participants in both conditions should come to the task of learning noun-picture mappings with strong learning of affix-noun pairs, and specifically, that this would lead to better noun-picture learning in the prefix condition than in the suffix condition. In addition to using a two-day procedure, where the intention is to take advantage of consolidating effects of sleep on vocabulary learning (Brown, Weighall, Henderson, & Gaskell, 2012; Henderson, Weighall, Brown, & Gaskell, 2012), the number of items was decreased by half, so that in Experiment 8 participants were only exposed to eight items in total.

6.4 Experiment 8a

6.4.1 Method

6.4.1.1 Participants

Forty participants (20 per condition) were recruited through Prolific Academic[1]. All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for their participation.

6.4.1.2 Stimuli

Same as Experiment 7, except that only half of the nouns and pictures were used (see Figure 6.7).

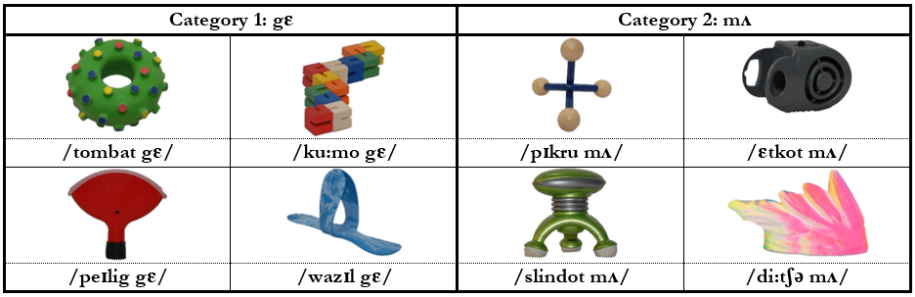


Figure 6.7: Sample training set. Images were taken from NOUN Database (Horst & Hout, 2016).

6.4.1.3 Procedure

Training. A two-day procedure was used. On Day 1, participants were instructed that they were going to hear sounds from an "alien language", and that their task was to focus and listen. Participants were then played eight unique noun-affix bigrams, 40 times each – this resulted in 320 trials distributed over 4 blocks (80 trials per block). After each block, participants were offered a short break, and the entire exposure session took approximately 15 minutes. At the end of the fourth block, participants were reminded that they would be invited to return to the study the following day. Approximately 24 hours after they completed Day 1, participants were sent an email containing a web-link to Day 2 of the study, and they had until the end of that day to complete the study. On Day 2, the training from Day 1 was repeated, and this was followed by another exposure session, where pictures were shown along with the audio stimuli. The timing and the presentation of the stimuli was identical to Experiment 7. As in Experiment 7, there were 160 trials in total, meaning 20 exposures per item.

Testing. Same as Experiment 7, except that there were eight trials on each of the noun-picture tests (due to there being eight trained items), as well as eight trials on the noun-affix test (four correct and four incorrect).

[1]This study had a high drop-out rate. An additional 20 participants were tested but they did not return to complete Day 2

Table 6.5: Experiment 8a: Noun-Picture Test Statistics.

Hypothesis	Contrast in lme	Mean difference	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect affix	1.91	1.01	2.87[1]	3.09	[0.95 : 3.03]	.06
Learning bigrams > no learning	Main effect bgrm learning	1.91	1.02	5.52[2]	1.89	[2.9 : >6.54]	.062
Greater bigram learning effect in Prefix	Affix by bgrm learning int.	-0.57	2.00	2.76[3]	0.49	[0.1 : 4.5]	.777
Greater noun-affix test effect in Prefix	Affix by correct bgrm int.	-0.21	0.82	0.52[3]	0.76	[0.1 : 1.8]	.802
Breaking down the affix-by-correct-bigram interaction by affix-condition							
Correct bigram better than incorrect	Correct bgrm effect in Prefix	0.53	0.51	0.85[4]	1.25	[0 : 4.46]	.296
	Correct bgrm effect in Suffix	0.85	0.46	0.53[5]	3.18	[0.47 : 1.22]	.065

[1]Based on an estimate from 7b (method A)
[2]Intercept from the same lme (method B)
[3]Twice the Intercept from same lme (method C)
[4]Suffix Intercept from same lme (method A)
[5]Prefix Intercept from same lme (method A)

6.4.2 Results

6.4.2.1 Noun-picture test

The data are shown in Figure 6.8 and the statistics are in Table 6.5. As predicted, and unlike in Experiment 7, there was evidence for a prefixing benefit in noun-picture learning (prefix: $M = 90.94\%$, $SD = 21.22\%$, suffix: $M = 76.56\%$, $SD = 28.09\%$).

The effect of incorrect noun-affix bigrams on performance in the second noun-picture test. There was ambiguous evidence for a greater effect of the noun-affix test on performance on the noun-picture test in the prefix compared to the suffix condition. Breaking down by affix-condition, however, showed substantial evidence for an effect of correct bigrams in the suffix condition, which is contrary to the prediction (correct: $M = 81.25\%$, $SD = 30.21\%$; incorrect: $M = 72.5\%$, $SD = 32.34\%$), whereas in the prefix condition the evidence was ambiguous (correct: $M = 91.25\%$, $SD = 23.33\%$; incorrect: $M = 87.5\%$, $SD = 22.21\%$).

The effect of noun-affix learning on noun-picture learning. The evidence that those participants who showed noun-affix learning were better at the noun-picture test than those who did not was ambiguous. The evidence was also ambiguous for a bigger difference between learning and no learning in the prefix condition (learning: $M = 94.89\%$, $SD = 13.06\%$; no learning: $M = 86.11\%$, $SD = 28.43\%$) compared to suffix condition (learning: $M = 85.62\%$, $SD = 22.45\%$; no learning: $M = 67.5\%$, $SD = 31.29\%$).

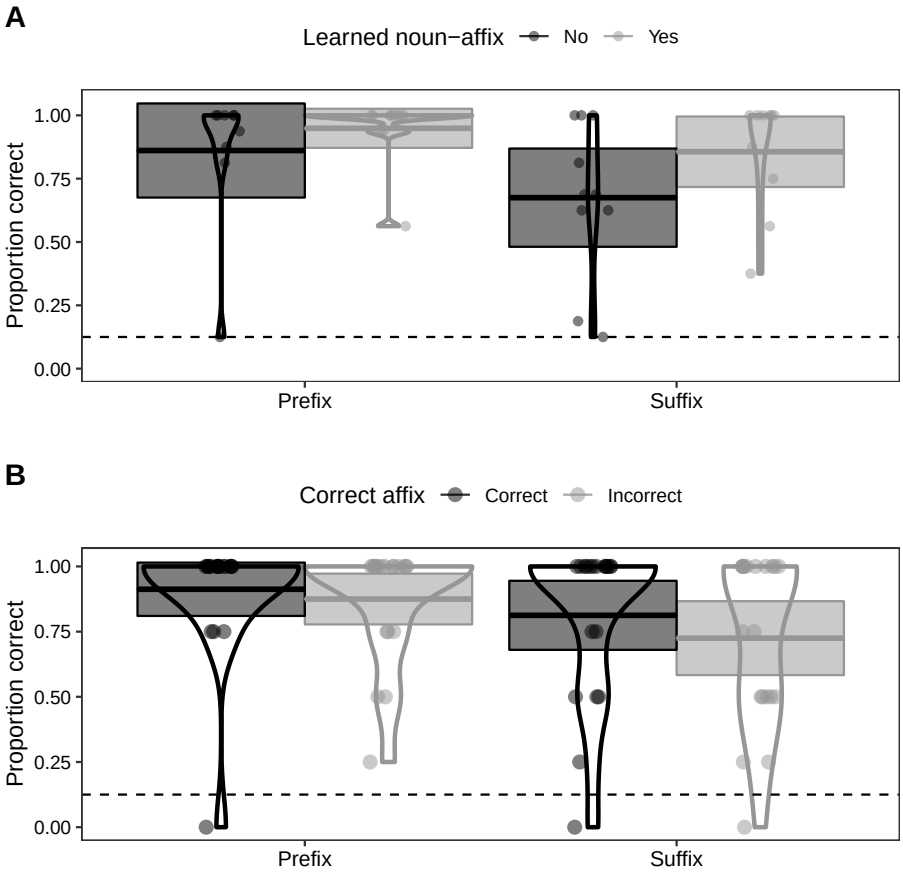


Figure 6.8: Experiment 8a: Panel A: average accuracy on the two noun-picture tests (combined) of participants who showed noun-affix learning (black) and those who did not (grey). Panel B: average accuracy on the second noun-picture test for those nouns which occurred with the correct affix in the noun-affix test (black) and those which occurred with the incorrect affix (grey). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance (0.125 or 1/8).

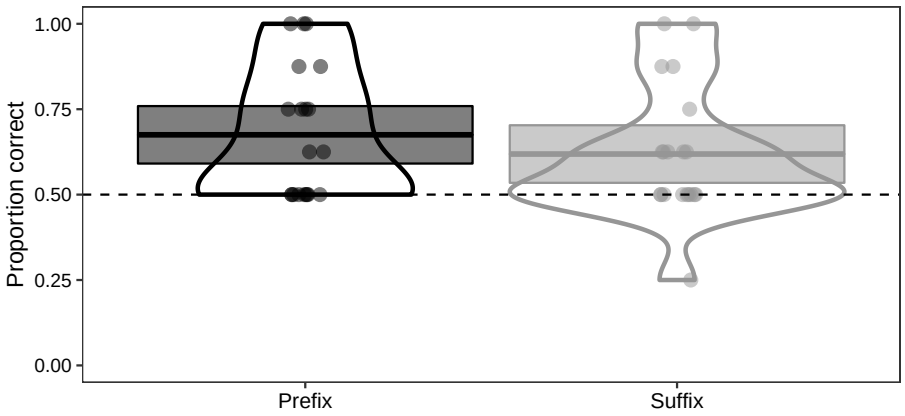


Figure 6.9: Experiment 8a: Proportion of correct responses on the Noun-affix test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line is chance-level performance.

Table 6.6: Experiment 8a: Noun-Affix Test Statistics.

Hypothesis	**Contrast in lme**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Suffix better than Prefix	Main effect of affix	-0.26	0.27	1.33[1]	0.11	[0.39 : ∞]	.475
Breaking down by affix-condition							
Prefix above chance	Prefix intercept	0.76	0.19	0.50[2]	542.66	[0.05 : >4.59]	<.001
Suffix above chance	Suffix intercept	0.50	0.19	0.76[3]	13.79	[0.09 : 4.37]	.007

[1]Based on an estimate from 7b (method A)
[2]Suffix Intercept from the same lme (method B)
[3]Prefix Intercept from same lme (method A)

6.4.2.2 Noun-affix test

The data are shown in Figure 6.9, and the inferential statistics in Table 6.6. There was evidence for no difference between conditions in the learning of noun-affix bigrams (prefix: $M = 67.5\%$, $SD = 19.19\%$, suffix: $M = 61.88\%$, $SD = 19.23\%$), and evidence for learning (above-chance performance) in both affix conditions.

6.4.3 Discussion

The aim of Experiment 8a was to induce better learning of noun-picture mappings in the prefix condition than in the suffix condition by improving the learning of the noun-affix bigrams. We aimed to improve noun-affix learning by reducing the number of trained items by half, and by doubling the exposure in a two-day paradigm, which allowed for a 24-hour consolidation period.

Modifying the paradigm did improve noun-affix learning. In Experiment 7, noun-affix learning was relatively low and there was no evidence that prefixing benefit in noun-picture

learning (the evidence was in support of the null). In Experiment 8a, on the other hand, noun-affix learning was better in both conditions, and, critically, there was evidence for a prefixing benefit in noun-picture learning. This pattern of results suggests that, when participants in both conditions show stronger learning of the noun-affix bigrams, this knowledge is more beneficial for noun-picture learning in the prefix condition than the suffix condition. This is in line with our theoretical approach, whereby prefixes make nouns more learnable by reducing the entropy of the noun.

As for our secondary hypothesis – that participants in the prefix condition should be affected more strongly by exposure to incorrect noun-affix bigrams than participants in the suffix condition – the evidence for this was ambiguous. Contrary to prediction, there was evidence for an effect of incorrect bigrams in the suffix condition, whereas in the prefix condition this was ambiguous. On the whole, we cannot draw conclusions with respect to this hypothesis.

Finally, with respect to noun-affix learning, as mentioned above, there was evidence for learning in both conditions. However, unlike Experiment 7, where the evidence for a suffixing advantage was ambiguous, Experiment 8a yielded evidence for no difference between the two affix conditions.

To summarize, Experiment 8a provided evidence for our key hypothesis – the prefix condition were better at noun-picture learning than the suffix condition, suggesting that learners are sensitive to the entropy-reducing effect of prefixes. Experiment 8b attempts to replicate this critical finding.

6.5 Experiment 8b

6.5.1 Method

6.5.1.1 Participants

Fifty-six participants (30 in Prefix and 26 in Suffix condition[2]) were recruited through Prolific Academic. All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for their participation.

6.5.1.2 Stimuli and Procedure

Same as Experiment 8a.

[2]As per the usual optional stopping procedure, our plan was to begin by inspecting 20 participants per condition. However, estimating a high drop-out rate, an additional 10 participants per condition were tested. Only four did not return to complete Day 2, and so all the data were analysed.

Table 6.7: Experiment 8b: Noun-Picture Test Statistics.

