Dissertation in part-fulfilment of the Master in Science
in Applied Linguistics and Second Language Acquisition (MSc ALSLA)

The role of prediction error in linguistic generalization and item learning: evidence from an online eye-tracking study

Haoyu Zhou

Word Count: 19212



Do androids dream of electric sheep?
Perhaps real sheep too. Or just sheep. Sheep is sheep.

献给我的姥姥，张学兰

Acknowledgements

While moving on to typing this dissertation out from doing the experiment (which I found very engaging and fulfilling, thanks to the great ideas and knowledge of Liz and Eva), I have been thinking about this one question for the past two months – why cannot writing be simply just writing. Why does writing have to come together with the constant struggles of digging ideas from the void, forming arguments to persuade someone, phrasing sentences to try expressing everything clearly, type, delete and repeat. It seems this is quite a negative experience for me then, if it is full of struggles, full of tears. However, I am grateful that I can get on this journey to complete a piece of writing, at least to some extent of completion, with the help of many people along the way.

I want to express great gratitude to my supervisors, Liz Wonnacott and Eva Viviani, for their patience in answering every question and need during this academic year, their guidance to my research and dissertation writing, and their timely help for solving any unexpected problems with me together. I have learned so much from working with them and their ideas would continuously inspire me in life no matter I proceed in academia or not.

I want to also thank my fellow peers, who have been my all-time encouragement and reliance: Tianyi, Coco, Edihanson Chen, the Last Supper, Natalie, Shuman, and Tse noodle pals, Bella, the Cornwall trip, Chenchen, Jianrui Pei, Nina Bleach Session and my household in 4 Ship Street. My participants as well, thanks for your time and dedication which made eye tracking work online.

At last, the love, understanding, and support from my mom and dad is nothing I can ever repay. We learn with negative evidence but this journey becomes way less daunting for me than it could be knowing you would always be by my side.

Contents

Abstract…… 5

List of Figures…….6

1. Introduction…… 7

2. Literature Review…….... 11

2.1 Different theoretical perspectives on language acquisition…… 11

2.2 The statistics in the language environment…… 15

2.3 The discriminative learning mechanism……19

2.4 Learning through predictions…… 25

2.5 The current study…….. 27

3. Experiment 1 – fribbles…… 30

3.1 Methods…… 30

3.1.1 Participants…….…… 30

3.1.2 Materials and design…….…. 30

3.1.3 Tracking eye-movements online…… 32

3.1.4 Procedure……... 33

3.2 Results…….35

3.2.1 Offline AFC test……. 36

3.2.2 Eye-tracking data……37

3.3 Discussion……... 44

4. Experiment 2 – cross-situational learning…….. 49

4.1 Methods…….49

4.1.1 Participants…… 49

4.1.2 Materials and design…….. 50

4.1.3 Procedure……51

4.2 Results…… 54

4.2.1 3AFC generalization…….. 54

4.2.2 3AFC pairing……..56

4.2.3 Eye-tracking data……58

4.3 Discussion…….. 62

5. General Discussion…… 66

5.1 Individual differences……..68

5.1.1 The current study……68

5.1.2 Adult L2 learning……71

5.2 Limitations…….. 72

5.2.1 The artificial language setting…….72

5.2.2 Online eye tracking……73

5.2.3 Sample size……74

6. Conclusion…… 75

7. References……...78

8. Appendices……..87

Abstract

The discriminative learning framework, grounded in principles of learning theory, was recently revitalized as a specific mechanism to carry out language acquisition within a domain-general, input-driven constructivist account. It specifies that language learning, rather than being guided by an innate linguistic system like Universal Grammar, could be constrained by obtaining negative evidence using prediction errors. Previous studies have tried to indirectly show the effect of prediction making on word-referent mapping by comparing the learning performance of two groups of learners who experienced language input with different input order conditions, where only one condition allowed cue competition. The current study applied eye tracking to the learning process and demonstrated direct evidence for the predictive behaviors during discriminative learning. Two learning experiments with a learning phase where cues predicted visual linguistic outcomes (labels) were conducted. Experiment 1 was a replication of an earlier word learning study – with eye tracking added – and showed that participants made predictive saccades towards positions of labels before they appeared for both high and low frequency conditions. Experiment 2 used a new cross-situational learning paradigm and did not demonstrate default fixations. Offline results showed weaker learning than previous studies and did not show above-chance accuracy for the low frequency trials. However, plot analysis of the individual variation for low frequency trials suggested that participants who failed to learn low frequency cues could have been persistently overgeneralizing in training. This overgeneralization pattern is discussed in the context of adult L2 learning.

List of Figures

1. The category structures employed in experiment 1 of Ramscar et al. (2010).

Experiment 1:

2. Demonstration of the stimuli set.
3. Demonstration of set up and timelines of the learning phase.
4. Demonstration of the procedure.
5. Demonstration of website interfaces during the 4AFC generalization test.
6. Proportion of correct trials in the AFC generalization test split by frequency conditions.
7. Proportion of looks to each region of interest in the fribble experiment split by frequency.
8. Proportion of looks to each region of interest split by learner group and experiment stage
9. The three competing regions of fixations following the low frequency fribble of label “dep”.

Experiment 2:

10. Demonstration of the stimuli set.
11. Procedure of Experiment 2.
12. Learning phase set up and timelines.
13. Interfaces for the test phase tasks.
14. Proportion of correct trials in the 3AFC generalization test split by frequency and session.
15. Proportion of correct trials in the 3AFC pairing test split by frequency and session.
16. Proportion of looks to each region of interest split by frequency and session.

1. Introduction

The logical problem of language acquisition is the claim that the environment input is too impoverished, underdetermined, and diverse to allow for children's internationalization of a full grammar system (Baker, 1979; Chomsky, 1959; 1980). Therefore, humans must be equipped with a biologically endowed linguistic system so that the development process could be guided and constrained (Chomsky, 1995). This innate account of language acquisition recognizes language as a special faculty within human bodies and constrains on what to/not to learn universally, providing an explanation for why we all converge on the same set of grammatical structures despite different input backgrounds. However, the nativist account has been criticized for ignoring individual variation in language development and in the lack of details as to the genetic constructions of the universal grammar system.

An alternative account, though, can be derived from the constructivist's view, which takes language as a result of knowledge construction by abstracting patterns (i.e. form-meaning pairs) from the linguistic input with domain-general mechanisms (Goldberg, 1995). The account is input-driven and thus would recognize the effect of different input backgrounds on the linguistic knowledge developed by learners. In addition, the ability to learn environmental patterns can be well corroborated with evidence of statistical learning, which has shown that both infants and adults are sensitive to the distributional information of linguistic structures in input so that they can conduct speech segmentation and generalize grammar to other new circumstances (e.g. Gomez; 2002; Saffran et al., 1996; Thompson and Newport, 2007). Though artificial language learning is a typical experimental design for these experiments (Gomez & Gerken, 2000), the positive results are not exclusive to these settings. Studies using a natural language setting for second language (L2) learners (Treffers-Daller & Calude, 2015; Treffers-Daller & Ziyan, 2016) or an unfamiliar language for monolingual infants (Pelucchi et al., 2009) have both demonstrated the effect of input exposure to their use or segmentation of words. These studies suggested that humans

have the ability to process probabilistic information in the environment to learn linguistic structures at different levels, which support the input-driven and domain-general account to language acquisition of the constructivist approach.

To further account for the exact learning process based on input processing of the linguistic environment, many connectionists have applied computational modellings (e.g. Elman, 1990; Rumelhart & McClelland, 1986). Among these models, the discriminative learning framework (as introduced by Ramscar & Yarlett, 2007; DL) explains the process of form-meaning mapping without the help of external corrections (c.f. Rumelhart & McClelland, 1986) or internal linguistic knowledge (i.e. UG). It proposes that language learning should be understood as a process of discriminating the *contexts* in which particular linguistic forms are expected. As such, the learning process is discriminative, error-driven and consist of cue competition:

"discriminative": The theory recognizes the complexity of the learning environment and believes that language knowledge is derived by discriminating informative cues from uninformative cues in the larger sets of input. This contrasts with the idea that relevant cues are represented in a universal grammar system which is genetically given.

"error-driven": The logical problem of language acquisition stresses that the input is underdetermined and learners therefore require negative evidence to constrain learning. In DL, negative evidence is provided from the discrepancies between what is expected and what is observed.

"cue competition": The discriminative process is then carried out using the mechanisms of uncertainty reduction, whereby all available cues compete to predict the outcomes of possible linguistic forms. Specifically, the value of cues increases when they predict an outcome correctly, and this occurs at the cost of other competing cues that lose predictive values. Likewise, the value will decrease if the cue fails to

predict an outcome – a prediction error occurs, but the other competing cues would be added more values in the meantime. Consequently, the informative cues will be distinguished by accumulating their values in predicting correct outcomes, while the competing cues lose their predictive values for being uninformative. The uninformative cues would become disassociated eventually from outcomes for which they are irrelevant.

Note that an important mechanism in the DL process is prediction making, which emphasizes that negative evidence could only be obtained under an input order where learners can make predictions with the competing cues and encounter errors. Ramscar et al. (2010), Nixon (2020), and Vujović et al. (2021) have all investigated the effect of input order by comparing the learning performances in generalization tests following two input conditions. In one ordering, the complex set of features (cues) were encountered first and then the outcomes (cue-outcome). In the other, the outcomes were presented first and then the features (outcome-cue). In this latter condition, the theory predicts that learners should not be able to make predictions with the possible cues and thus cue competition cannot occur. The participants performed as expected and could only generalize appropriately after the exposure of training in the cue-outcome condition.

Although the researchers conducting these experiments assumed the better learning performance should be a result of learners making predictions about a stimulus with competing cues and expecting to see an outcome, this predictive behavior was not observed directly. Thus, the current study applied eye-tracking methods in the training sessions to try to provide on-line evidence of prediction making for DL. Primarily, the study aimed to replicate the results of cue-outcome condition in Ramscar et al. (2010), in which participants viewed complex visual referents (cues) directly followed by one of three labels (outcomes). We tracked eye-movements between referent offset and label onset to see if participants looked at the positions of target label *before* they appeared on the screen. This would indicate that predictions

were made based on the cues in the objects they have seen. Overall, two experiments were conducted to adopt two different learning paradigms – a direct replication of Ramscar et al. (2010) and a conceptual replication in a new cross-situational learning paradigm (Yu & Smith, 2007; CL). The second paradigm presents individual objects separately in a scene as competing cues and this make it potentially possible to also have an eye-tracked outcome-cue condition. However, for the current experiment, we only attempted the first step which was to look at eye-movement patterns in the cue-outcome ordering for both experiments.

Many studies have demonstrated prediction making for language processing in native speakers (e.g. Delong et al., 2005; Trueswell et al., 1993), and to some extent in L2 speakers (e.g. Grüter et al., 2014), although prediction performance is generally weaker than that observed in the studies of L1 language processing (Grüter et al., 2021). Consider that the current study implemented artificial language stimuli which could be similar to an unfamiliar L2 for the adult participants, the predictive behaviors presented by the two experiments could potentially shed light on the findings of less prediction making in L2 speakers (e.g. Martin et al., 2013). Additionally, the study is among the first studies on language acquisition that carried out eye tracking with a web-based algorithm, Webgazer.js (Papoutsaki et al., 2016). The successful operation and meaningful results would serve to validate such a novel approach for its implementation in future studies.

To set a clearer stage for the current study, chapter 2 will review relevant concepts and literature in more details, and specify the research hypotheses. The two experiments with their methods, results and brief discussions will be covered in chapter 3 and 4. Chapter 5 will provide a general discussion on the interpretation of unexpected results, implications, and limitations. The study will be concluded in chapter 6.

2. Literature Review

This chapter starts with describing the historical background of different theoretical approaches of language acquisition, highlighting their alignment and disparity. Existing findings regarding human statistical learning ability and connectionist models will then be reviewed, followed by the re-emergence of DL theory from its original formulation under behaviorism. Finally, the chapter will discuss the empirical debate on prediction making in L2 acquisition contexts. The aim of this study and key hypotheses will also be elaborated at the end of the chapter.

2.1 Different theoretical perspectives on language acquisition

At the start of the 20th century, behaviorism was a dominant approach to understanding human learning. Assuming the cognitive processing mechanism – what is happening inside the "black box" of learning – was impossible to decode, researchers treated learning as *behaviors* that can be observed with an association to the environment. They believe the associative process of human learning is shaped by positive and negative reinforcements (*the stimulus-response theory* or *operant conditioning*; Skinner, 1938; Thorndike, 1898) and applied this approach to language development. Particularly, Skinner (1957) explained speech as the product of reinforcing the accurate output with language models in the environment, and discouraging erroneous utterances using corrections made by others. However, this sole dependence on interaction with input to achieve a fully developed language system was strongly refuted by Chomsky (1959). He argued learners' environmental input is not sufficient for them to acquire all possible grammar structures (Chomsky, 1980), referring to this as the "poverty of the stimulus" (POS) and proposed instead an innate system of structural rules – later termed as Universal Grammar (UG) to drive language acquisition. As such, Chomsky's review became hugely influential and led to behaviorism being abandoned by mainstream linguistics.

The POS argument highlighted several aspects of language development to support

the existence of UG. Firstly, language development seems to be generative. Under the behaviorist perspective, learning is an imitating behavior to the available linguistic forms in the environment. Thus, learners should not be able to generate forms they have not encountered. However, extensive examples of overgeneralization errors produced by children show that this is not the case. Since mistakes such as *write-ed* and *mouse-s* are rarely presented in the adults' output (MacWhinney & Snow, 1985; Ramscar et al., 2013), children's ability to creatively produce unencountered forms cannot be explained by the simplistic input-driven account of behaviorism. Secondly, the nativists noted the environmental input lacks *negative evidence* to inform learners what they should or should not produce (Brown & Hanlon, 1970; Gold, 1967). Taking the word *mouse* again as an example, the behaviorists explain learning of *mice* as the correct plural form of *mouse* by attributing to the positive reinforcement with adults' production of *mice*. However, they would not be able to clarify how children unlearn their overgeneralization errors of *mouse-s* because no negative reinforcements could be carried out if the form does not exist in the input. Also, children cannot rely on corrective feedback from their parents because researchers have found that direct corrections to child speech were not usually provided (Brown & Hanlon, 1970). Moreover, in experimental cases where parents do make an effort to correct children's grammar, the corrections were either ignored (1a) or wrongly interpreted (1b).

1a. Child: Want other one spoon, Daddy".
Father: You mean, you want THE OTHER SPOON".
Child: Yes, I want other one spoon, please, Daddy.
Father: Can you say "the other spoon"?
Child: Other… one…spoon.
Father: Say… "other".
Child: Other.
Father: Spoon.
Child: Spoon.
Father: Other…spoon.
Child: Other…spoon. Now give me other one spoon?
(Adapted from Brain, 1971)

1b. Child: Nobody don't like me.
Mother: No, say "nobody likes me".
Child: Nobody don't like me.
[Eight repetitions of this dialogue follow.]
Mother: No, now listen carefully, say "NOBODY LIKES ME".
Child: Oh! Nobody don't likes me.
(McNeill, 1969)

Consider the poor availability of sufficient and interpretable negative evidence described above, the input should not be able to constrain children's productions by itself. Therefore, nativists proposed that humans must be biologically endowed with knowledge about the principles of human languages, UG. In this way, children can map the specific input in the environment on to UG and set parameters correspondingly to develop their own language systems (Chomsky, 1995; Hyams, 1987).

A third point the nativists made was that language input is idiosyncratic, which specifies that language developments of learners with different input environment converge to a common grammatical system (Crain et al., 2019). This point is a controversial support for UG because learners do not necessarily develop the same grammar. In Street and Dąbrowska (2010), native English speakers were tested for their comprehension of simple grammar structures with sentences such as "Every mug has a toothbrush in it" and "The girls was hugged by the boy". The participants were recruited from different backgrounds of academic attainment (AA) and were found to perform differently in the task of sentence-picture mapping. Specifically, the high AA group performed better than the low AA group (separated by formal education more or less than 11 years) for all three structures tested, actives, passives and simple locatives with quantifiers. Under UG, the comprehension differences would not be expected because learners should have gained comparable amounts of linguistic knowledge with the universal guide of what to acquire. In the meantime, studies also demonstrated the variation of language development in other syntax structures (e.g. Chipere, 2001; Dąbrowska, 2008; Indefrey & Goebel, 1993), learning

styles (i.e. developmental pathways; Peters, 1977), and learning rates (Fenson et al, 2007). With this evidence, it is clear the nativist approach is not fully capable of accounting for the individuality and divergence of language acquisition, suggesting the innate account is not as compelling as it claims.