Hypothesis	Contrast in lme	Data	Mean difference	SE	H1	B	Robustness region	p
Prefix better than Suffix	Main effect affix	8b	-0.22	1.06	2.87[1]	0.3	[2.56 : ∞]	.838
		Joined	0.56	0.84	7.57[2]	0.2	[1.45: ∞]	.786
Learning bigrams > no learning	Main effect bigram learning	8b	0.04	1.07	9.36[2]	0.12	[3.1 : ∞]	.971
		Joined	0.97	0.85	7.57[2]	0.37	[0 : 3.05]	.315
Greater bgr learn. effect in Prefix	Affix by bigram learning int.	8b	0.05	2.14	4.68[3]	0.42	[0 : 6.13]	.983
		Joined	-0.69	1.71	3.79	0.31[3]	[0 : 1.35]	.517
Greater correct-bgr effect in Prefix	Affix by correct bigram int.	8b	0.11	0.63	0.52[4]	0.85	[0 : 2.08]	.86
		Joined	-0.20	0.49	3.94[2]	0.09	[0 : 3.6]	.147
Breaking down the affix-by-correct bigram interaction by affix-condition								
Prefix better with correct	Correct bigram effect in Prefix	8b	0.01	0.32	0.09[5]	0.97	[0 : 0.9]	.974
		Joined	0.14	0.27	0.32[5]	0.92	[0 : 0.2]	.262
Suffix better with correct	Correct bigram effect in Suffix	8b	0.09	0.35	0.01[6]	1.01	[0 : 1.2]	.804
		Joined	0.32	0.28	0.14[6]	1.43	[0 : 0.2]	.353

[1]Based on an estimate from 7b (method A)
[2]Intercept from the same lme (method B)
[3]Twice the intercept from the same lme (method C)
[4]Same effect from 7a (method A)
[5]Correct bigram effect in Suffix (method A)
[6]Correct bigram effect in Prefix (method A)

6.5.2 Results

6.5.2.1 Noun-picture test

The data are shown in Figure 6.10 and the statistics are in Table 6.7. Contrary to prediction, there was evidence for no item-learning benefit of prefixing compared to suffixing (prefix: $M = 84.12\%$, $SD = 26.91\%$, suffix: $M = 87.64\%$, $SD = 23.51\%$).

The effect of incorrect noun-affix bigrams on performance in the second noun-picture test There was ambiguous evidence for the predicted greater effect of exposure to incorrect bigrams on performance in the prefix condition (correct: $M = 83.67\%$, $SD = 28.21\%$; incorrect: $M = 84\%$, $SD = 27.55\%$) compared to the suffix condition (correct: $M = 87.5\%$, $SD = 27.26\%$; incorrect: $M = 88.04\%$, $SD = 27.25\%$).

The effect of noun-affix learning on noun-picture learning There was evidence that participants who performed above-chance on the noun-affix test *were not* more accurate in the noun-picture test than those participants who performed at chance-level or below. There was ambigous evidence for a stronger effect of learning noun-affix bigrams on noun-picture learning in the prefix condition (learning: $M = 86.83\%$, $SD = 23.16\%$; no learning: $M = 80.68\%$, $SD = 31.27\%$) compared to suffix condition (learning: $M = 87.5\%$, $SD = 24\%$; no learning: $M = 87.85\%$, $SD = 23.43\%$).

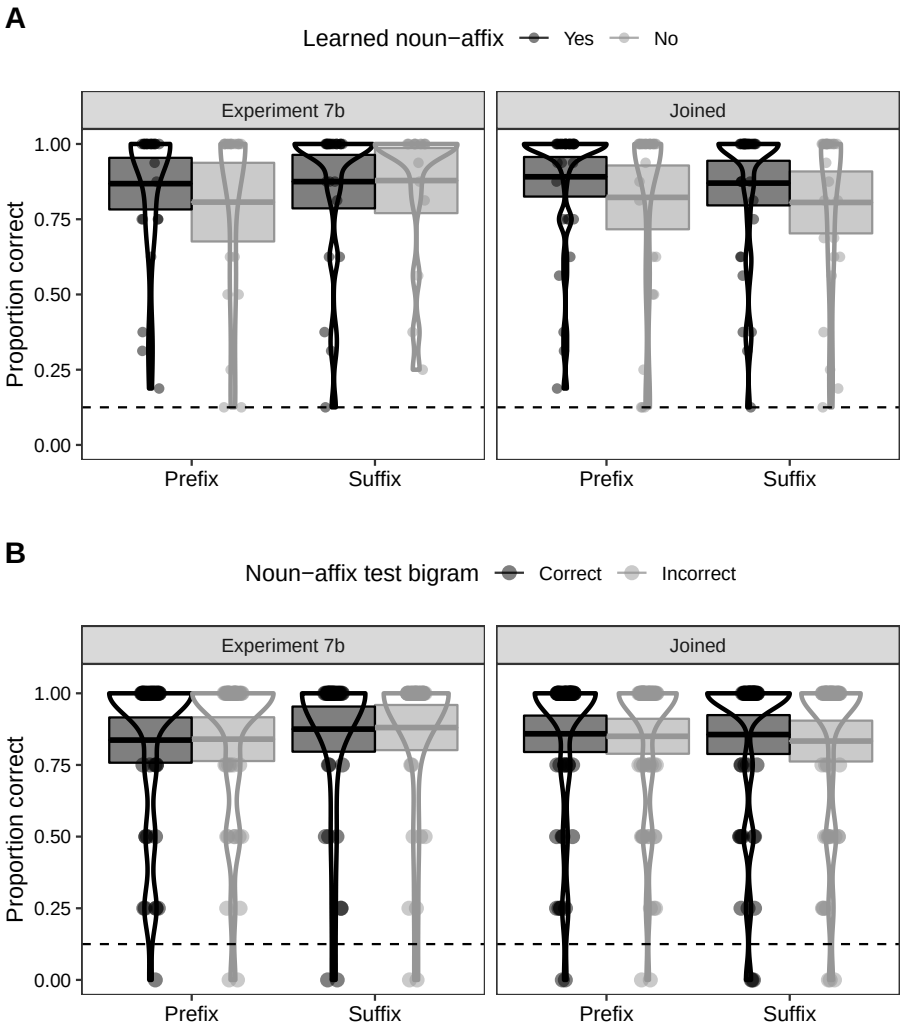


Figure 6.10: Experiment 8b: Panel A: average accuracy on the two noun-picture tests (combined) of participants who showed noun-affix learning (black) and those who did not (grey). Panel B: average accuracy on the second noun-picture test for those nouns which occurred with the correct affix in the noun-affix test (black) and those which occurred with the incorrect affix (grey). Data are from Experiment 8b (left) or combined with 8a (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line indicates chance-level performance (0.125 or 1/8).

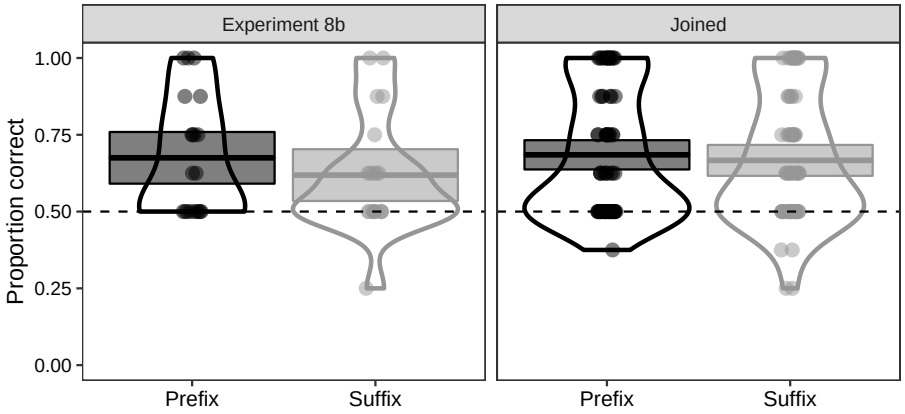


Figure 6.11: Experiment 8b: Proportion of correct responses on the Noun-affix test in Experiment 8b (left) or combined with 8a (right). Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. Dashed line is chance-level performance.

Table 6.8: Experiment 8b: Noun-Affix Test Statistics.

Hypothesis	**Contrast in lme**	**Data**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Suffix better than Prefix	Main effect of affix	8b	-0.26	0.27	1.33[1]	0.11	[0.39 : ∞]	.34
		Joined	-0.09	0.17	0.81	0.15[2]	[0.33 : inf]	.599
Breaking down by affix condition								
Prefix above chance	Prefix Intercept	8b	0.76	0.19	0.50[3]	542.66	[0.06 : >4.59]	<.001
		Joined	0.85	0.12	0.76[3]	2.8×10^{9}	[0.02 : >4.59]	<.001
Suffix above chance	Suffix Intercept	8b	0.50	0.19	0.76[4]	13.79	[0.09 : 4.37]	.007
		Joined	0.76	0.12	0.85[4]	2×10^{7}	[0.03 : >4.59]	<.001

[1]Based on an estimate from 7b (method A)
[2]Intercept from the same lme (method B)
[3]Suffix intercept from the same lme (method A)
[4]Prefix intercept from the same lme (method A)

6.5.2.2 Noun-affix test

The data are shown in Figure 6.11 and the statistics are in Table 6.8. There was evidence for no difference between conditions in noun-affix learning, and this was true for combined data, too. COmparing each condition to chance showed evidence for learning both in the prefix condition ($M = 67.5\%$, $SD = 19.19\%$) and in the suffix condition ($M = 61.88\%$, $SD = 19.23\%$), and this was also true for combined data.

6.5.3 Discussion

Experiment 8b was a replication of Experiment 8a. The most important finding from Experiment 8a which we aimed to replicate was a prefixing advantage in noun-picture learning. However, Experiment 8b showed evidence for no prefixing advantage – there was no difference between the affix conditions in noun-picture learning. Combining the data from Experiments 8a and 8b also showed evidence for the null. I therefore conclude

that Experiment 8 showed evidence that the prefix condition is not better than the suffix condition in noun-picture learning.

With respect to our secondary hypothesis for noun-picture learning, where poorer performance on the second noun-picture test in the prefix condition was predicted due to exposure to incorrect noun-affix bigrams in the noun-affix test (and no difference between the two noun-picture tests in the suffix condition), as evidenced by an affix-by-correct-bigram interaction, both Experiment 8a and 8b showed ambiguous evidence for this, but with combined data, there was evidence for no interaction.

Finally, in Experiment 8a, the evidence that those participants who showed learning of noun-affix bigrams were better at noun-picture learning than those participants who did not was ambiguous, but in Experiment 8b there was evidence for no effect. Combined data suggested ambiguous evidence, with the Bayes Factor of 0.37 being in the direction of more evidence for the null. As in Experiment 8a, the evidence for an affix-by-bigram-learning interaction was ambiguous in Experiment 8b, but combined data suggested evidence for the null.

To summarize, Experiments 7 and 8 suggest evidence that reduced entropy of the noun in the prefix condition does not have an effect on the learning of the "meaning" of the noun. Additional analyses also suggested no relationship between noun-affix learning and noun-picture learning, and no evidence that this relationship is different across the two affix conditions. This is contrary to our prediction, which is based both on our mathematical analysis of entropy (Section 3.2.6), and on the work of Dye and colleagues (2017, 2018), who proposed that prenominal gendered articles and adjectives reduce the entropy of the upcoming noun, which in turn makes the noun easier to process. From this, we hypothesised that this processing advantage will lead to better learning of the "meaning" of the noun (better recall of the picture which co-occurred with the noun) (see Section 1.4.2 for evidence that lower entropy reduces cognitive load), but Experiments 7 and 8 provided no evidence for this. However, before ruling out this hypothesis, it is worth considering two possibilities: (1) that our paradigm did not induce the processing advantage to begin with, or (2) that, if the processing advantage was present, our paradigm was not sensitive to its effects on learning.

Beginning with (1), it appears that the noun-picture mappings in Experiments 7 and 8 were highly learnable in their own right and that noun-affix learning has little effect on this. In Experiment 7, noun-affix learning was poor, but noun-picture learning was relatively high; in Experiment 8, noun-picture learning was near-perfect, and while noun-affix learning was stronger than in Experiment 7, it was still not very high (62% in the suffix condition and 67% in the prefix condition). In fact, the increased noun-picture learning in Experiment 8 is more likely to be caused by increased exposure to a smaller training set in that experiment compared to Experiment 7, than by slightly better noun-affix learning – this is corroborated by additional analyses which showed no indication that noun-affix learning was related to noun-picture learning. This suggests that the noun-picture mappings in this experiment were highly learnable in-and-of-themselves, and that the entropy reduction by the prefix may not have been necessary, or even beneficial, for the

learning of noun-picture mappings. Why was this the case?

Recall that in every trial, the picture was presented first; at this point eight nouns were possible (or 16 in Experiment 7). After this, the affix was played in the prefix condition, and in theory this should have reduced the possible set of candidates in half, thus smoothing the entropy of the utterance over the prefix and the upcoming noun. However, it is possible that the visual objects and the nouns in these experiments were sufficiently discriminable and learnable, to the point where the affixes could be ignored. Given the strong noun-picture learning, it is possible that, relatively quickly in the course of learning, once participants saw the picture, the picture itself was enough to reliably predict the noun without the help of the affix (in the prefix condition). Were this the case, it would explain why there was no difference between the conditions, as the picture was presented first in both conditions. It is possible that the affix would be more useful for noun learning in more complex learning contexts, for example, when multiple possible referents are on-screen at the same time, or when the utterance is more complex and the affix is more salient (in this study, the affixes were monosyllabic). A related possibility is that the benefit of the affix could be observed earlier in learning, before noun-picture learning consolidates in both conditions (recall that in Chapter 5 I discussed how depending on where on the learning trajectory participants are when they are tested may lead to differences in the effects of order). One way to address this would be to collect data at each learning trial (in all experiments so far in this thesis data are only collected at test) – I return to this point below.

Turning to (2), in addition to our mathematical analyses of entropy smoothing, multiple psycholinguistic experiments have shown that listeners use the information provided by prenominal articles to reduce the search space for the upcoming noun

One difference is that previous work used paradigms in which the entire search space was presented to the participants visually on-screen, and the gendered article was used to reduce the uncertainty about the referent of the upcoming noun. In Experiments 7 and 8, however, rather than multiple referents being present, only the correct referent was presented on-screen (this design matches Arnon and Ramscar, 2012), meaning that the prefix reduced the uncertainty only about the upcoming artificial label. In other words, rather than the whole search space being presented visually, in these experiments, the search space was the listener's mental lexicon. While in theory the entropy-reducing effect of prefixes is not limited to what is immediately present in the listener's visual field, it is possible that this effect is harder to capture with auditory labels than with visual objects – this might be because participants find visual objects are more perceptually salient or more concrete.

Another, more critical difference is that previous studies were conducted in participants' native languages, to which they had years of exposure – in other words, there was no learning involved. In the present study, on the other hand, noun-affix learning was below 70% at best, whereas presumably participants in previous literature had much stronger knowledge of which nouns could follow which gendered articles in their native language. Study 3 of this thesis took this idea further by examining this effect on learning a new miniature "language". As previous work did not involve learning, the key measures were on-line

measures of processing such as eye movements and reaction time. In our study, we were interested in learning as captured by accuracy – however, it is possible that this approach to measuring learning was not appropriate for capturing the effect of interest. Specifically, learning was tested through measures administered post-exposure. This means that there are no data regarding the dynamics of learning as it occurred, and so inferences about the mechanisms of learning are made retroactively, based on participants' accuracy on offline tests. The forced-choice offline tests that were used are explicit and meta-cognitive, and as such they might be interfering with what was learned during exposure and they might be tapping into explicit learning mechanisms that could obscure the subtle, implicit differences in the dynamics of learning between the two conditions that we are interesting in measuring (see Malmberg, Criss, Gangwani, & Shiffrin, 2012; Siegelman et al., 2017, for discussions of the limitations of offline measures of statistical learning).

These differences are addressed in Experiment 9, where item-learning is explored using a paradigm in which data are collected during exposure, and where the referent of learning is presented together with one or more competitors – the cross-situational learning paradigm (Yu & Smith, 2007). In this paradigm, a target referent and one or more competitors are present on-screen at a given trial. A label is played, and participants are asked to select the correct referent. Participants do not receive feedback, and at first their responses will be based on guessing, however by tracking cross-trial co-occurrence statistics between labels and referents, they become increasingly more confident in their responses; this paradigm has been used in numerous studies (e.g., Medina, Snedeker, Trueswell, & Gleitman, 2011; Ramscar, Dye, & Klein, 2013; Roembke & McMurray, 2016; K. Smith, Smith, & Blythe, 2011; L. Smith & Yu, 2008; Vouloumanos & Werker, 2009; Yu & Smith, 2007). In addition, making the search space more concrete to participants by displaying multiple objects on-screen, another advantage of this paradigm is that it allows us to observe learning not only through a test administered post-training, but also through the training itself. This is a richer data set, since we can observe learning as it unfolds from one trial to the next. As each training trial provides a datapoint, this also increases the statistical power of the experiment. Displaying more than one referent in a single trial may make noun-picture learning harder, and therefore learners in the prefix condition may rely on the prefix more than in previous experiments. Finally, the paradigm also allows us to test for more nuanced effects of prefix by manipulating the type of competitors that occur with the target (more detail in Section 6.6.1).

6.6 Experiment 9

6.6.1 Rationale and predictions

In this experiment, the same training set was used as in Experiment 8, meaning that participants learned the same pairs of nouns and pictures and nouns and affixes as in that Experiment (the particular assignments were randomised again), with the key difference being the learning procedure in the exposure phase: whereas in Experiment 8, participants were shown pictures and played nouns and affixes in specific orders, but no action was

required on their end, in Experiment 9, in each trial during exposure, each target picture appeared with two competitors, and participants were asked to click on the picture they thought matched the label they were played. Following typical cross-situational learning procedure, no feedback was given.

Critically, there were three types of trials: in the first type, both foils occurred with the same affix as the target image (*both foils same affix*); in the second, one foil occurred with the same affix as the target image and the other occurred with the opposite affix (*one foil opposite affix*); finally, in the third type of trial, both foils occurred with the opposite affix (*both foils opposite affix*). With this design, we intended to capture the effects of prefixing in a more robust way: the entire search space is immediately available in the visual field, and when participants hear the prefix, depending on the competitors, the prefix reduces the search space to one or two items (as opposed to three), and, we hypothesise, makes the correct noun-picture mapping easier to learn. On each type of trial, the prefix reduces the entropy of the upcoming noun by 50% (as half of the nouns occur with one affix and half with the other). However, on *both foils same affix trials*, the prefix does not decrease the uncertainty about the correct picture. In *one foil opposite affix*, once the affix is heard, only two of the images on-screen remain the possible correct answer, and in *both foils opposite affix*, the correct image can be identified based on the prefix alone. In the suffix condition, on the other hand, when the noun is played, each picture is an equally likely response. Even though participants can only indicate their response after the suffix finished playing, meaning they can in principle still use the affix to eliminate competitors, the entropy of the noun remains the same on every trial type, which should make it harder to learn the correct noun-picture mapping.

Note that in this study "dummy" affixes were introduced. In the prefix condition, this was a suffix, which was played after the noun and was the same on every trial. In the suffix condition, the dummy affix was a prefix, which was played before the noun, but was identical on every trial, and therefore carried no information about the noun. The dummy affixes mean that the difference between conditions is not the auditory nature or structure of the utterance – in both conditions participants heard the same sequence: *affix-noun-affix* – but the way in which uncertainty is distributed across the utterance. We believe this adjustment to the paradigm provides a more precise and robust test of our key predictions.

As in the rest of this chapter, we predicted that participants in the prefix condition will be better at noun-picture learning (more accurate at choosing the correct image) than participants in the suffix condition. While key data come from the cross-situational training task, the same noun-picture test used in Experiments 7 and 8 was included at the end of Experiment 9 for comparison. We also predict that the prefix condition will be faster than the suffix condition, however given that mouse-clicks were used rather than button presses (these were deemed more appropriate given that there were three pictures to choose from), reaction time data are treated as complementary, but not critical for the hypothesis. This paradigm also allowed us to test more nuanced effects of prefixing by varying the degree of informativeness of the prefix. We therefore predict that participants in the prefix condition will be more accurate on *both foils opposite affix* than on *both foils same affix* trials, and

that this difference will be bigger in the prefix condition than in the suffix condition (as evidenced by an interaction). We predict the same effect for *one foil opposite affix* versus *both foils same affix*, but our theoretical approach predicts a smaller effect here.

6.6.2 Method

6.6.2.1 Participants

Eighty-two participants (41 per condition) were recruited through Prolific Academic. All participants were adult monolingual native speakers of English with no known language impairments, hearing, or vision impairments. Participants were randomly allocated to one of the two affix conditions. They provided informed consent and were paid for their participation.

6.6.2.2 Stimuli

Same as Experiment 7.

6.6.2.3 Procedure

Cross-situational learning test. This test consisted of four blocks of trials with 84 trials in each block. On a single trial, three images would appear on-screen at the same time. After 500ms, a label was played. Participants were instructed to click on the picture they thought matched the label, and were told that they would have to guess at first, but that they should soon become more confident in their responses. Participants' response was followed by a blank screen and after 1000ms, a new trial would begin.

Item-learning test. Same as the test used in Experiments 7 and 8.

6.6.3 Results

Analyses were not pre-registered due to time constraints, though the same procedures for choosing the H1 were followed as elsewhere in the thesis. Optional stopping was used, with the rule to stop collecting once there is substantial evidence for/against the key prediction – more accurate performance on the cross-situational learning test in prefix condition. The data were inspected at 20 and at 40 participants per condition.

6.6.3.1 Cross-situational training: Accuracy

The inferential statistics are in Table 6.9 and the data are shown in Figure 6.12. There was evidence for no predicted prefixing advantage (prefix: $M = 81.21\%$, $SD = 13.78\%$; suffix: $M = 78.95\%$, $SD = 17.03\%$). There was evidence that accuracy in both conditions improved by block. However, there was evidence that participants in the prefix condition did not improve faster than the suffix condition.

With respect to the different effects of trial type on performance in the two affix conditions, we predicted a greater difference between the trials on which both foils had the same gender as the target and those where both foils had the opposite gender in the prefix

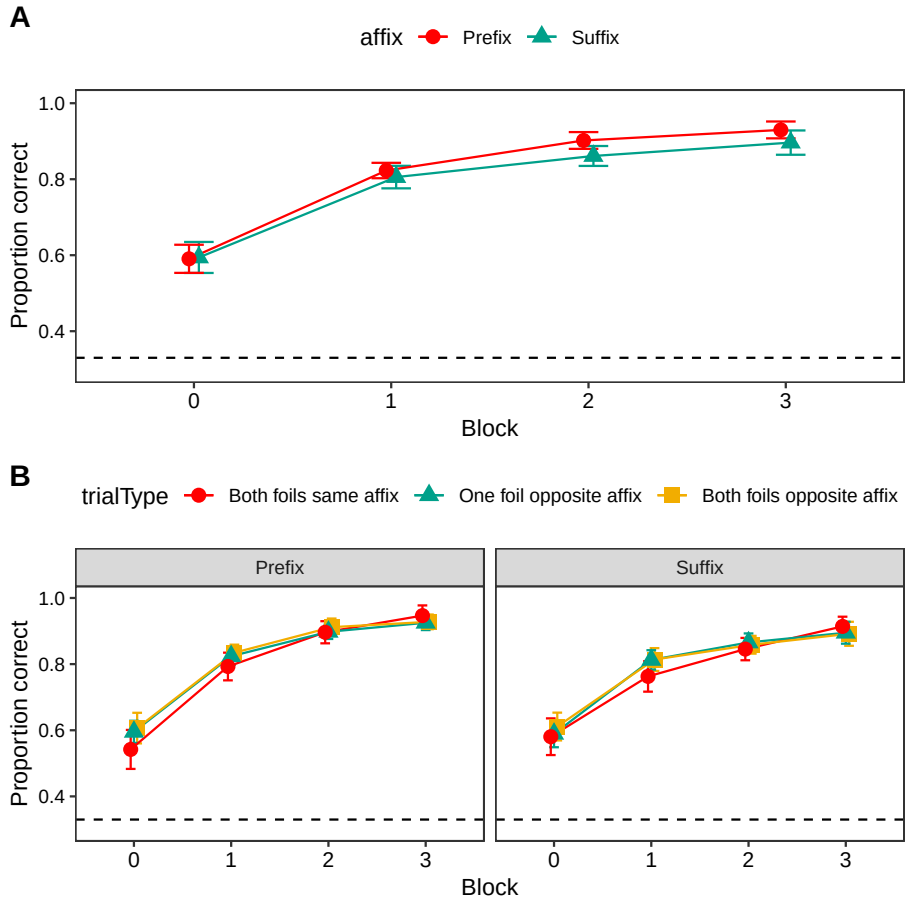


Figure 6.12: Experiment 9. Panel A: Average proportion of correct responses per block in the cross-situational learning test for the prefix (red circle) and the suffix condition (green triangle). Panel B: Average proportion of correct responses in the cross-situational learning test for the prefix (left panel) and the suffix condition (right panel), broken down by trial-type. Error bars represent 95% CI around the mean. The dashed line is chance-level performance (0.33 or 1/3)

condition (foils opposite affix: $M = 82.27\%$, $SD = 3.26\%$; foils same affix: $M = 79.9\%$, $SD = 4.21\%$), than in the suffix condition (foils opposite affix: $M = 79.48\%$, $SD = 3.57\%$; foils same affix: $M = 77.53\%$, $SD = 4.17\%$). However, the evidence for this was ambiguous. The evidence was also ambiguous for a greater diffeernce between trials on which one foil had the same gender as the target, and those where both foils had the opposite gender in the prefix condition (one foil opposite affix: $M = 81.02\%$, $SD = 3.46\%$) compared to the suffix condition (one foil opposite affix: $M = 79.03\%$, $SD = 4.14\%$) (Figure 6.12 plots this data). Analysing the two affix conditions separately, we also found that these two effects were ambiguous in both conditions.

6.6.3.2 Cross-situational training: RT

For the analysis of reaction time, only those trials on which participants gave a correct response were considered, and responses longer than 4000ms were removed (this threshold was determined through visual inspection of the distribution of the data). For each participant and for each block, the values that fell above/below 2.5 SD for that participant in that block were removed. The data are shown in Figure 6.13 and the inferential statistics

Table 6.9: Experiment 9: Cross-Situational Learning Test Statistics.