Other than the controversy over convergence of grammar development, the most problematic side of UG is that it cannot fulfill the basic criteria of scientific theories – falsifiability. As formulated by Popper (1959), a theory can only be scientific if it can be challenged and even proven false. However, there has been no clear agreement for UG on what the principles and parameters of grammar structures are despite decades of research (Dąbrowska, 2015; Smith, 1999). Thus, it is difficult to falsify if researchers can come up with alternative theories in a post-hoc manner (e.g. Cipriani, 2015; Newmeyer, 2008). Additionally, there is little empirical evidence directly supporting the biological existence of such an *innate* language system (Sampson, 2005). In fact, nativists themselves assumed the parameter setting to be invisible (Newmeyer, 2008). It is also vague how parameters could be part of the innate universal structure if they were proposed to vary across languages (c.f. Newmeyer, 2008). Consequently, it is hard to find counterevidence for UG as the theory does not have clear limitations of its scale and content. Considering the problems of the nativist account, an alternative perspective would be the cognitivists who suggest that language learning relies on our general cognitive abilities, such as information processing, attention allocating, and memory encoding (Bates, 1979; Piaget, 1999). With domain-general constraints on language development, the account can not only go beyond the behaviorist approach, providing a more detailed depiction of the "black box" between input and output, but also exceed nativism in the way that it can be falsified with the concrete functions of the cognitive abilities.

The constructivist approach, though similar to behaviorism in assuming that language learning relies on domain-general learning mechanisms, extended upon cognitivism and re-envisioned the role of input. Proponents of this approach believe that

knowledge of languages is actively synthesized by learners from their individual experiences with linguistic input from the environment. For example, Ellis (2003) proposed a constructivist account of learning as extracting "chunks" (collocation of linguistics units) through the sequential frequency information in their input. Tomasello (2003) explained children's categorical acquisition of nouns as the abstraction of common properties shared by all kinds of nouns they have encountered. Under the constructivist approach, individual variation in acquisition of grammar, or any other linguistic structure, could be attributed to the different experiences of learners with their input and environment. Furthermore, unlike the assumption of UG, it can be verified by examining a central ability of humans claimed by constructivists like Ellis (2003) and Tomasello (2003) – the ability to learn linguistic structures by abstracting patterns from the input. This ability to construct linguistic knowledge from the input is reviewed in the following section using the literature on statistical learning.

2.2 The statistics in the language environment

After a few months from birth, humans can already apply inductive reference to infer the world around them for object information such as pattern of shapes, location of appearance, property of populations and categorization (e.g. Canfield & Haith, 1991; Duffy et al., 2006; Johnson et al., 2003; Xu & Garcia, 2008). In fact, the first study that clearly established this ability to learn from input information was conducted in the language acquisition context. Saffran, Aslin and Newport (1996) asked if infants at 8 months could segment words from artificial speech samples based on only statistical sequencing of sounds. They generated a continuous speech stream with some three-syllable nonsense words repeated in random order (i.e. bidaku/padoti/golabu/bidaku) and ensured the speech stream contained no prosodic cues – the only way to segment the words was through the transitional probabilities that were higher within words (e.g.daku or pado) and lower between words (e.g. ku/pa). After exposing the 24 infants (from American-English speaking background) to the speech stream in monotone for 2 minutes, researchers found that

after training, infants could successfully discriminate between “word” stimuli (e.g. bidaku) and “nonword” stimuli (e.g. dadola; syllables strings that has not occurred in the input). This discrimination was shown by infants listening significantly longer to the nonwords which was taken as demonstrating a novelty preference. Moreover, they repeated the experiment with another 24 infants from the same background, but with “part-word” (e.g. bidaku) instead of the “nonword”. The “part-word” stimuli were created by taking the first two syllables from a “word” and the final syllable from another “word” so that it would be more difficult to discriminate than “nonword” in terms of the transitional probabilities between the syllables. However, the infants again showed significantly longer listening times for “part-words” than “words”, confirming the infants’ ability to track transitional probabilities in their input. The results of the two experiments indicated that human learners are sensitive to the statistics in languages. This seminal work re-vitalized the exploration of input driven approaches to language learning and established the study of *statistical learning* among researchers in psycholinguistics and developmental science.

Following the work of Saffran et al. (1996), studies have tried to reveal the wealth of cues in the linguistics environment for different levels of language learning. These include the distributional cues in syllables (e.g. Werker et al., 2007), morphology (e.g. Rumelhart & McClelland, 1986), word categories (e.g. Swingley, 2005), and syntactic structures (e.g. Martí et al., 2019). For example, in Mintz et al. (2002), researchers analyzed the distributional information of word grammatical categories with four corpora of child-directed speech extracted from the CHILDES database. For the distributional information of each word in the data, they recorded the frequency of the most frequent 200 words that come before and after each words, and created a string of frequency numerals as the vector reference for the words’ immediate contexts – their co-occurrence patterns with the other words. By comparing the similarity of these context references, Mintz et al. (2002) showed that words could be successfully categorized into nouns and verbs. This result is in line with findings from the corpora analyses which suggested that words with the same

syntactic categories tend to occur in similar linguistic contexts (e.g. Brill et al., 1990).

On the other hand, studies of statistical learning have also tried to demonstrate that learners can make use of the cues uncovered in the input. At the word level, infants were found to have tracked segmentation cues including transitional probability of phonemes (Friederici & Wesels, 1993), allophones (Jusczyk et al., 1999), and stress patterns (Jusczyk et al., 1993; 1994). Beyond word segmentation, Maye, Werker and Gerken (2002) demonstrated that statistical distribution of phonetic variation in speech signal can affect how infants at 6-8 months discriminate speech sounds. For sentence structures, Thompson and Newport (2007) also found that adults can segment phrases in sentence stimuli based on transitional probabilities between words or word classes. In addition, consistent with the ample evidence identifying the patterns of syntactic information in the input, many studies explored the use of transitional probabilities in segmenting of linguistic structures to learn underlying grammar (e.g. Tamminen et al., 2015; Wonnacott, 2011; Wonnacott et al., 2008; 2012). For example, Gomez (2002) conducted a study with adults and 18 months old infants and investigated their learning of nonadjacent transitional probabilities (i.e. the dependency between two nonadjacent elements in a linguistic string) when regular decencies of adjacent elements fluctuate with set-size differences. Consider examples like change of auxiliaries and inflectional morphemes to modify tense (*have* writt*en*) and subject verb agreement in longer sentences (The *ice-cream* in the fridge *is* mine), transitional probabilities of nonadjacent elements could be important in learning certain types of syntactic structures. Gomez (2002) manipulated the two types of artificial stimuli to only differ in non-adjacent dependencies, and the set-size differences by adjusting the size of pool from which they drew elements to form the stimuli. The results showed that both adults and infants increased discrimination of non-adjacent transitional probabilities in novel strings after brief exposures (18 minutes for adults, 3 minutes for infants) to the type of acoustic stimuli with higher variability (i.e. drew from larger set), indicating that humans can process other source

of statistical patterns even when adjacent transitional probabilities are not stable in a highly variable input.

Note that most of the statistical learning studies mentioned have applied artificial languages to focus on particular distributional properties while controlling for the possible confounding factors of a natural language setting. Thus, there have been concerns with the relevance of laboratory results to real life scenarios (Saffran, 2009; Johnson & Seidl, 2009), and further literature has aimed to demonstrate learners' sensitivity to probabilistic information in natural languages as well (e.g. Jusczyk et al., 1993; Jusczyk, Luce & Charles-Luce, 1994). Later studies such as Pelucchi et al. (2009) worked with infants from monolingual English backgrounds and obtained positive results for their investigations on word segmentation for an unfamiliar language, Italian. Another example study in a L2 context was conducted by Treffers-Daller and colleagues (2015; 2016) who investigated the influence of statistical learning on L2 use of motion verbs. They found that after getting more exposure in the target language via study abroad or language classes, learners of either a target language related to their first language (English-French) or a totally different L2 (Mandarin-English) had all adjusted their frequency of use for manner and path verbs in L2 towards the frequency in the input. These results align with the other findings that learning of syntactic structures is influenced by the distributions of input (e.g. Casenhiser & Goldberg, 2005; Wonnacott et al., 2012; 2017). Some other studies continued to use manipulated artificial stimuli but showed that they were treated like natural language. This is the case presented in both Saffran (2001) and Graf Estes et al. (2007) which demonstrated that once infants at 8 or 18 months had learned nonsense words via statistical learning from an artificial language, they treated them as novel words and integrated them appropriately into their natural language use (i.e. into English) in the tests.

Taken together, this body of research for statistical language learning illustrates that humans can indeed learn regularities at different levels of the language system, which

supports the account of constructivists. However, these studies do not aim to investigate the exact mechanisms which underpin the statistical learning process. That is, the empirical evidence discussed so far tells us what information is in the input, and whether learners are sensitive to this information, but we also need to know "how" learners make use of the input for language learning without innate constraints. The review continues in the following section by discussing the *DL theory* which has been proposed as a candidate mechanism.

2.3 The discriminative learning mechanism

One methodology to explore the learning mechanisms behind statistical learning is computational modelling. A classic study by Rumelhart and McClelland (1986) applied its neurocomputational theory of parallel distributed processing (PDP) to simulate children's past tense acquisition. Inspired by findings in neuroscience that human brain activity involves connections among a network of neurons, the connectionists modelled learning as incremental changes in the connection strength between processing units (equivalent to neurons in the neuro-networks) based on learners' interaction with the input regularities (Joanisse & McClelland, 2015). In this way, the cognitive abilities including language processing can be accounted by the constant adjustment of the "neuro-networks". For example, learning of past-tense morphology can thus be interpreted as a process of deriving associative strength for formation of "rules" from correct inflections in the input (Rumelhard & McClelland, 1986). Elman (1990, 1993) also provided an account of the simple recurrent networks (SRNs) to learn lexical categories inductively by grouping words that usually occur at similar locations in the input together. These models provided psychological plausibility to the constructivist account as they implemented learning mechanisms mimicking the brain activities. They also demonstrated possible ways that humans can learn some of the linguistic rules (e.g. past tense) without innate knowledge.

However, when it comes to the details of each model, there are still some limitations. Firstly, many of these models are set up to rely on the feedback from an external

teacher. For example, in Rumelhard and McClelland (1986), when the PDP model tried to create the past-tense form of a stem and make errors, it would adjust its weights on the basis of seeing the "correct" forms. It therefore assumed a situation in which learners have access to a teacher who corrects them when they make errors, but at least in first language (L1) acquisition this is generally not the case as discussed above (Brown & Hanlon, 1970; Marcus, 1993). Secondly, most connectionist models are set up to map between different linguistic forms such as turning a stem into an inflected form, but in natural language learning, linking forms to *meanings* is a more primary task (Christiansen & Chater, 1999; Quine, 1960). In contrast, *DL* (Ramscar & Yarlett, 2007) is a theoretical framework that addresses these problems. As described in chapter 1, this mechanism works by adjusting predictive values of informative cues to disambiguate them for a certain linguistic form. Under the DL framework, language learning without UG can still be constrained by obtaining negative evidence from prediction errors. As learners make predictions with possible cues for a certain outcome, they also encounter errors from uninformative cues. The predictive value of these cues would decrease when errors are made and in turn allow the competing cues to increase their informative strengths. Alternatively, correct predictions of the cue-outcome pairs would increase the value of that cue and make its competing cues lose weights on the balance. For example, when learners try to name a picture of multiple mice, they could make the prediction "mouses" according to the plurality cue of multiple "mouse" presented. However, after they hear "mice" from adults' input, which does not match their prediction, "mouse" with plurality cues would be downweighed. In this way, prediction errors would work as negative evidence to help the learners reduce the strength of uninformative cues and ultimately get them disassociated. The informative cues can also be gradually distinguished after they have accumulated way more predictive values comparing to the uninformative counterparts.

Note that this procedure of cue competition via value adjustment resembles the positive and negative reinforcement of learning behaviors under behaviorism. In fact,

DL is grounded in the principles of learning theory, which was developed to describe animal learning under the behaviorist paradigm (Reynolds, 1961). Although the behaviorist approach was originally rejected by mainstream linguistics (i.e. Chomsky, 1959), there is now a revival of the idea that some of the principles might be able to explain learning of complex systems, including language, when they are embedded in computational systems (Gallistel, 2002; Ramscar & Yarlett, 2007; Rescorla & Wagner, 1972). Particularly, with the error-driven procedure, DL of meanings (Ramscar et al., 2010) or linguistic structures (Ramscar & Yarlett, 2007; Ramscar et al., 2013) can be both accounted without external corrections or inbuilt linguistic rules. Moreover, the predictive value of cues constantly changing in these models has analogues in brain activity: the firing patterns of dopamine neurons in the monkey brains for under- and over-predicted learning trials (Waelti et al., 2001) are similar to the process of cue competition, suggesting a strong biological base for the DL framework.

Another major advantage of the DL theory is that it can make specific and testable predictions about when DL should/should not occur, such as the effect of input order on the learning process and structure generalization. Specifically, Ramscar et al., (2010) simulated word-object mapping with the discriminative model and found that learning was best when objects with complex features were presented to learners *before* the word label (object-label), rather than vice versa (label-object). This is because learners could only get access to enough candidate cues for predictions making and cue competition under the input order of object-label. Taking the word "fox" in English for illustration, when learners see the animal fox visually, they would notice a set of cues regarding the color, the size or other features that could compete to be informative for the label "fox". During their multiple encounters of these features and occasions where the word "fox" was heard, they would downweigh the cues such as *long tails* or *omnivorous* if they predicted to hear the word "fox" and it did not occur. Gradually, the most informative cues (e.g. *triangular face* or *red fur*) would be discriminated by accumulating strength from their frequent correct

predictions towards “fox”. In contrast, if learners instead generally heard the word “fox” before seeing the animal, they would not make predictions for the word on the basis of its cues. Thus, they would not be able to discriminate the uninformative cues because no cue competition occurs and no negative evidence can be obtained.

To test this hypothesis about the effect of input orders, Ramscar et al. (2010) conducted an experiment with alien objects and included three label categories, dep, wug, and tob, which are all English nonwords. Each label was paired with two alien objects that differ in their features (figure 1). Critically, body type (i.e. shape and color) was a highly salient but uninformative cue and the categories were designed such that for low frequency objects, they shared body type with a high frequency object of a different label. Under this design, it was necessary to dissociate the body type cue in favor of the informative discriminative cues to correctly learn the label for low frequency objects. Similar to the “fox” example above, the only way to dissociate an uninformative cue in DL is by reducing its predictive values using prediction errors. However, this could not be achieved in a context where participants encountered labels before objects. When objects appear first, participants would make predictions based on the available object features to allow cue competition for their predictive values towards the upcoming label. Conversely, when labels appear first, no prediction errors would be made on the basis of the object cues because the set of features associated with the label are directly provided following its appearance. Without prediction errors, the uninformative cues could not be dissociated under cue competition. Therefore, participants should only be able to show learning of the low frequency cues under the object-label input order.

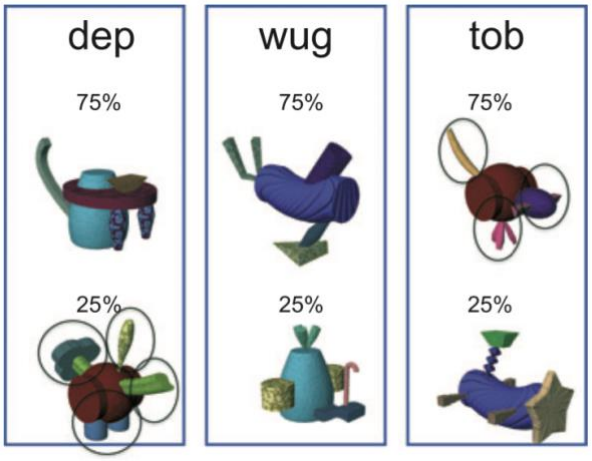


Figure 1. The category structures employed in experiment 1 of Ramscar et al. (2010). The alien objects that served as stimuli were the *fribbles* retrieved from the website of Michael Tarr's lab: http://wiki.cnbc.cmu.edu/Novel_Objects. Note that body type does not discriminate between categories. Thus, using this cue of a low frequency "dep" might lead to an erroneous identification of "tob", for which the body type has occurred more frequently with. The sets of discriminating features that need to be learned in order to successfully distinguish the label categories are also circled on the low-frequency object of ''dep'' and the high-frequency object of ''tob''. The image is adopted from figure 7 in Ramscar et al. (2010).

Ramscar and colleagues recruited 32 undergraduate students from a US university and asked them to learn "the names of alien species". An additional control category with binary pairing of a label "bim" and the color blue was added to ensure a high baseline learning. Learning of label and cues was indicated with a post-training generalization test (i.e. 4 alternative forced-choice tests). It used novel fribbles with the same features as those associated with the different categories of features in training, which required the participants to learn word-referent (cue) pairing to the extent that they can generalize with it in new circumstances. Results showed that only participants who were exposed to object-label training achieved an above-chance accuracy rate in the post-test and supported the hypothesis of different learning outcomes resulting from the order of input.