Hypothesis	**Contrast in lme**	**Mean difference**	**SE**	**H1**	**B**	**Robustness region**	**p**
Prefix better than Suffix	Main effect of affix	-0.01	0.36	3.24[1]	0.11	$[1 : \infty]$	.981
Learning improves by block	Main effect of block	1.43	0.10	1.08[2]	7.45×10^{10}	[0 : >5.28]	<.001
Greater improvement by block in Prefix	Affix by block interaction	0.05	0.19	1.43[3]	0.17	[0.69 - inf]	.787
Both-foils-opposite affix vs both-foils-same affix (contrast 1)							
Greater effect of trial-type in Prefix	Affix by contrast1 int.	0.00	0.13	0.13[4]	0.71	[0 : 0.39]	.965
Both foils opposite better than *same* in Prefix	Effect of contrast1 in Prefix	0.07	0.10	0.12[5]	1.09	[0 : 0.6]	.465
Both foils opposite better than *same* in Suffix	Effect of contrast1 in Suffix	0.07	0.09	0.07[6]	1.26	[0 : 0.55]	.442
Both-foils-opposite affix vs one-foil-opposite affix (contrast 1)							
Greater effect of trial-type in Prefix	Affix by contrast2 interaction	0.13	0.10	-0.00[7]	0.99	[0 : 1.24]	.201
Both foils opposite better than *one opposite* in Prefix	Effect of contrast2 in Prefix	0.12	0.08	0.07[6]	2.09	[0 : 1.32]	.131
Both foils opposite better than *one opposite* in Prefix	Effect of contrast2 in Suffix	-0.02	0.07	0.12[5]	0.44	[0 : 0.16]	.81

[1]Intercept from same lme (method B)
[2]Third of the intercept from same lme (method E)
[3]Main effect of block from the same lme (method C)
[4]Affix-by-contrast2 interaction (method A)
[5]Effect of contrast2 in Prefix (method A)
[6]Effect of contrast1 in Prefix (method A)
[7]Affix-by-contrast1 interaction (method A)

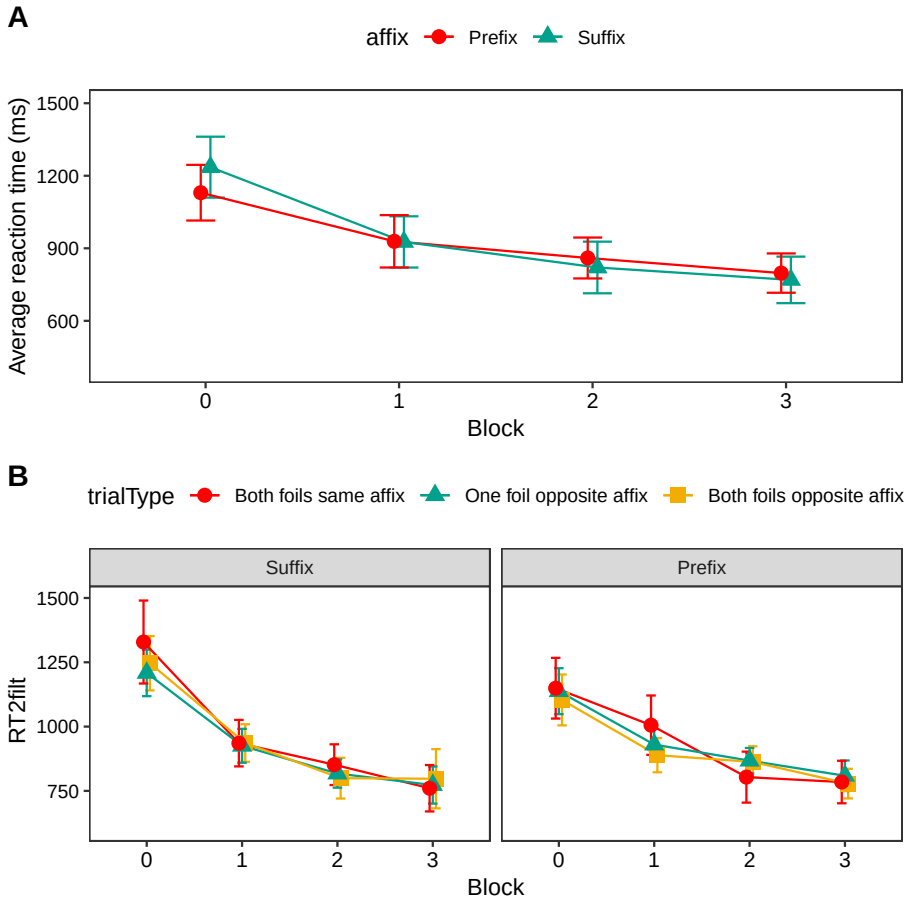


Figure 6.13: Experiment 9. Panel A: Average reaction time per block in the cross-situational learning test for the prefix (red circle) and the suffix condition (green triangle). Panel B: Average reaction time per block in the cross-situational learning test in the prefix (left) and the and suffix condition (right) broken down by trial type. Error bars are 95% CI around the mean.

are in Table 6.10.

The evidence for faster performance in the prefix compared to the suffix condition was ambiguous (prefix: $M = 911.19$ms, $SD = 208.71$ms; suffix: $M = 901.41$ms, $SD = 205.1$ms), however, there was strong evidence that the participants' speed did not decrease faster in the prefix condition than in the suffix condition. The evidence that the prefix condition was more strongly affected by the types of trials than the suffix condition was ambiguous.

6.6.3.3 Item learning test

The data are shown in Figure 6.14. There was evidence for no difference between conditions in item-learning (prefix: $M = 92.68\%$, $SD = 20.34\%$, suffix: $M = 85.67\%$, $SD = 26\%$), ($\beta = 0.915$, SE $= 1.293$, $z = 0.708$, BF $= 0.251$, $p = .479$).

6.6.4 Discussion

The aim of Experiment 9 was to show that a processing advantage of hearing a prefix before the noun leads to better learning of the "meaning" of the noun (better accuracy at choosing the correct picture given a noun label) compared to the suffix condition where an

Table 6.10: Experiment 9: Cross-Situational Learning Test Statistics for Reaction Time.

Hypothesis	Contrast in the lmer	Mean difference	SE	H1 estimate	BF	Robustness region
Prefix faster than Suffix	Main effect of affix	0.01	0.05	0.16[1]	0.34	$[0 : \infty]$
Greater improvement by block in Prefix	Affix by block interaction	0.04	0.02	-0.12[2]	0.07	$[0 : >5.28]$
Both foils opposite better than *same* in Prefix	Effect of contrast1	-0.01	0.03	0.01[3]	0.94	$[0.69 : \infty]$
Both foils opposite better than *one opposite* in Prefix	Effect of contrast2	-0.03	0.02	0.01[3]	0.68	$[0 : 0.39]$

[1]Half the maximum effect (method B)
[2]Main effect of block (method C)
[3]Main effect of affix (method C)

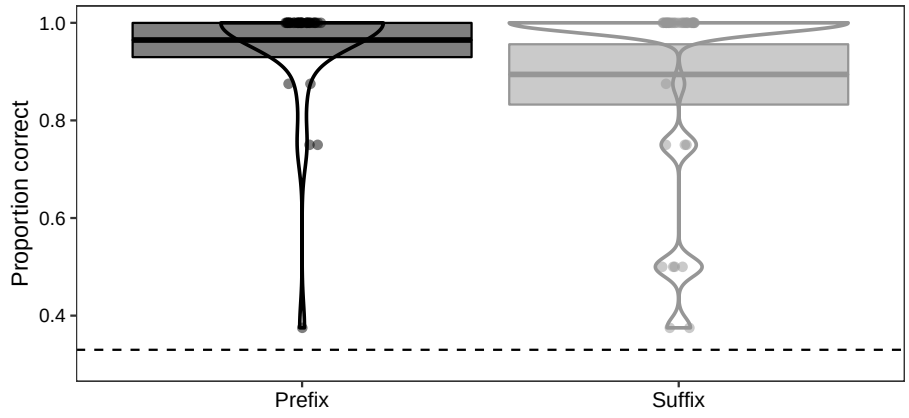


Figure 6.14: Experiment 9: Proportion of correct responses in the Item-learning test. Points show by-participant means, and violins show the kernel probability density of participants' means. Horizontal lines in each violin are by-condition means, with the boxes showing 95% CI around the means. The dashed line is chance-level performance (0.125 or 1/8).

informative prefix was not played. However, there was found evidence for the null – there was no difference between the two affix conditions with respect to accuracy on the cross-situational learning test, but accuracy was high in both conditions (approaching ceiling in the final block). Similarly, there was evidence that performance was not faster in the prefix compared to the suffix condition. The accuracy and reaction time results thus suggest that participants in the prefix condition did not exploit the information in the prefix during processing. This is corroborated by the fact that there was no clear evidence for an effect of trial type. If participants used the information carried by the prefix, there would be better performance on trials where the prefix eliminates competitors compared to when it does not – however we did not see this, but the evidence was ambiguous, meaning that no clear conclusions can be drawn. Finally, contrary to prediction, there was evidence that the prefix condition were not better than the suffix condition on the noun-picture test (the same test used in Experiments 7 and 8), but with high learning in both affix conditions. This finding – even though contrary to our original prediction – is consistent with Experiments 7 and 8, where there was high noun-picture learning in both conditions, but evidence for no difference between the conditions. The findings across the three experiments are discussed in the General Discussion of Study 3.

6.7 General Discussion of Study 3

The aim of this chapter was to test the prediction that hearing a prefix before hearing a noun reduces the uncertainty about the noun. Specifically, in the experiments reported here, one half of the items occurred with one prefix, and the other half occurred with the other; therefore, as soon as a prefix is heard, this eliminates half of the possible candidates for the upcoming noun. Reduced uncertainty about the noun makes the noun easier to process, which in turn, we hypothesised, should make it easier for participants to remember the picture that the noun occurred with (its "referent"). In the suffix condition, on the other hand, the noun is always heard first, at which point the entropy of the noun is maximal (any noun could occur); however, once the noun has been played, only one of the two suffixes is possible, and therefore the forward transitional probability of the suffix given the noun is 1. From this, we predicted that participants in the prefix condition would be better at the learning of noun-picture mappings, whereas participants in the suffix condition would be better at noun-affix mapping. Note, therefore, that while sufficient amount of noun-affix learning is necessary in both conditions, this is expected to be stronger in the suffix condition, but to only be beneficial to noun-picture learning in the prefix condition. Each hypothesis is discussed in turn below.

Across three large-scale experiments, there was no consistent evidence for a prefixing advantage in noun-picture learning. In Experiment 7a and the replication Experiment 7b, in which participants learned 16 nouns in total (eight per affix), the evidence for a prefix advantage was ambiguous, however, combined data suggested evidence that the prefix condition was not better than the suffix condition (evidence for the null). In Experiment 8, a two-day procedure was used, in which participants learned eight nouns in total over

two days. Experiment 8a showed evidence for better performance in the prefix condition. However, the replication Experiment 8b showed evidence for no prefixing advantage, and combining the data from the two experiments also showed evidence for the null. The performance in both conditions was near-ceiling, however. Finally, Experiment 9 implemented a cross-situational learning procedure, and this too showed evidence for no difference between the two affix conditions. Taken together, the three studies suggest that there was no evidence that the prefix condition was better at noun-picture learning that the suffix condition.

This pattern of results is contrary to our prediction, which is based on a converging body of work. Specifically, Dye and colleagues (Dye et al., 2017, 2018) demonstrated in corpus analyses of German and English that prenominal gendered articles and adjectives reduce the uncertainty of the noun. In addition to this, numerous studies showed that reduced entropy is related to less cognitive effort in processing – from shorter reading times (S. L. Frank et al., 2015), to distinguished patterns of brain activity (Boston et al., 2011; S. L. Frank, 2013; N. J. Smith & Levy, 2013). In an artificial language learning study, Arnon and Ramscar (2012) demonstrated better noun-picture learning in a condition in which nouns were taught together with their gendered articles, rather than in isolation. In addition to this work, our own calculations of entropy in our experimental conditions showed that only in the prefix condition is the entropy of the noun reduced; in the suffix condition, on the other hand, the suffix cannot reduce the entropy of the noun. This body of work strongly suggests that previous parts of the utterance reduce the entropy of the following parts (that is, smooth the entropy over the whole utterance), which helps processing and possibly learning.

However, much of the evidence for this effect is indirect with respect to learning – it comes from experiments with child or adult speakers of their first language, or corpus analyses of speech in people's first language – that is, from speakers who have amassed some critical amount of knowledge about the statistical distributions of linguistic forms. In the experiments in this thesis, predictions are made under the assumption that participants have learned the underlying probability distribution of nouns and affixes, that is, that they have learned which nouns occurred with which affix, and that they are using this knowledge to predict the upcoming noun. However, a closer analysis of participants' performance suggests that this may not have been the case. The learning of noun-affix co-occurrences was tested using a test in which a trained noun was played either with the correct affix or the opposite affix, and participants indicated whether the label they heard was correct or incorrect. In Experiment 7a, there was evidence for above-chance performance on this test in the suffix condition alone, whereas in the prefix condition the evidence was ambiguous. In Experiment 7b, there was evidence for learning in both conditions, and the same was true for combined data. We also analysed whether those participants who performed above chance on the noun-affix test were better at noun-picture learning than those who performed at chance, and found ambiguous evidence in Experiment 7a, and evidence for the null in Experiment 7b and with combined data. Therefore, when there was learning of the noun-affix co-occurrences, this was low (lower than 55% in both conditions), and there was no

evidence that this knowledge helped noun-picture learning in either condition. Given that our predictions are made under the assumption that participants rely on the knowledge of the probability of noun-affix co-occurrence to identify the noun, the predicted effect can only be observed when participants do indeed learn noun-affix co-occurrences to an extent that is useful in noun processing.