Similar ordering effects in learning have also been found in experiments looking at learning of non-native acoustic cues in Nixon (2020) and affixes in Vujović et al. (2021). Consistent with the design of Ramscar et al. (2010), the two studies compared

between two ordering conditions to demonstrate effect of DL. In Nixon (2020), each acoustic stimulus (word) was formed by a combination of base-syllables and tones (i.e. ‘tsche’ rise or ‘phe’ fall). Critically, the words were assigned to semantic categories such that the same base-syllable was associated with multiple categories, and thus the base-syllables could be uninformative cues equivalent to the body type of fribbles in Ramscar et al (2010). Participants would need to dissociate these to find the true discriminating cues, tones, for the semantic outcomes. In this experiment, the discriminative order presented the speech cues (words) before the semantic outcomes (pictures), which expected a learning advantage in “word-picture” rather than “picture-word” order. Results did indicate a better learning in this discriminative order as well and demonstrated that the effects shown in these experiments for word learning are not focusing on the ordering of “words” and “pictures” per se, but more so on using an order that allows the appropriate cue competition over the available features.

In Vujović et al. (2021), the two input orders came from contrasting prefixing and suffixing – where the affix comes before or after the word stem. Participants were then expected to learn the discriminating cues for an affix better from the stem or visual object under the suffix condition, when they hear affix after the audio of stem label with its visual representation of fribbles (the semantics). In this experiment, affixes were the outcome. The consistent vowels categorizing each stem (e.g. “/mi:b/” or “ti:p/”) and the discriminating semantic features of fribbles were both informative cues for the affixes, which characterized the co-occurrence of phonological and semantic cues for grammar generalization in natural language settings (MacWhinney et al., 1989). They were both competing with the overlapping body type of fribbles, which again worked as the uninformative cues. As expected, a benefit for the discriminative suffixing ordering in learning low frequency items was found in the post-training generalization tests.

These studies further confirm that DL order leads to stronger performance, particular for low frequency cues when there is a salient, high frequency competitor. However, while these results are in line with Ramscar et al. (2010), and the general hypothesis of DL theory, they did not provide *direct* evidence for the mechanism of learning in that theory, i.e. that learners make prediction when they see a cue and show expectation of seeing the corresponding linguistic outcome. Therefore, the current thesis aimed to find concrete evidence that learners do indeed make prediction errors during learning. This finding would be particularly important since participants of the two current experiments were all adults with little exposure to the language stimuli, and there is evidence that while we make abundant predictions when processing in our native language, prediction making is less clear for adult L2 learners (e.g. Grüter et al., 2014). Before the proceeding two experiments of the thesis, the subsequent section reviews this literature on human predictive processing during in natural languages.

2.4 Learning through predictions

Previous literature has provided abundant evidence indicating that native speakers can make predictions based on their experience of patterns in their language (e.g. Dahan et al., 2000; DeLong et al., 2005; Hare et al., 2003; Juliano & Tanenhaus, 1993; Lew-Williams and Fernald, 2007). Many studies implemented eye tracking to gain on-line behavioral evidence for prediction making. For example, Trueswell et al. (1993) tested participants in a reading task with sentence stimuli featured in two types of verb-biased structures. Direct-object biased verbs like *observed* tend to take noun phrases as complements (e.g. the friend *observed* our house), whereas sentential biased verbs like *insist* are usually followed by sentence complements (e.g. the friend *insisted* (that) our house was unusual). By monitoring fixation durations during the reading of disambiguating word (was) in 2a and 2b, the participants were found to have increased their reading time for 2a when no complementizer (i.e. that) was following the direct-object biased verb, but not for the sentences with sentential biased verb in 2b. This finding of early sentence processing being influenced by verb

information indicates that participants generated presumptions based on the verb bias – "the solution" should be a direct object that ends the sentence when no complementizer is presented to suggest a following sentence complement. Thus, when participants continued to see "was" which did not fit their prediction, they needed more time to process. Meanwhile, Snedeker and Trueswell (2004) tracked looking time of 5-year-old children while giving them instructions to interact with objects on the table using differently biased verbs. They found similar sensitivity to verb bias information in guiding children's processing of the instruction, showing that predictive processing is common to language speakers of all age.

2a. The student *forgot* (that) the solution <u>*was*</u> in the back of the book.
(direct-object bias)

2b. The student *hoped* (that) the solution <u>*was*</u> in the back of the book.
(sentential bias)
(Both 2a and 2b are adapted from Trueswell et al., 1993)

While evidence in L1 is well established, research extending the exploration to L2 learning has obtained mixed results. In comparison to native speakers, L2 learners were found to have reduced ability to generate expectations during sentence and discourse processing (Grüter et al., 2014; Martin et al., 2013). For example, Mitsugi and MacWhinney (2016) reported a lack of prediction in Japanese L2 learners in processing the morphosyntactic structures such as case markers. They tracked eye-movements of both native and L2 learners of Japanese while asking them to explore whether the images displayed on the computer screen could apply to the Japanese sentence stimuli they heard. The predictive behavior was indicated by fixations to the theme object in response to the stimuli. Results showed that when a sentence was played with case markers that suggest a particular theme object will follow, native speakers looked more at the theme object during the prediction period. However, no significant difference in the fixations was observed for L2 learners. Dijkgraaf et al. (2017) adopted a similar paradigm to test unbalanced Dutch-English bilinguals and

English monolinguals. They heard simple sentences and viewed one of two versions of a visual scene: the constraining condition had only one object that would be appropriate following the verb, whereas all four objects in the neutral condition would be appropriate. They showed that bilinguals (L2 speaker for English) made a higher proportion of fixations to the target object in the constraining rather than neutral scene before hearing the theme object in the sentences, thus indicating that predictive processing was conducted by the L2 speakers. Additionally, this fixation behavior was also found to be comparable to the predictive processing performed by monolinguals in the same test. However, since this study was conducted with relatively proficient bilinguals compared to the other studies (Dijkgraaf et al., 2017: around 10 years of English learning in school; Mitsugi & MacWhinney, 2016: average 4.25 years of Japanese courses), the inconsistent findings in L2 prediction processing could thus be caused by the proficiency difference in participants.

It seems that literature presents a mixed picture as to whether L2 learners can make predictions in language processing as readily as the native speakers. Considering the artificial language studies which have claimed to demonstrate DL (Nixon, 2020; Ramscar et al., 2010; Vujović et al., 2021) worked with adult participants who were new to the artificial languages in the experiments, and who received a relatively small amount of exposure, it is questionable whether these participants *were* making predictions about the upcoming outcomes on the basis of cues when processing this unfamiliar language within a short time. This is tested in the current study.

2.5 The current study

This review has discussed how constructivists have provided an alternative to the nativist account in which language development is constrained by extracting distributional information from the environmental input. The account receives support from findings of statistical learning research and shows that not only is useful probabilistic information provided by the input, but also human learners can pick up on it. DL theory provides a potential mechanism for this input-driven learning process

which relies on prediction errors to prompt the constant adjustment of associative strength between cues and outcomes in the linguistic environment. So far, past experimental research has mainly focused on proving that learning is better in conditions where stimuli are presented in an order that theoretically allows learners to make predictions with cues (i.e. Nixon, 2020; Ramscar et al., 2010; Vujović et al., 2021). However, learning performances were compared in a post-training "offline" test so that the better learning in discriminative condition groups explained in terms of prediction making was only an assumption. Moreover, a review of the literature on native and L2 predictive processing suggested though adult L2 learners do engage in predictive processing, they may not do so as much as native speakers. This further motivated the need to establish direct *on-line* evidence for predictive processing in the experiments which have claimed to show DL. Therefore, the current study conducted two experiments with adults participants to investigate their use of prediction errors in learning artificial language stimuli.

Experiment 1 repeated the condition with the discriminative ordering in Ramscar et al. (2010), where participants see object before the label. The experiment applied eye tracking during the learning session to monitor on-line predictive behaviors of DL by tracking participants' fixations on the screen. Critically, both objects and labels were presented visually as in Ramscar et al. (2010), but each of the three category labels occurred at different fixed position on the screen. We want to see prediction making indicated by the difference in proportion of fixations when participants look at the designated position of a particular label before the label appear. Given that the design is relatively exploratory and incorporates a novel web-based eye-tracker (due to the restrictions of COVID-19), a pipeline for data analysis was used to avoid overinterpretation of spurious results: firstly, we sought to observe a baseline looking pattern in a control condition where the label position is unpredictable and participants were expected to only look at the label *after* it appears on the screen. Following the design of Ramscar et al. (2010), there were high and low frequency cues to incorporate cue competition. Once the baseline looking pattern was

established, we would check if participants look at the predicted label position on the basis of high frequency cues *before* the target label appears. Finally, if the desired saccades were found in both control and high frequency conditions, we would expect participants to make the same predictions with the low frequency cues. An offline test resembling the post-training generalization test in Ramscar et al. (2010) was also included and participants were expected to obtain above-chance accuracy for both the high and low frequency trials.

Experiment 2 repeated Experiment 1 in a different linguistic paradigm and the details are explained in chapter 4.

3. Experiment 1 - fribbles

3.1 Methods

3.1.1 Participants

33 proficient English speakers with variant language backgrounds were recruited through Prolific (www.Prolific.co), and among colleagues and students in University of Oxford. All participants were over 18 years old with normal or corrected to normal vision and reported no history of neurological or learning impairments. They gave consent to participation and were reimbursed for their time.

3.1.2 Materials and design

The current study replicated the object-label condition of Experiment 1 in Ramscar et al. (2010) and was conducted online through the Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine et al., 2018). Participants were taught 4 pseudo word labels ("wug", "tob", "dep" and "bim") for some novel pictures of the alien objects – *fribbles* (see figure 2). Images of *fribbles* were the same used in Ramscar et al. (2010) and were retrieved from Michael Tarr's lab (http://wiki.cnbc.cmu.edu/Novel_Objects). The label "bim" was used as a control condition that paired with only blue objects to indicate the baseline learning performance. The participants were then tested to see if they could generalize cues to the correct labels.[1]

[1] In both Experiment 1 and Experiment 2, an additional contingency judgement test was conducted after the 4AFC generalization test to examine participants' ratings of word labels paired with novel fribble objects. Participants rated the pairing on a scale from -100 (perfectly disagree) to 100 (perfectly agree). This tested the same information as the 4AFC test using a different method and the results showed qualitatively the same pattern of results as the 4AFC test. Considering the amount of data in the current study and the length of this dissertation, this test is left out of this report following supervisor's advice.

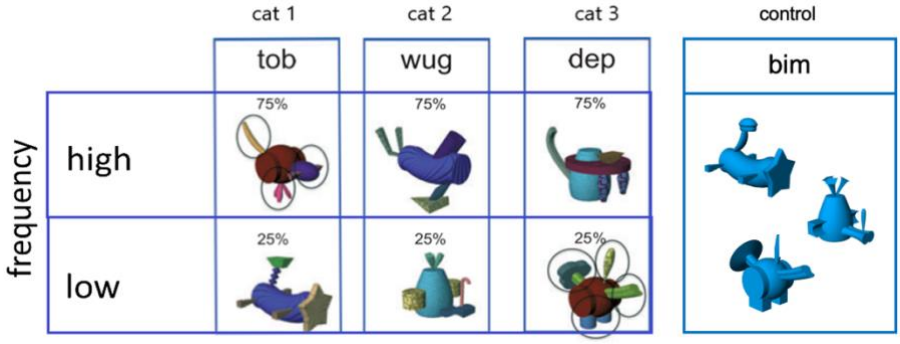


Figure 2. *Experiment 1:* Demonstration of the stimuli set. Each label category consists two sub-categories of fribbles which are associated to their respective labels at different frequencies (high or low). Informative features, predictive of the label category, are circled in green. The control fribbles that are associated to label "bim" all shared the blue color feature. The image is adapted from figure 7 in Ramscar et al. (2010).

In each trial of the training phase, participants were presented with fribbles followed by word labels that appear consistently in unique positions on the screen (left, right or top-center). The timings are shown in figure 3 and are based on Ramscar et al. (2010). If participants make predictions on the basis of fribbles' visual features, such as body types (attached structures and colors; figure 2), they might predict particular labels, and this predictive behavior can be observed by tracking the fixation spots of participants before the labels appear. The core design was identical to Ramscar et al. (2010): each word label was paired with two fribbles, one occurring at a higher frequency of 75% and the other one at 25%. The low frequency fribble shared a high-saliency but uninformative feature (i.e. body type) with the high frequency fribble of another label category. For instance, in figure 2, both the low frequency fribble of "dep" and the high frequency fribble of "tob" possessed a red rounded central body structure. As discussed in section 2.3, the frequency manipulation stimulated the cue competition and required the participants to dissociate uninformative cues to get the correct mapping of word label and object features.

In addition, the control fribbles associated with the label "bim" were expected to be well learned because they were all in a uniform blue color. Since the experiments

were conducted online where participants were not observed in the same way they would be in lab as in the original experiment, the control category could serve to filter out any participants who were not paying attention as they would do badly on control trials in the offline tests. Other than this purpose, the control condition was used as a way of testing the eye-tracking paradigm. In contrast to the experimental stimuli, the control label "bim" did not occur in a fixed position as the other test labels and thus could not be predicted. This design allowed us to see the default pattern of looking when participants could not predict the upcoming label.

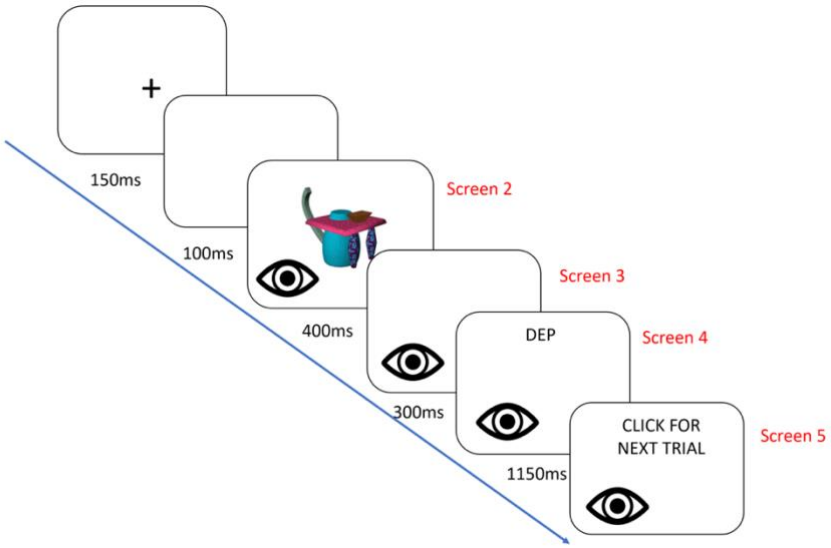


Figure 3. *Experiment 1:* Demonstration of set up and timelines of the learning phase. The time frames where eye-movements were tracked and marked with the eye symbol. The duration of each screen is marked below.

3.1.3 Tracking eye-movements online

To provide a direct measure of the discriminative learning process, eye movements were tracked from the fribble onset (Figure 3: screen 2) to 150ms after the label offset (Figure 3: screen 5), for a total duration of 2 seconds per trial. The embedded eye-tracker function of Gorilla allows eye tracking via participants' web cameras by utilizing the library Webgazer.js (Papoutsaki et al., 2016). The model package can be executed on participants' browsers to predict their real-time eye-gaze positions

through the facial information detected by their web cameras. Before applying to the learning phase of the current experiment, the function was piloted for its recording precision and data validity. Raw data of fixations by time with a resolution of 60Hz were sorted using R (R Core Team, 2020) to compute the proportion of fixations by time in three regions of interests, corresponding to the three word label positions where learners may conduct predictive saccades. To ensure the quality of the eye-tracking data, participants were required to calibrate with the standard 9-point display before and throughout the learning phase of the experiment. They were also asked to keep their head still in front of the webcam, to not blink more than necessary, and to try staying in a bright environment. Since all the eye-movements information was estimated from the X and Y coordinates that are translations of the webcam images, no sensitive information about the participants was accessed or saved in the process of eye tracking.

3.1.4 Procedure

Data collection was ethically approved by the Social Sciences and Humanities Interdivisional Research Ethics Committee of University of Oxford with the reference number R73484/RE001. A demonstration of the general experiment procedure is provided in figure 4.