With this in mind, Experiments 8a and 8b were designed – here, the number of items was reduced by half, and a two-day procedure was used, in which participants were first exposed to noun-affix pairs without pictures on Day 1, and on Day 2 they were exposed to noun-affix pairs again but with accompanying pictures. The idea was that participants would show strong learning of noun-affix pairs during this extended exposure, and that this knowledge would facilitate noun-picture learning in the prefix condition. While the extended exposure to noun-affix pairs did somewhat improve noun-affix learning in both conditions (judging by higher means and evidence for above-chance learning in both conditions), the evidence that this boosted noun-picture learning in both conditions was ambiguous in Experiment 8a, and there was evidence for the null in Experiment 8b; evidence with joined data was ambiguous, with more evidence in the direction of the null.

Therefore, across the four experiments, the learning of noun-affix co-occurrences was not clearly related to the learning of noun-picture pairs. The reasons for this are discussed at length in Section 6.5.3. One possibility was that, given how strong noun-picture learning was, particularly in Experiments 8a and 8b, the noun-picture mappings were so learnable that participants could ignore the information carried by the affix. In addition to this, our paradigm only involved off-line forced choice tests, which may not be suitable for capturing this subtle effect, in which the prefix smooths the entropy of the noun in real-time. A paradigm which provides trial-by-trail data was more appropriate, and with this in mind, in Experiment 9, a cross-situational learning paradigm was used. Participants were presented with multiple picture on each trial and a single noun-affix label, and were asked to choose the correct picture at each trial. However, here there was no evidence for a prefixing advantage in noun-picture learning, with evidence for the null. This is potentially important considering strong evidence from previous work that participants use parts of the utterance as it unfolds in real-time to identify the referent of the upcoming noun (Dahan et al., 2000; Lew-Williams & Fernald, 2007, 2010). Given the strong converging evidence that the fact that prenominal articles and adjectives reduce the entropy of the upcoming noun facilitates the processing of the noun, it is worth reflecting on why the work in this thesis is inconsistent with this effect.

One possibility is that much of the previous work that empirically demonstrated this faciliatory effect used eye-tracking, and data about participants' eye movements were used as evidence for the processing benefit. It is possible that this kind of effect is harder to capture with measures of accuracy used in this work, which are more explicit by nature (however, we did not find evidence for the effect with reaction time either, though we used mouse-clicks rather than button presses, which might be noisier). Therefore, future work should aim to use implicit measures of learning to capture the effects of entropy on noun-picture processing and learning – I return to this in the General Discussion in Chapter

Another possibility is that the type of learning observed in this study does not benefit from potential processing advantages in the prefix condition (although the lack of processing-related measures in this study makes it difficult to pull apart these two possibilities). Unlike in natural language, the nouns in this study may have been learned via more explicit learning mechanisms; it is possible that participants quickly realised that the middle part of the utterance (the noun) is what uniquely identifies the picture, and explicitly focused on that part of the utterance. To the best of my knowledge, the studies reported in this thesis are the first to study the effect of entropy smoothing on learning, and more work is necessary to develop a paradigm that is better suited to capture this effect, including more implicit on-line measures of learning, and possibly also more complex learning tasks (which would make the information in the affix more useful in learning).

Note that there have been artificial language learning experiments which explored the role of entropy in learning, but rather than manipulating the distribution of entropy across an utterance within a single trial, this work manipulated the entropy of the whole stimulus set. For example, Hendrickson and Perfors (2019) exposed participants to a cross-situational learning task in which the distribution of words had high entropy, where the frequency of individual words was uniform, or low entropy, where the frequency of individual words followed a Zipfian distribution, such that there were a few highly frequent words, and many low frequency words (therefore the distribution was more predictable). Hendrickson and Perfors (2019) found overall better learning in the low-entropy condition, even for the words which were equally frequent in the two conditions (see Kurumada, Meylan, & Frank, 2013, for related findings in artificial speech segmentation). A difference between this work and the present is that, in the present study, each individual utterance/noun+picture pair had the same probability, and therefore the distribution of the whole stimulus set had high entropy. What we manipulated, therefore, was not the entropy of the entire stimulus but how the information was distributed across a single trial; rather than there being high- or low- entropy trials (nouns), in our study there were trials with even or skewed distribution of information (entropy) within the trial. This leaves the possibility that, while individual nouns with higher entropy are more learnable (as Hendrickson & Perfors, 2019, demonstrated), utterances with more even entropy may not be, or that a more subtle, implicit paradigm is required to capture this effect.

Our second hypothesis was that participants in the suffix condition would show better learning of noun-affix mappings than participants in the prefix condition. This prediction comes from the observation that in the suffix condition, the forward transitional probability of the suffix given the noun is 1 (in the prefix condition, on the other hand, the forward transitional probability of the noun given the prefix was 0.125 in Experiment 7 and 0.25 in Experiment 8). As with the other prediction, this too relies on the assumption that participants learn noun-affix co-occurrence probabilities in the first place. The evidence for the suffixing benefit was ambiguous in Experiment 7a, whereas in the replication Experiment 7b, there was evidence for the null (combined data were ambiguous). In both Experiments 8a and 8b, where noun-affix learning was on average stronger than in Experiments 7a and

b, there was evidence for no difference between the conditions in noun-affix learning (note that this hypothesis was not tested in Experiment 9). Therefore, across the four experiments, noun-affix learning was generally low, and was not better in the suffix condition compared to the prefix condition. It is possible that participants in both conditions were not attending to the affix sufficiently. In the suffix condition, this may be because the suffix did not provide any additional information about the picture – all the information was carried by the noun. In the prefix condition, the prefix did provide information about the noun, but the picture itself may have been salient and relatively easy to associate with the noun, that the information about the prefix could be ignored. This highlights an important point – while a purely mathematical analysis may indicate reduced entropy, this may not necessarily (or at all) mean a processing or learning advantage, if, for example, humans do not attend to the stimulus in question. A similar point was made in Chapter 5, where predictions were made about a naïve learner in idealised conditions, but where the evidence may be "obscured" but other cognitive and extra-cognitive factors involved in human learning.

To summarize, Study 3 did not find evidence for a facilitatory effects of prefixing in noun learning. This finding is interpreted in terms of the design of the paradigm.

Chapter 7

General Discussion

This thesis tested the predictions of the discriminative learning theory of language learning, formulated by Ramscar and colleagues (Baayen et al., 2011; Dye et al., 2017, 2018; Ramscar & Yarlett, 2007; Ramscar et al., 2010; Ramscar, Dye, & McCauley, 2013; Ramscar, Dye, & Klein, 2013). This theory is based on insights from decades of empirical research in the psychology of learning (learning theory) and information theory. Critically, under this framework, processes that lead to learning are specified and modelled mathematically, which allows the researcher to formulate precise, testable hypotheses, something that existing theories of language learning have lacked to some extent (as discussed in Chapter 1). Specifically, discriminative learning approach to language views language use, encompassing both learning and processing, as a fundamentally probabilistic process of reducing the uncertainty about the form and the meaning of the utterance, by discriminating informative from uninformative cues. Discriminative learning is driven by competition between cues for predictive value, which can be tested by directly manipulating the amount of cue competition a learning situation provides. This was done in the current thesis. In a series of artificial language learning experiments, the order in which participants were presented with "nouns" and "affixes" was manipulated, such that in the suffix condition, nouns were followed by affixes, whereas in the prefix condition they were preceded by affixes. From the principles of cue competition and prediction error, I hypothesized that participants in the suffixing language would be better at generalizing the statistical structure of the artificial languages to novel items, whereas participants in the prefix condition would show better item learning, that is, better learning of trained instances. Each of these hypotheses is discussed in turn.

7.1 Is there a suffixing advantage in generalization?

In one of the seminal papers in this framework, Ramscar et al. (2010) demonstrated that verbal labels provide fewer perceptual cues, and therefore fewer opportunities for prediction error via cue competition, than complex visual objects, events and states in the environment. Therefore, learning by predicting verbal labels from features of complex referents (objects, events, states) should provide greater cue competition that predicting these features from relatively simple verbal labels. In their experiment with adult learners, Ramscar

et al. (2010) found that when learners were first shown a picture of a novel referent and then played a verbal label (e.g., "That was a wug"), learners were more accurate at generalizing the label to novel referents, compared to when learners first heard the label (e.g., "This is a wug") and were then shown the referent. This finding suggests that learners in the feature-label condition were better at learning the discriminating cues that group novel referents with the label (e.g., all wugs have legs and a trunk, whereas all nizes have wings and tails) than were learners in the label-feature condition, due to greater cue competition in that condition.

In this thesis, the prediction that feature-label learning leads to better generalization than label-feature learning was tested in a different learning context. Specifically, following related work (Ramscar, 2013), feature-label learning was taken to correspond to suffixing, whereby the affix follows the features of the stem morpheme it attaches to, whereas label-feature learning corresponds to prefixing, whereby the affix precedes the features of the stem morpheme. From this we predicted that learning a suffixing language should lead to better generalization than learning a prefixing language. Studies 1 and 2 of this thesis presented a series of artificial language learning experiments in which adult learners were exposed to artificial languages in which "nouns" were either preceded by "affixes" (prefix condition) or followed by them (suffix condition). Each noun was uniquely matched with a novel visual referent. Half of the items (noun-picture pairs) occurred with one affix and another half with the other affix. Affix occurrence was predicted by semantic and phonological cues – nouns which shared certain phonological features and whose referents shared certain visual features occurred with the same affix. Here, the critical question was, given a new item, will participants be able to group it with the correct affix on the basis of the semantic and phonological features? The discriminative learning approach predicts that the suffix condition will be better than the prefix condition at generalization, because of the greater cue competition in this condition, which allows for the dissociation of uninformative cues.

The findings of this thesis broadly support the predictions of the discriminative learning framework with respect to generalization, but also demonstrate that these predictions are more nuanced than previous work may have assumed. Specifically, while all experiments in this thesis except one (Experiment 6) found order effects in generalization, these effects were not always in the predicted direction, showing that it is not the case that feature-label learning is better than label-feature learning by default. Below I interpret these differences in the direction of the effect as a consequence of cue structure in different learning contexts, and the extent to which different structures facilitate cue competition and prediction error. I conclude that, rather than by the order of features and labels or referents and affixes as such, differences in generalization are best explained by differences in cue competition, the core learning mechanism in the discriminative learning framework. Another important insight from this thesis is that modelling cue competition is not straightforward, and that researcher assumptions about the amount of cue competition in a given learning context are still assumptions that must be tested with human learners. Each of these insights is discussed in turn below.

In Study 1, generalization was better in the prefix condition than in the suffix condi-

tion, which was contrary to what was predicted from the performance of computational models trained using a discriminative learning algorithm (Experiment 1). However, further thinking about the input sets used in this study raised the possibility that we did not model the learning context appropriately, and that our predictions may not have been accurate. In particular, in this study, the one cue which consistently predicted affix occurrence – body shape of the aliens – may also have been the most salient cue, and other cues may have been insufficiently salient to compete with that cue for predictive value. In the hypothetical absence of cue competition, this task becomes one of learning two simple correlations: shape1 predicts affix1 and shape2 predicts affix2 (and the other way round in the prefix condition), and this kind of learning may be stronger in the prefix condition (although our theory does not predict nor explain this finding). Furthermore, the design of the stimulus set was such that, in order to learn the correct affixes, it was not necessary to "unlearn" frequent but uninformative cues, as the most frequent cues were also most informative. Therefore, correct generalization in Study 1 did not critically depend on learning from negative evidence (prediction error). To further elucidate this, in Chapter 5, I re-visited the computational model presented in Experiment 1. Specifically, I used a different learning metric (the softmax function) which primarily captures the learning of the predictive cues (rather than the "unlearning" of unpredictive cues), and found that this metric shows better accuracy in the prefix simulation compared to the suffix simulation, much like what was observed with human learners. However, while the results from Study 1 inspired this modelling work, this is not to say that the modelling in turn explains the results from Study 1. The extent to which the prefixing advantage in Study 1 is due to that design favouring learning from positive evidence is unclear – our present method does not allow us to verify how exactly humans encoded the training input (and in turn how best to model it) and whether or not there indeed was cue competition. It is also possible that other aspects of the study caused the prefixing advantage. For example, there was a high incidence of explicit learning in this study, as inferred from responses to a questionnaire at the end of the study. As discussed in Section 3.5, explicit learning may have been encouraged by the self-paced training paradigm used in this study. If learning was largely explicit in this study, the predictions of the discriminative learning framework are not relevant, and the findings of this study (including, if learning was indeed explicit, why were there more/better explicit learners in the prefix condition?) remain largely unclear. To elucidate this, future work should replicate Study 1 with a speeded, rather than self-paced paradigm. A prefixing advantage in that study would suggest that, in a learning context in which there is no negative evidence, prefix learning is indeed better. This would, to the best of my knowledge, be a novel contribution to the discriminative learning framework.

Therefore, in Study 1, the prefix condition was better than the suffix condition, which is contrary to what we predicted. One possibility is that the design of Study 1 involved little prediction error (in which case, as the follow-up modelling in Chapter 5 showed, a prefixing advantage may occur), and that the suffixing advantage is only observed when learning critically depends on prediction error, which was the case in Ramscar et al. (2010). With this in mind, Study 2 borrowed the design used in Ramscar et al. (2010). In particular,

participants were exposed to an artificial language in which 25% of the items which occurred with one affix had the same body shape as 75% of the items that occurred with the opposite affix. To learn the discriminating feature that groups these 25% of the items (the low type-frequency items) with the correct affix, learners had to "unlearn" body shape as a frequent, salient, but unpredictive cue. Critically, this (un)learning can only occur through prediction error. As predicted, participants in the suffix condition were able to correctly generalize both low type-frequency (LF) and high type-frequency (HF) items, whereas participants in the prefix condition were significantly better with HF items, and generalization of LF items was poor. In our analyses, this suffixing advantage was tested as an affix-by-type-frequency interaction, where evidence for the interaction means that better performance with HF items compared to LF items (type-frequency effect) is stronger in the prefix condition compared to the suffix condition (the evidence for the interaction was found in two out of three experiments in Study 2).