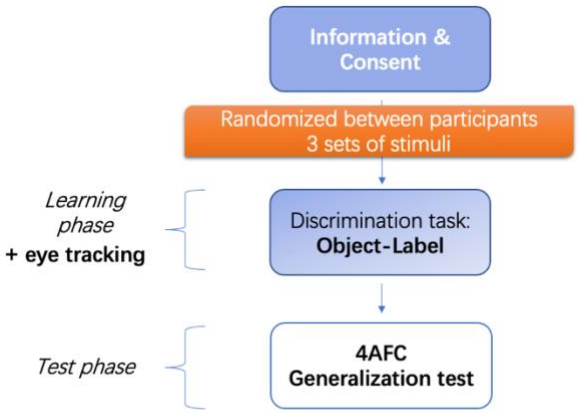


Figure 4. *Experiment 1:* Demonstration of the procedure.

Participants were instructed to remember the names of the novel objects presented to them in the "alien language learning" game. They were informed that their eye-movements would be tracked through their webcam and that they would be asked to recall the objects and names they learned during the 30-minutes experiment.

Learning. The learning phase started with 6 practice trials to familiarize them with the procedure of seeing a fribble image (different from the test trials) followed by each of the three test labels. After participants passed through the first calibration of the eye-tracker, they proceeded to go through 160 training trials in total. The trials consisted of 40 control trials, 90 high frequency trials and 30 low frequency trials, corresponding to the 75% and 25% design for the frequency conditions (table 1), and were randomized into 6 blocks with breaks in between. Demonstrated in figure 3, the participants were first shown one fribble object at the center of the screen for 400ms. Then, the paired word label would be presented at the designated position for 1150ms following an intermission screen of 300ms. After the word label disappear, the participants were instructed to actively click on the screen to proceed to the next trials. Three different sets of stimuli were randomized between participants to minimize confounding effects caused by the appearance of particular fribble objects or their pairing with the label category.

Task	**Control**	**High frequency**	**Low frequency**	**Total**
Learning	40	90	30	160
4AFC	8	24	24	56

Table 1. *Experiment 1:* Trial arrangement.

4AFC generalization test. To further examine participants' learning in the post-training tests, a 4 AFC test was conducted to see if they could generalize the feature-label pairings they had learned to novel fribbles. The offline test had no practice trials. The total 56 trials consisted of 8 trials for the control condition and 24 for each

frequency condition (table 1). For each trial, the participants were asked to pick the correct object for a word label from four novel fribbles (figure 5). The correct fribble can be identified based on the discriminative features. The other fribbles were all foils: for a low frequency trial, there were a novel control fribble and two other low frequency fribbles that were each associated with one of the other two labels; for a high frequency trial, it was a control fribble and two high frequency fribbles each associated with a different label; the control trial presented three novel fribbles that each associated with a test label other than the novel control fribble. Participants were also informed before the start of the test that they would have 3 seconds to respond for each trial. Each trial screen automatically switched to the next after 3500ms if no response was received to encourage participants' use of implicit knowledge rather than explicit reasoning.



Figure 5. *Experiment 1:* Demonstration of website interfaces during the 4AFC generalization test. The grey bar at the top indicated the remaining time.

3.2 Results

Since the offline performances can give a general indication as to whether participants have learned the object-label pairing or not, the results of 4AFC generalization test were presented first (section 3.2.1). Evidence of anticipatory looks at the upcoming label locations before the labels appear could be explored with the eye-tracking data obtained from the learning phase (section 3.2.2). All data were sorted and analyzed in R (R Core Team, 2020).

3.2.1 Offline AFC test

Data collected from the 33 participants were first filtered according to their performance in the trials where a control fribble ("bim") was the target. 4 participants scored lower than or equal to 80% correct for these control items were excluded. Since another participant had an unknown technical problem and did not complete the 4AFC task, there were 28 participants left for further analysis with an average accuracy of 97% on the control trials. For the remaining participants, analyses were conducted over the trials where one of the low or high frequency fribbles was the target (i.e. excluding trials where the control fribble was the target). Note that each trial consist of four competing object options for the presented label, a target fribble, two other fribbles associated with the other labels, and a control fribble. The subject group were expected to almost never select the control fribble for these trials because participants who did not learn the control fribbles have been filtered out. Therefore, 0.16% of the total trials where they still mistakenly chose the control fribbles were removed so that the chance level for performance was 33%. Another 4% of the remaining trials were excluded because there was no response due to participants having run out of time.

The data are plotted into a violin plot with the mean and a boxplot included for the descriptive statistics (figure 6). It seems while mean participant performance was above chance for both high and low conditions, the performance for high frequency trials was better. Since the dependent variable is binomial, a logistic mixed effect model was used to predict the performance accuracy with frequency as the fixed factor. A random intercept for participant and random by participant slope for frequency (i.e. the within participant effect) were also included. As predicted, results showed significant better performance for the high frequency than low frequency trials (beta = 3.09, SE = 0.48, $p < .001$), meaning that participants can match more novel fribbles belonging to a high frequency category with the correct word label. To investigate whether the performance for low frequency items was above chance, we re-ran the model with just the low frequency items so that the intercept could

correspond to the mean accuracy within only low frequency items. The model was tested to see if it was significantly different from chance (i.e. - 0.69 in log odds which corresponds to probability of 1/3). The result was not significant (beta = 0.13, SE = 0.39, p = .73), indicating that on average participants were picking the correct low frequency fribble more than predicted by chance.

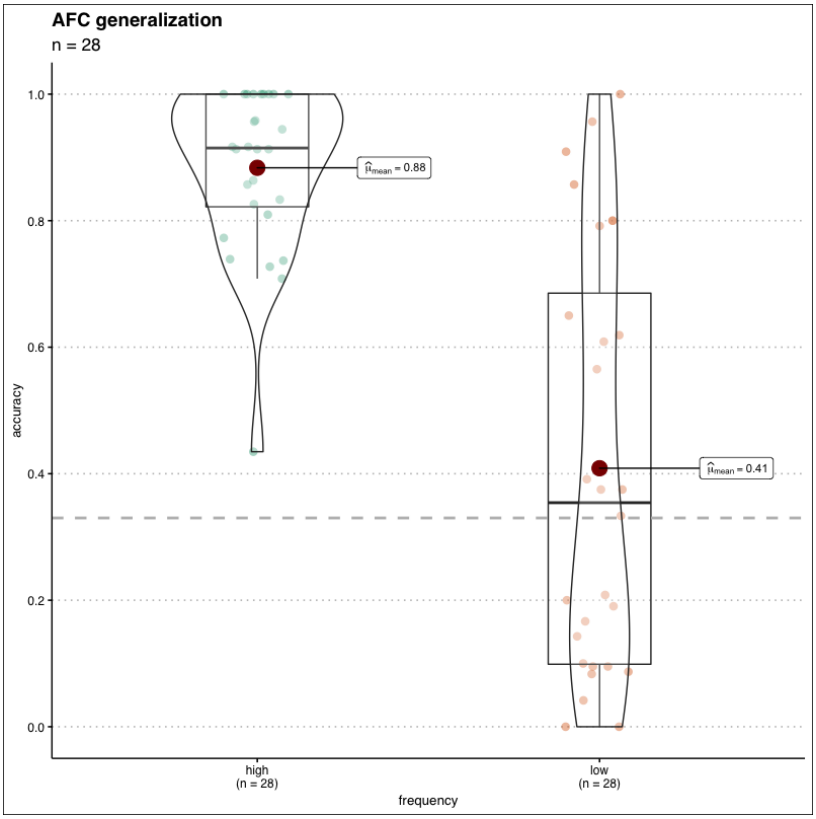


Figure 6. *Experiment 1:* Proportion of correct trials in the AFC generalization test split by frequency conditions. The colored points show the mean accuracy by participant, with the kernel probability density demonstrated by the violins. The boxes mark the interquartile range around the median of accuracy (black lines) for each frequency conditions. The grey dotted lines show the chance level at 33% (control trials are omitted).

3.2.2 Eye-tracking data

The eye-tracking data were extracted between fribble offset (time 0ms) and label

offset (1450ms), with label onset at 300ms[2]. To ensure data validity, participants' performance in the offline test was used as a reference. The same 4 participants who scored less than or equal to 80% correct for control trials in the 4AFC generalization test were excluded so that all the participants counted in the analysis (n = 29) met the baseline learning criteria. Data are plotted separately for the control, high and low frequency trials in the time course graphs shown in figure 7. Data for each of these trials were analyzed according to the pipeline in section 2.5.