Study 2 provided evidence that, under such conditions where prediction error is critical for generalization – with LF items – the suffix condition is better than the prefix condition. This finding is consistent with the discriminative learning framework (and Ramscar et al. 2010 in particular). However, Study 2 did not find evidence for an overall suffixing advantage. In fact, while this was not tested statistically, the prefix condition was often numerically better than the suffix condition with HF items. This is consistent with the prefixing advantage from Study 1, which I tentatively attributed to lower prediction error in that study compared to the LF items in Study 2, where a suffixing advantage was observed. Thinking about prediction error in Study 2, body shape, a highly frequent and salient cue, though not overall fully predictive of affix usage, was predictive 75% of the time, that is, with all HF items. Learning HF items in generalization did not require the "unlearning" of body shape as an unpredictive cue via prediction error to the same extent as LF items did; in fact, body shape was a useful cue with HF items. It is therefore possible that the potential prefixing advantage with these items is due to the fact that HF generalization did not rely on prediction error to the same extent as LF learning did. One might think of HF items in Study 2 as being placed between Study 1 and LF items in Study 2 on a prediction-error continuum, where prediction error was least critical in Study 1, and most critical with LF items in Study 2. As we move along this continuum, suffixing learning, that is, learning contexts which facilitate prediction error via cue competition, become increasingly more successful, and in turn, prefixing learning becomes weaker. Future work should test this possibility in a series of experiments manipulating the amount of prediction error required. This would include a replication of Study 1 but with a speeded paradigm (to discourage explicit learning). It may also be useful to do a speeded version of Study 1 using fribbles instead of the aliens, to gain a better understanding of the amount of cue competition provided by each type of stimuli.

Studies 1 and 2 of this thesis raise the possibility that, when learning from negative evidence is not critical, label-feature learning may be better. A related point is the observation that there may be contexts in which label-feature learning is better because, in those contexts, verbal labels provide more cue competition than do features. Nixon (2018) recently

demonstrated this in an experiment with native speakers of English who were exposed to syllables from Southern Min Chinese, which were associated with simple coloured shapes – a red circle, a yellow triangle, and a blue square. Each base syllable was produced with two different tones, and it was the tone, rather than the syllable, that was reliably predictive of the shapes. For example, a red circle occurred with syllable "phe" 75% of the time, but 25% it occurred with syllable "tshe"; critically, in both cases, the tone of the syllable was rising. Therefore, to learn the label for that particular shape, participants had to ignore the more salient and more frequent cue, the syllable, and learn the less salient and less frequent cue, the tone (recall that these were native speakers of English, a non-tonal language). This design is thus analogous to Ramscar et al. (2010) and to Study 2 of this thesis, with the critical difference being that verbal labels generated greater error signal than visual referents. The labels involved tones, as well as syllables which overlapped across visual features; visual referents, on the other hand, only contained two perceptively discriminable features (shape and colour), which did not overlap across different tones. At test, participants were shown a trained shape and played three labels: for example, a red circle was shown, and alongside the syllable "tshe" with a rising tone, the two other LF labels were played. Critically, one of the two competitor labels was the syllable "phe", which occurred with the red circle 75% of the time, but the syllable was played with a low frequency tone (falling). If participants relied on frequency of co-occurrence, they should be more likely to select "phe" than "thse". Nixon found that this was the case for participants in the feature-label condition, but not in the label-feature condition; in this condition, participants were more likely to select the correct item on the basis of the tone. This result is consistent with the suffixing advantage with LF items in Study 2 – in both cases, richer cue structure provided more competition error than simpler cue structure, and this resulted in better generalization. Note that this finding cannot be attributed to the (expected) better item-learning in the label-feature condition (as the test used trained items, though produced by new speakers): if this was the case, we would expect better performance in this condition with HF items, too – but this was not the case. However, it is still true that a more appropriate test of generalization would involve presenting participants with a new visual object, perhaps the same shape but with different colour, and observing whether the participants match this shape to a HF syllable but incorrect tone, or with the correct tone but a LF syllable. This would be an interesting future project. Nixon's work shows that, rather than necessarily being the case that the feature-label order always facilitates discriminative learning, this is determined by cue structure – predicting simple outcomes from more complex cues provides more cue competition and prediction error than the other way round, which results in learning. This is consistent with the findings in this thesis, which did not observe an advantage of feature-label learning as such, but only when cue competition was critical for generalization.

Finally, while Experiments 4 and 5 of Study 2 showed evidence that suffixing leads to equally successful generalization of HF and LF items, Experiment 6 found better generalization of HF items than LF items in both affix conditions. This is contrary to the predictions of the discriminative learning framework and the discriminative learning com-

putational model presented in Experiment 3. However, in Chapter 5 I revisited this model and demonstrated that a type-frequency effect also occurs in the suffix condition in earlier stages of the learning. If the model is tested at asymptote (as it was in Experiment 3), however, the type-frequency effect disappears in the suffixing condition. In Chapter 5 I therefore made the point that, even though our predictions are based on models which are trained to asymptote, it is not possible to know where in the learning trajectory our participants are at the point when we test them. Experiment 6 contained the largest input set of all experiments – 32 items, compared to 16 in Experiment 4 and eight in Experiment 5 – but participants received the same amount of exposure in each experiment. It is possible that learning in Experiment 6 was harder compared to the other experiments due to the increased set, and that it was therefore slower. Had participants been given more exposure in this experiment, it is likely that the type-frequency effect in suffixing would have gone away. While a further experiment is necessary to test this possibility, Experiment 6 demonstrates that, even when a mathematical model is available to test predictions on, fine-tuning how this translates to human performance often requires multiple iterations of the paradigm.

To summarize, this thesis has shown that cue competition and prediction error have important consequences for learning in a series of artificial learning experiments. Ramscar et al. (2010) demonstrated that feature-label learning leads to better generalization than label-feature learning as it provides more cue competition. This thesis adds to this work by showing that, when prediction error was critical for generalization, feature-label learning was better than label feature learning. The thesis also raised the possibility that there may be learning contexts in which cue competition is not critical for generalization, and in those contexts label-feature learning may in fact better than feature-label learning. However, as I discussed above, the insight that, where there was better label-feature (prefix) learning in this thesis, it was due to cue competition not being critical for generalization, is one interpretation of the results. To incorporate it into the discriminative learning theory more formally, future work is needed to precisely manipulate the importance of cue competition across learning contexts, and observe order effects across these contexts. An important next step would involve investigating how these different learning contexts scale up to human language more broadly. It is likely that very few linguistic generalizations are perfect correlations of the kind used in Study 1 of this thesis, and that instead generalization fundamentally depends on the ability to discriminate between informative and uninformative cues (which critically depends on prediction error). Modelling different linguistic generalizations in terms of cue structure, and the extent to which different cue structures provide prediction error, may provide a parsimonious account of linguistic generalization in different learning contexts, languages, and levels of linguistic organization.

7.2 Is there a prefixing advantage in item-learning?

This thesis tested another prediction – that, while suffixing leads to better generalization, prefixing leads to better item-learning. This prediction comes from viewing language as

uncertainty reduction, both about the meaning of the message (by discriminating between informative and uninformative cues) and about the form of the message (by predicting what the speaker is most likely to say next). Specifically, applying information-theoretic approaches to language has suggested that linguistic forms self-organizesuch that information content (the entropy or uncertainty of the message) is distributed evenly across the whole utterance (A. F. Frank & Jaeger, 2008; Jaeger & Levy, 2007; Jaeger, 2010). Dye and colleagues (2017, 2018) demonstrated that prenominal gendered articles in German and prenominal adjectives in English smooth the uncertainty of the utterance over the article-noun (or prenominal adjective-noun) pair. They showed that in an English utterance *Look at the* NOUN the entropy of the noun is high; however, the entropy of the noun is reduced in *Look at the cute little* NOUN – in this context, a much smaller sub-set of nouns is likely to occur (e.g., *baby/kitten/puppy*). Therefore in the second utterance the entropy of the noun is reduced by virtue of entropy being distributed across the adjectives and the noun. This may be particularly important in language processing as nouns tend to be the largest part-of-speech group, and therefore also the group with highest uncertainty. In the context of the experiments presented in this thesis, this means that, in the prefix condition, once the learner hears the prefix, the possible number of upcoming noun+picture pairs is reduced by 50% (as half of the nouns occur with one affix, and half with the other affix). From this we predicted that better learning of both affix+noun and noun+picture mappings would occur in the prefix condition compared to the suffix condition, as only in the prefix condition was the noun+picture pair easier to process due to reduced entropy.

This hypothesis was tested in all the studies in this thesis by playing participants a trained affix+noun pair, showing them two or more trained pictures (or one trained and one novel picture), and asking them to select the picture that matches the label (an additional item-learning test was used in Study 2). In Study 1, there was evidence for the prediction – there was better item-learning in the prefix condition compared to the suffix condition. While originally this was taken as evidence for our theory, neither of the subsequent studies found consistent evidence for this. In Study 2, there was no evidence for a prefixing advantage, but overall item-learning was relatively weak in that study. This learning was attributed to the nature of the stimuli used in the study. These stimuli, *fribbles*, are particularly useful for studying "category" learning, as individual items have salient visual similarities; distinguishing individual items in the same "category" from each other, which was necessary for the item-learning test, was in retrospect consider much harder (see Section 4.7.2. for a lengthy discussion). Study 3 therefore used stimuli which were more suitable for individual item-learning; however, across all experiments in that study, no consistent evidence for a prefixing advantage was found -– there was evidence for this advantage in Experiment 8a, but this did not replicate in Experiment 8b, which was a replication study with a larger sample. Recall that our prediction, whereby the prefix can smooth the entropy over the affix-noun pair, only apply if learners learn the underlying probability distribution of nouns and affixes in the language, and generate responses according to that distribution. In Study 3, there was strong noun-picture learning, but noun-affix learning was poor, and further analyses showed no indication that participants' learning of which nouns co-occur

with which affixes is related to how well they learn the noun-picture mappings. Therefore, noun-picture mappings in Study 3 study were learned via a mechanism different to the one our theory predicts. The reasons for this, and suggestions for how future work should approach them, are discussed at length in the Discussion of Study 3.

Given the findings of Studies 2 and 3, the prefixing advantage in item-learning in Study 1 is difficult to interpret. Note that this study did not include a test of the learning of trained noun+affix pairs as the one in Study 3, and therefore we cannot test whether the prefixing advantage in that experiment was due to better learning of noun+affix pairs in that condition. A test in which the knowledge of noun+affix labels was tested in Study 1 was such that participants were presented with a trained picture, and two noun+affix labels, where the target label matched the picture, and in the foil label the affix was wrong. Therefore while the test could be done by matching nouns to correct affixes, it could also be done by matching pictures to correct affixes, and it is not possible to pull these options apart. In addition, the test could be done through item-knowledge, but also by abstracting away from individual items and mapping abstract discriminating features of trained items to the affixes (by contrast, this was not possible in Study 3, where there were no features of the nouns or pictures which consistently matched them with either affix). This test did show a prefixing advantage, but it was unclear whether this reflected better item-knowledge in that condition, or better learning of discriminating features (this is discussed in Experiment 2b in more detail).

Finally, note that in Study 1, learning was better in the prefix condition not just with respect to item-learning, but also with respect to generalization. While in Chapter 5 I identified some theoretical conditions under which a prefixing advantage in generalization is expected, it is also possible that Study 1, which was self-paced rather than timed, induced a different learning mechanism compared to Studies 2 and 3, which made learning generally better in the prefix condition. Were that the case, predictions of the discriminative learning mechanism would not apply. While this is possible, it is not clear what kind of alternative mechanism that would be, nor why it would be better in the prefix condition. To cast some light on this finding, future work should aim to: 1) do a timed version of Study 1 to test the hypothesis raised in Chapter 5, namely that the prefix condition was better in Study 1 due to there being one salient, fully consistent cue (and no need for learning from negative evidence), and 2) include the same test of noun+affix learning as the one used in Study 3. This work would help get a better understanding of the learning mechanisms which caused the prefixing advantage in Study 1, both with respect to item-learning and generalization.

To summarize, Studies 2 and 3 found no evidence for better item-learning in the prefix condition, while several aspects of Study 1 make it difficult to interpret better item-learning in the prefix condition in this study. On the whole, therefore, I conclude that this thesis did not provide (clear) evidence that prefixes smooth the entropy over the prefix-noun pair and make the noun easier to learn.

7.3 Methodological contribution of the thesis

This thesis has two key methodological contributions.

First, most of the experiments in this thesis were replicated, with pre-registered hypotheses and analyses plans. This is particularly important in light of the "replication crisis", whereby researchers have not been able to replicate many published effects in psychological literature, which is in part caused by the original studies being underpowered (see Section 2.2 for detail). Pre-specifying our predictions means that any results that were contrary to the prediction can be discussed and interpreted, but cannot be accredited to the theory in a post-hoc manner. In this thesis, the importance of replication was particularly striking in Experiment 8. Here, the predicted effect was observed in Experiment 8a, but the effect went away (with evidence for the null) in a replication experiment with a larger sample (Experiment 8b), suggesting that the finding of Experiment 8a may have been Type I error.

Second, with respect to statistical analysis, the thesis used an innovative method of modelling the H1 estimate in cases where there are no past data available – this particular issue has made many researchers reluctant to incorporate Bayes factors into their analyses. Turning to Bayes factors may be particularly valuable for researchers who work with children and infants, but also in studies of statistical learning more generally, where data can be noisy due to a lack of robust measures (Siegelman et al., 2017), as unlike traditional p-values, Bayes factors allow the researcher to distinguish between noisy findings and actual evidence for the null. Where evidence for the null is found, this is theoretically relevant; where, on the other hand, ambiguous evidence is found, this may help researchers identify aspects of their paradigm which are picking up noise (as was the case in this thesis on several occasions). Furthermore, the choice of H1 was almost always pre-registered, thus eliminating the possibility of experimenter bias – for example, picking an H1 after receiving the data, with the goal of getting a Bayes factor greater than 3. The use of robustness regions in this thesis also contributes to the transparency of the approach, as others have raised concerns about the effect of the choice of H1 on the resulting inference, but the robustness regions allow readers to evaluate for themselves how the choice of the H1 affects the inferences that were drawn. In the domain of artificial language learning, to the best of my knowledge, only two other studies used Bayes factors (Samara et al., 2019; Wonnacott et al., 2017). My hope is that this thesis encourages more researchers in the area of (child) language learning to replicate studies, and that they find the statistical approach in this thesis valuable in their own work.

7.4 Limitations and future directions

Throughout the thesis, limitations of specific designs and potential useful areas for future work have been identified. Here I address some more general limitations of the work presented in the thesis and broader implications for future work.

One possible concern in Studies 1 and 2, which looked at generalization, is that the timing of stimuli was somewhat unnatural, in that pictures were only present on-screen

while nouns were played with a blank screen showing during the affix. However, this was necessary so that in the prefix condition the semantic features available to the noun were not already present – and thus available to act as "cues" for the prefix – until the noun was actually encountered (the timing in the suffix condition was designed to be as closely matched as possible to the prefix condition). It is important to note, however, that other work which found support for the key discriminative learning principles (Ramscar, 2013) did not include this temporal separation of pictures and affixes. While such timing was theoretically motivated in this proof-of-concept work, future work should aim to develop more naturalistic learning paradigms.

Another limitation of this work is that the tests of learning largely involved "offline" (post-exposure) tests, with no information about learning as it unfolded in real time. This is particularly important in Study 2, where the pattern of results is interpreted as supporting the discriminative learning theory, however, this in only indirectly inferred from scores on forced-choice tests. In theory, it is possible that some other learning mechanism yielded the same pattern of results. Future work should therefore adapt the existing paradigm to use online measures which can be a more direct indication of learning. 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; Hyönä, Tommola, & Alaja, 1995; Just & Carpenter, 1993). The prediction here would be that, if the hypothesis is correct and prediction error is the driving mechanism for learning in the suffix condition, greater pupil dilation should be expected on LF trials than HF trials, as these generate greater error signal. This would provide more direct support for the theory, and it would also inform the dynamics of "unlearning" uninformative cues (currently this was possible to observe in the computational simulations but not with human learners). Developing more implicit measures may also allow to adapt the current paradigm for use with child learners. This is important since ultimately the interest is in first language acquisition, and given evidence that adult and child language learning may be different (Culbertson et al., 2017, 2019; Hudson Kam & Newport, 2005).

The only experiment in which trial-by-trial data were collected is Experiment 9, in which a cross-situational learning paradigm was used (Yu & Smith, 2007). The advantage of this paradigm is that a datapoint is obtained at every trial – thus it was possible to track learning over time. However, here as well the main measure was accuracy (reaction time with mouse clicks was a secondary measure and is potentially less precise than key presses, which were deemed inappropriate for a 3-alternative forced-choice test). As the effect of prefixing on item-learning may be subtle, especially given that the nouns and the pictures were already highly learnable in their own right, it is possible that a more implicit measure such as eye tracking would reveal an indicative pattern – for example, it is possible that participants in the prefix condition would show reduced looks to the opposite-affix competitors once the prefix has been played. This is particularly likely given that this effect has been reported with speakers exposed to similar prenominal cues from their first language. However, it may be possible that affixes and nouns which are longer in duration would be required to capture this effect. This adaptation would be fairly easy to

implement in the existing paradigm, and future work should aim to do this. To summarize, I have identified several important directions for future work that could help clarify the validity and robustness of the results presented in this thesis, as well as elucidate the places where no evidence for the theory was observed. In the meantime, these results provide in principle evidence that cue competition over features of the noun can lead to a different patterns of learning in suffixing compared to prefixing.

Finally, a necessary next step would be to investigate more systematically how discriminative learning feeds into higher-level cognitive processes which shape language learning, such as social cognition, for example. Here, it is necessary to integrate discriminative learning with existing findings about the way humans learn language from social cues. Given the findings of this thesis, as well as previous work, it is likely that discriminative learning would provide a unified account of the lower-level learning that underpins these kinds of cognitive mechanisms. While not in the focus of this thesis, this question has important implications for broad theories of human learning and cognition.

7.5 Conclusion

This thesis tested the predictions of a novel theoretical account of language learning and processes, the discriminative learning framework (see Ramscar et al., 2010), which views language learning as a process of uncertainty reduction by discriminating informative from uninformative cues. The advantage of this approach over other existing approaches is that it models the underlying learning mechanism precisely, by returning to the basic principles of classical learning theory – cue competition and prediction error. This approach makes specific predictions about the effect of the order of learning events on linguistic generalization and item learning, and these predictions were tested in a series of artificial language learning experiments. This thesis found support for the discriminative learning framework with respect to generalization but not item-learning, which is discussed in terms of the limitations of the existing experimental paradigm. On the whole, the thesis provided a further test of discriminative learning as a precise theory of language learning, and identified some fruitful avenues for future research.

References

Abbot-Smith, K., Lieven, E., & Tomasello, M. (2001). What preschool children do and do not do with ungrammatical word orders. *Cognitive Development*, *16*(2), 679–692.

Abney, S. (1996). Statistical methods and linguistics. *The balancing act: Combining symbolic and statistical approaches to language*, 1–26.

Aguado-Orea, J., & Pine, J. M. (2015). Comparing different models of the development of verb inflection in early child spanish. *PLoS ONE*.

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language*, *26*(2), 339–356.

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development*, *67*(2), 635–645.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*.

Ambridge, B. (2010). Children's judgments of regular and irregular novel past-tense forms: New data on the english past-tense debate. *Developmental Psychology*.

Ambridge, B. (2018). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*.

Ambridge, B., & Lieven, E. V. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge: Cambridge University Press.

Ambridge, B., Pine, J. M., & Rowland, C. F. (2012). Semantics versus statistics in the retreat from locative overgeneralization errors. *Cognition*, *123*(2), 260–279.

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

Annau, Z., & Kamin, L. J. (1961). The conditioned emotional response as a function of intensity of the us. *Journal of Comparative and Physiological Psychology*, *54*(4), 428.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 1–20.

Arnon, I., & Clark, E. V. (2011). Why brush your teeth is better than teeth - children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*.

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.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*.

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324.

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.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390–412.

Baayen, R. H., Milin, P., Djurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438.

Baguley, T., & Kaye, W. (2010). Review of: Understanding psychology as a science: An introduction to scientific and statistical inference, by z. dienes. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 695–698.

Bailey, T. M., & Hahn, U. (2001). Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language*.

Baldwin, D. A. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental psychology*, *29*(5), 832.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations: Research article. *Psychological Science*. doi: 10.1111/j.1467-9280.2008.02075.x

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beckman, M. E., & Edwards, J. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development*.

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.

Berko, J. (1958). The Child's Learning of English Morphology. *WORD*.

Blything, R. P., Ambridge, B., & Lieven, E. V. (2018). Children's Acquisition of the English Past-Tense: Evidence for a Single-Route Account From Novel Verb Production Data. *Cognitive Science*.

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in neuroscience*, *12*, 152.

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.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, *138*(3), 389.

Braine, M. D. (1963). On learning the grammatical order of words. *Psychological Review*. doi: 10.1037/h0047696

Brooks, P. J., Braine, M. D., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, *32*(1), 76–95.

Brooks, P. J., & Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, 720–738.

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

Brown, H., Weighall, A., Henderson, L. M., & Gaskell, M. G. (2012). Enhanced recognition and recall of new words in 7-and 12-year-olds following a period of offline consolidation. *Journal of experimental child psychology*, *112*(1), 56–72.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form* (Vol. 9). John Benjamins Publishing.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Bybee, J., & Moder, C. L. (1983). Morphological Classes as Natural Categories. *Language*.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered reports at aims neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4–17.

Chomsky, N. (1957). *Syntactic structures*. Walter de Gruyter.

Chomsky, N. (1959). *Chomsky, n. 1959. a review of bf skinner?s verbal behavior. language, 35 (1), 26–58.* JSTOR.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT press.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*(1), 1–15.

Chomsky, N. (1994). *Bare phrase structure* (Vol. 8). MIT Press Cambridge (MA).

Chomsky, N. (1995). *The minimalist program*. MIT press.

Chomsky, N., & Lasnik, H. (1993). Principles and parameters theory. *Syntax: An international handbook of contemporary research*.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and brain sciences*, *31*(5), 489–509.

Clark, E. (2007). Morphology in language acquisition. In A. Spencer & A. Zwicky (Eds.), *The handbook of morphology* (pp. 374–89). Oxford: Blackwell.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior*, *12*(4), 335–359.

Coleman, J., & Pierrehumbert, J. B. (1997). Stochastic phonological grammars and acceptability. *Computational phonology Third meeting of the ACL special interest group in computational phonology*. doi: 10.3109/13682820109177934

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, *14*(4), 597–612.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132.

Croft, W. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

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.

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*(2), 268–293.

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82.

Culbertson, J., & Schuler, K. (2019). Artificial Language Learning in Children. *Annual Review of Linguistics*, *5*, 353–373.

Dabrowska, E. (2008). The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. *Journal of Memory and Language*.

Dahan, D., Swingley, 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.

Daugherty, K., & Seidenberg, M. (1992). Rules or connections? the past tense revisited. In *Proceedings of the fourteenth annual conference of the cognitive science society* (pp. 259–264).

de Heide, R., & Grünwald, P. D. (2017). Why optional stopping is a problem for bayesians. *arXiv preprint arXiv:1708.08278*.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Macmillan International Higher Education.

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*, 781.

Dienes, Z. (2016). How bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of mandarin lexical tones. *PeerJ*, *7*, e7191.

Dryer, M. S., & Haspelmath, M. (2013). *Wals online*. Leipzig: Max Planck Institute for

Evolutionary Anthropology. Retrieved from `http://wals.info/`

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.

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 (icslp 96)* (Vol. 3, pp. 1393–1396).

Dye, M., Milin, P., Futrell, R., & Ramscar, M. (2017). A functional theory of gender paradigms. In *Perspectives on morphological organization* (pp. 212–239). Brill.

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.

Edwards, J., Beckman, M. E., & Munson, B. (2004). The Interaction between Vocabulary Size and Phonotactic Probability Effects on Children's Production Accuracy and Fluency in Nonword Repetition. *Journal of Speech, Language, and Hearing Research*.

Egger, M. D., & Miller, N. E. (1962). Secondary reinforcement in rats as a function of information value and reliability of the stimulus. *Journal of Experimental Psychology*, *64*(2), 97.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, *24*(2), 143–188.

Ellis, N. C. (2006). Selective attention and transfer phenomena in l2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194.

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

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, *33*(4), 547–582.

Elman, J. L., Bates, E. A., & Johnson, M. H. (1998). *Rethinking innateness: A connectionist perspective on development* (Vol. 10). MIT press.

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*, 30-69.

Erickson, L. C., & Thiessen, E. D. (2015). *Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition.*

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.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? an empirical support

from us states data. *PloS one*, *5*(4).

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158–180.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science?s aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant behavior and development*, *10*(3), 279–293.

Finch, S., & Chater, N. (1992). Bootstrapping syntactic categories. In *Proceedings of the sixteenth annual meeting of the cognitive science society* (p. 820-825). Bloomington, IN.

Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, *7*(6), 585–594.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30).

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science*, *5*(3), 475–494.

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.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language*, *39*(2), 218–245.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of Wordlikeness: Effects of Segment Probability and Length on the Processing of Nonwords. *Journal of Memory and Language*, 481-496.

Frisch, S. A., Large, N. R., Zawaydeh, B. A., & Pisoni, D. B. (2001). Emergent phonotactic generalizations in English and Arabic..

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*(12), 1128.

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341.

Gagliardi, A., & Lidz, J. (2014). Statistical insensitivity in the acquisition of tsez noun classes. *Language*, 58–89.

Gallistel, C. R. (2002). Frequency, contingency and the information processing theory of conditioning. *Frequency processing and cognition*, 153–171.

Gambi, C., Pickering, M. J., & Rabagliati, H. (2016). Beyond associations: Sensitivity to structure in pre-schoolers? linguistic predictions. *Cognition*, *157*, 340–351.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 110.

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

Gathercole, S. E., Frankish, C. R., Pickering, S. J., & Peaker, S. (1999). Phonotactic Influences on Short-Term Memory. *Journal of Experimental Psychology: Learning Memory and Cognition*.

Gathercole, S. E., Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental psychology*. doi: 10.1037/0012-1649.33.6.966

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological Memory and Vocabulary Development During the Early School Years: A Longitudinal Study. *Developmental Psychology*.

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.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1), 3–55.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: The quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3-4), 305–320.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Goodman, N. D., Frank, M. C., Griffiths, T. L., Tenenbaum, J. B., Battaglia, P. W., & Hamrick, J. B. (2015). Relevant and robust: A response to marcus and davis (2013). *Psychological Science*, *26*(4), 539–541.

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

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the bayesians got their beliefs (and what those beliefs actually are): comment on bowers and davis (2012).

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology*, *41*(4), 313–360.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review*, *106*(3), 491.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, *111*(3), 662.

Havron, N., de Carvalho, A., Fiévet, A.-C., & Christophe, A. (2019). Three-to four-year-old children rapidly adapt their predictions and use them to learn novel word meanings. *Child development*, *90*(1), 82–90.

Hawkins, J. A., & Gilligan, G. (1988). Prefixing and suffixing universals in relation to basic word order. *Lingua*, *74*(2-3), 219–259.

Hay, J. F., Pelucchi, B., Estes, K. G., & Saffran, J. R. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106.

Hebb, D. O. (1949). *The organization of behavior*.

Henderson, L. M., Weighall, A. R., Brown, H., & Gaskell, G. M. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental science*, *15*(5), 674–687.

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a zipfian environment. *Cognition*, *189*, 11–22.

Hoffman, M. F., & Walker, J. A. (2010). Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change*.

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior research methods*, *48*(4), 1393–1409.

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.

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*(6), 876–909.

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.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jaeger, T. F., & Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849–856).

Jeffreys, H. (1961). *Theory of probability*. London: Oxford University Press.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences*, *96*(13), 7592–7597.

John, M. F. S., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*(1-2), 217–257.

Jones, L. (1967). English phonotactic structure and first language acquisition. *Lingua*, *19*, 1-59.

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*, 169.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language*, *33*(5), 630–645.

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.

Kam, X. N. C., Stoyneshka, I., Tornyova, L., Fodor, J., & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive Science*.

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: Miami University Press.

Karmiloff-Smith, A. (1981). *A functional approach to child language: A study of determiners and reference* (Vol. 24). Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Konorski, J. (1948). Conditioned reflexes and neuron organization.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*(3), 439–453.

Labov, W. (1966). The social stratification of english in new york city.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*.

Lany, J., & Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development. *Developmental Science*, *14*(5), 1207–1219.

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

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.

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.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science*, *14*(6), 1323–1329.

Lieven, E. V., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of child language*, *24*(1), 187–219.

Lightfoot, D. (1991). *How to set parameters: Arguments from language change*. MIT Press.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*.

Mackintosh, N. (1965). Selective attention in animal discrimination learning. *Psychological Bulletin*, *64*(2), 124.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk (third edition): Volume I: Transcription format and programs, Volume II: The database. *Computational Linguistics*.

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

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542.

Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science*.

Marchman, V. A. (1997). Children's productivity in the english past tense: The role of frequency, phonology, and neighborhood structure. *Cognitive Science*.

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language*, *21*(2), 339–366.

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

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the society for research in child development*, i–178.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121–157.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*.

Masicampo, E., & Lalande, D. R. (2012). A peculiar prevalence of p values just below. 05. *The quarterly journal of experimental psychology*, *65*(11), 2271–2279.

Maslen, R. J., Theakston, A. L., Lieven, E. V., & Tomasello, M. (2004). A dense corpus study of past tense and plural overregularization in english. *Journal of Speech, Language, and Hearing Research*.

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed

speech. *Cognitive Science*.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2005). The role of frequency in the acquisition of english word order. *Cognitive Development*, *20*(1), 121–136.

Matthews, D. E., & Theakston, A. L. (2006). Errors of omission in English-speaking children's production of plurals and the past tense: The effects of frequency, phonology, and competition. *Cognitive Science*.

Maurits, L., Navarro, D., & Perfors, A. (2010). Why are some word orders more common than others? a uniform information density account. In *Advances in neural information processing systems* (pp. 1585–1593).

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does ?failure to replicate? really mean? *American Psychologist*, *70*(6), 487.

McClelland, J. L., & Rumelhart, D. E. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, *2*, 216–271.

McCloskey, M. (1991). Networks and Theories: The place of connectionism in cognitive science. *Psychological Science*, 387-395.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of M athematical Biophysics*, *5*(4), 115–133.

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.

Milin, P., Divjak, D., Dimitrijević, S., & Baayen, R. H. (2016). Towards cognitively plausible data science in language research. *Cognitive Linguistics*, *27*(4), 507–526.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, H. (2017). Discrimination in lexical decision. *PloS one*.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition*, *30*(5), 678–686.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*. doi: 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*(1-2), 139–167.

Mirković, J., Seidenberg, M. S., & Joanisse, M. F. (2011). Rules versus statistics: Insights from a highly inflected language. *Cognitive Science*, *35*(4), 638–681.

Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal ofAutism and Developmental Disorders*, *20*(1), 115–128.

Munson, B. (2001). Phonological Pattern Frequency and Speech Production in Adults and Children. *Journal of Speech, Language, and Hearing Research*. doi: 10.1044/

1092-4388(2001/061)

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*(4), 329–336.

Nelson, D. G. K., Jusczyk, P. W., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant behavior and development*, *18*(1), 111–116.

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, et al. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468.

Nixon, J. S. (2018). Effective acoustic cue learning is not just statistical, it is discriminative. In *Interspeech* (pp. 1447–1451).

Nosek, B. A., & Lakens, D. (2014). *Registered reports.* Hogrefe Publishing.

OpenScienceCollaboration, et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Pavlov, I. P. (1927). Conditioned reflexes. *London: Oxford*.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, *80*(3), 674–685.

Pereira, F. (2000). Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *358*(1769), 1239–1253.

Pérez-Pereira, M. (1991). The acquisition of gender: What spanish children tell us. *Journal of Child Language*, *18*(3), 571–590.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *Journal of Child Language*, *37*(3), 607–642.

Perruchet, P., & Desaulty, S. (2008). A role for backward transitional probabilities in word segmentation? *Memory & cognition*, *36*(7), 1299–1305.

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological science*, *21*(12), 1894–1902.

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.

Pine, J. M., & Lieven, E. V. (1997). Slot and frame patterns and the development of the determiner category. *Applied psycholinguistics*, *18*(2), 123–138.

Pinker, S. (2013). *Learnability and cognition: The acquisition of argument structure*. MIT press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*.

Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science*.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2005). Articulatory planning is continuous

and sensitive to informational redundancy. *Phonetica*, *62*(2-4), 146–159.

Potter, C. E., & Saffran, J. R. (2017). Exposure to multiple accents supports infants? understanding of novel accents. *Cognition*, *166*, 67–72.

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

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*.

R, C. T., et al. (2013). R: A language and environment for statistical computing.

Ramscar, M. (2010). Computing machinery and understanding. *Cognitive Science*, *34*(6), 966–971.

Ramscar, M. (2013). Suffixing, prefixing, and the functional order of regularities in meaningful strings. *Psihologija*, *46*(4), 377–396.

Ramscar, M. (2019). Source codes in human communication. *arXiv preprint arXiv:1904.03991*.

Ramscar, M., & Dye, M. (2011). Learning language from the input: Why innate constraints can?t explain noun compounding. *Cognitive Psychology*, *62*(1), 1–40.

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

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of "mouses" in adult speech. *Language*, 760–793.

Ramscar, M., & Robert, F. (2015). Categorization (without categories). In E. Dabrowska & D. Divjak (Eds.), *Handbook of cognitive linguistics* (p. 75-99). Berlin: De Gruyter.

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.

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.

Räsänen, S. H., Ambridge, B., & Pine, J. M. (2016). An Elicited-Production Study of Inflectional Verb Morphology in Child Finnish. *Cognitive Science*.

Raviv, L., & Arnon, I. (2018). Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning. *Cognition*, *181*, 160–173.

Reber, A. S. (1967). Implicit Learning of Artificial Grammars. *Journal of Verbal Learning and Verbal Behavior*, 855-863.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

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.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). Distributional learning of subcate-

gories in an artificial grammar: Category generalization and subcategory restrictions. *Journal of Memory and Language*.

Rescorla, R. A. (1967). Pavlovian conditioning and its proper control procedures. *Psychological Review*, *74*(1), 71-80.

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.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist*, *43*(3), 151.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Reuter, T., Emberson, L., Romberg, A., & Lew-Williams, C. (2018). Individual differences in nonverbal prediction and vocabulary size in infancy. *Cognition*, *176*, 215–219.

Reynolds, G. S. (1961). Behavioral contrast. *Journal of the Experimental Analysis of Behavior*, *4*(1), 57–71.

Rodd, J. (2017). Lexical ambiguity. *Oxford handbook of psycholinguistics*.

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.

Roeper, T., & Williams, E. (1987). *Parameter setting* (Vol. 4). Springer Science & Business Media.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Rouder, J. (2019). On the interpretation of bayes factors: A reply to de heide and grunwald.

Rouder, J. N. (2014). Optional stopping: No problem for bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308.

Rubino, R. B., & Pine, J. M. (1998). Subject-verb agreement in Brazilian Portuguese: What low error rates hide. *Journal of Child Language*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 533-536.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*.

Samara, A., Singh, D., & Wonnacott, E. (2019). Statistical learning and spelling: Evidence from an incidental learning experiment with children. *Cognition*, *182*, 25–30.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*.

Savičiute, E., Ambridge, B., & Pine, J. M. (2018). The roles of word-form frequency and phonological neighbourhood density in the acquisition of Lithuanian noun morphology. *Journal of Child Language*.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Schwab, J. F., Casey, L.-W., & Goldberg, A. E. (2018). When regularization gets it wrong: Children over-simplify language input only in production. *Journal of Child Language*, *45*(5), 1054–1072.

Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science*.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive science*, *23*(4), 569–588.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning*, *7*(2-3), 161–193.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shannon, C. E. (1956). The bandwagon. *IRE Transactions on Information Theory*, *2*(1), 3.

Siegel, S., & Allan, L. G. (1996). The widespread influence of the Rescorla-Wagner model. *Psychonomic Bulletin and Review*.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*.

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining ?learning? in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive science*, *42*, 692–727.

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.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444–449.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: Mediated transfer or position learning? *Journal of Experimental Psychology*, 580-588.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain*

Sciences, *11*(1), 1–23.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology*.

Stavrakaki, S., & Clahsen, H. (2009). The perfective past tense in Greek child language. *Journal of Child Language*.

St Clair, M. C., Monaghan, P., & Ramscar, M. (2009). Relationships between language structure and language learning: The suffixing preference and grammatical categorization. *Cognitive Science*, *33*(7), 1317–1329.

Steedman, M. (2002). Connectionist and symbolic representations of language. In *Encyclopedia of cognitive science.*

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance?or vice versa. *Journal of the American statistical association*, *54*(285), 30–34.

Taft, M. (1988). A morphological-decomposition model of lexical representation. *Linguistics*, *26*(4), 657–668.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, *28*(1), 127–152.

Thothathiri, M., & Snedeker, J. (2008). Syntactic priming during language comprehension in three-and four-year-old children. *Journal of Memory and Language*, *58*(2), 188–213.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*(3), 209–253.

Tomasello, M., & Haberl, K. (2003). Understanding attention: 12-and 18-month-olds know what is new for other persons. *Developmental Psychology*, *39*(5), 906.

Trueswell, J. C., & Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of Memory and Language*. doi: 10.1006/jmla.1998.2565

Trueswell, J. C., Sekerina, I., Hill, N. M., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*. doi: 10.1016/S0010-0277(99)00032-3

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-Specific Constraints in Sentence Processing: Separating Effects of Lexical Preference From Garden-Paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10 .1037/0278-7393.19.3.528

Twilley, L. C., & Dixon, P. (2000). Meaning resolution processes for words: A parallel independent model. *Psychonomic Bulletin and Review*. doi: 10.3758/BF03210725

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*.

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.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and Syllable Stress: Implications for the Processing of Spoken Nonsense Words. *Language and Speech*, *40*(1), 47–62.

Vouloumanos, A., & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental psychology*, *45*(6), 1611.

Wagner, A. R. (1969). Stimulus selection and a ?modified continuity theory.? *The Psychology of Learning and Motivation*, *3*, 1–41.

Wagner, A. R., Logan, F. A., & Haberlandt, K. (1968). Stimulus selection in animal discrimination learning. *Journal of Experimental Psychology*, *76*(2p1), 171.

Waxman, S. R., Lidz, J. L., Braun, I. E., & Lavin, T. (2009). Twenty four-month-old infants? interpretations of novel verbs and nouns in dynamic scenes. *Cognitive Psychology*, *59*(1), 67–95.

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua*, *106*(1-4), 23–79.

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), 1–14.

Wonnacott, E. (2013). Learning: Statistical mechanisms in language acquisition. In *The language phenomenon* (pp. 65–92). Springer.

Wonnacott, E., Boyd, J. K., Thomson, J., & Goldberg, A. E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language*, *66*(3), 458–478.

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.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*(3), 165–209.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in cognitive sciences*, *8*(10), 451–456.

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

Zubin, D. A., & Köpcke, K. (1981). Gender: A less than arbitrary grammatical category. In *Papers from the regional meeting of the chicago linguistic societyl.*

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".

1. 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?

2. **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)?

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

4. **If you did notice any patterns, did you notice any exceptions to the patterns?**

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