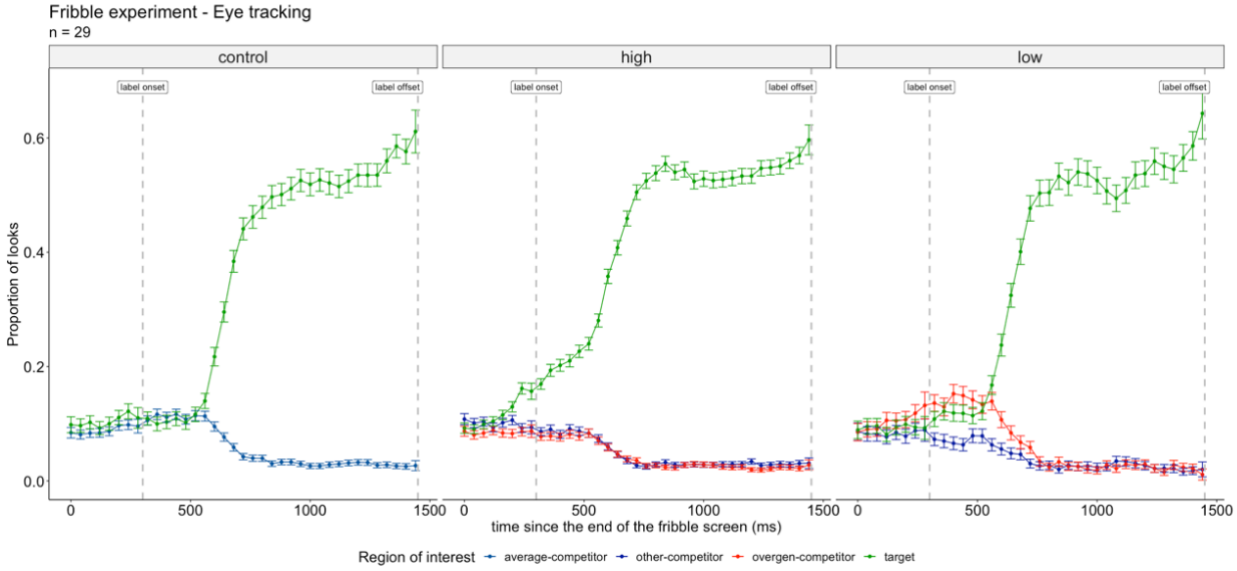


Figure 7. *Experiment 1:* Proportion of looks to each region of interest in the fribble experiment split by frequency. The label onset and offset are signaled at 300ms and 1450ms with the dotted vertical lines. The continuous data were binned every 40ms, with error bars marking 95% confidence interval (CI).

a) "Did the behavior in control condition meet default expectations?"

The default behavior should be demonstrated by a higher proportion of looks at the target regions than the other regions (competitor regions) starting from 200ms after the label onset (allowing 200ms to program an eye-movement; Hallett, 1986). This

[2] Although the eye-movement was tracked from the start of the fribble screen, the data before fribble offset were excluded because no predictions made were related to the object screen. Thus, data analysis treated fribble offset as 0ms for the time line of the eye-tracking data.

difference in fixation towards the target regions should be significantly higher when comparing to the difference in looks at the two regions before label onset.

Looking at figure 7, the lines show a difference in the proportion of looks to the target and two competitor regions within 200ms of the label onset (300ms – 500ms) in the control condition, indicating that participants have met the baseline assumption to proceed data analysis for the other predictions. This observation was confirmed in the results of statistical analysis. The number of looks at the target region was compared to that at the competitor region for both before (0ms – 500ms) and after label onset (500ms – 1450). To see if there was a greater proportion of looks to the target region in each time window, empirical logit transformation (Barr, 2008) was used to calculate the difference in gaze probability for the target region over the average of the two competitors. The difference value would be 0 if participants looked equally at target and competitors in this region, positive if they look more at the target, and negative if they look more at the competitors. T-tests were applied to examine if this difference value was significantly above chance. One-sample t-tests showed a non-significant difference in proportion of looks between competitor and target before label onset ($t_{(29)} = 0.91$, $p = .37$), but the results were significant for the post label onset period ($t_{(29)} = 13.21$, $p < .001$). A following paired t-test also showed that these values were significantly different in the two windows ($t_{(29)} = -11.438$, $p < .001$), confirming that participants had meet the default expectation of looking more at the target label after it appeared.

b) "Did they predict the upcoming label on the basis of the high frequency cues?"

After having observed the default behavior in (a), the inspection can proceed to checking predictive behaviors. For high frequency trials, the participants should look at the target position more than either of the competitor label position in the time period up to 200ms after the label appears (again allowing 200ms to program an eye-movement). Meanwhile, there should be no difference in looks to the two competitors.

Turning to the data plotted for high frequency trials (figure 7), the proportion of looks for the target region seem to be consistently higher than that for the two competitor regions even before label onset, suggesting that participants were making predictions about the upcoming label during the learning phase. To confirm this, we focused on the time frame before label onset (0ms – 500ms). Proportion of looks were aggregated for each of the region of interests (target, overgeneralization competitor and the other competitor) and gaze probability difference scores were computed for the differences between target and overgeneralization competitor, target and the other competitor, and the two competitors. Sample t-tests were used to investigate if these differences were above chance. Results showed significant differences between target and both competitors (target vs. overgeneralization competitor: $t_{(29)} = 3.42$, $p < .05$; target vs. the other competitor: $t_{(29)} = 2.28$, $p < .05$), but not between the two competitors ($t_{(29)} = -1.63$, $p = 0.11$). The results thus met the prediction for high frequency trials and confirmed that participants did show the predictive behaviors of looking at the target labels before they appeared.

c) ***"Did they learn the low frequency cues?"***

Since (a) and (b) are met, the analysis run on the high frequency trials can be repeated with the low frequency condition. Specifically, participants should look more at the target than either of the competitors in the region up to 200ms after label disappears. However, due to cue competition from the high saliency features (body shape and color) that occurred more often with the alternative label, the participants might show some overgeneralization (especially since the time of learning is short), and therefore look more at the overgeneralization competitor – the one that shared the high saliency features with the target label – than the other competitor.

To test this, the same difference scores were computed as for high frequency trials and the same t-tests were run to compare these scores to chance. The one sample t-tests showed no significant difference between the target and the overgeneralization competitor ($t_{(29)} = 0.14$, $p = 0.89$). However, significant effects were found both for

the difference between the target and the other competitor ($t_{(29)} = 2.58$, $p < .05$), and between the two competitors ($t_{(29)} = 2.20$, $p = < .05$). Since participants looked significantly more at both the target label region and the over-generalization competitor than the other competitor, they were showing evidence of predictive saccades to both regions, and therefore should have taken into account both the low frequency discriminative cue (which predicts the target) and the high salience uninformative cue (which predicts the overgeneralization location).

Follow up analyses on low frequency trials

Although there was evidence of prediction making on the basis of both discriminative cues and salient non-discriminative cues, these effects were averaged across 29 participants. Recalling the offline data which showed extensive individual variation (figure 6) that some participants obtained above-chance accuracy for low frequency trials but some did not, the sub-group of participants who performed poorly in the 4AFC could be the same group who were overgeneralizing during learning. In addition, the eye-tracking data plotted in figure 7 and analyzed with the t-tests was also averaged across trials of the experiment. Since participants were learning through each trial, there could be difference in how much they overgeneralize at early or later stages of the learning phase. Therefore, we looked into the performance during training for participants who learned the items in low frequency condition by the end of the learning task (as measured in the offline test), and the ones that did not, and how their performances changed over time. The two learner groups were categorized by looking at the proportion of times they chose the correct label over the overgeneralization competitor in the offline AFC test for low frequency items. Participants who scored above chance were classified as learners of the low frequency cues. Those who scored below chance were grouped as the over-generalizers. In addition to the one participant who did not have a score for the AFC test, five more participants were removed because their responses were inconsistent with those of a separate contingency test (see footnote 1, section 3.1.2). The details of coding and categorization are provided in Appendix 3.

Since the categorization identified only 10 learners and 13 over-generalizers in the dataset, no statistical analyses were applied, and the eye-movement data were inspected in a plot (figure 8). The learning phase was separated into 3 stages to capture the difference in participants' performance over time. Looking at the graphs for stage 1, participants who are over-generalizers in the offline test seem to have a strong tendency to look at the overgeneralization competitor position from the start – the line for overgeneralization competitor (red line) looks higher than the other two lines around label onset (200ms – 500ms). Looking across the three stages, it seems they never stopped this overgeneralization throughout the whole learning phase. Conversely, those who have been confirmed by the offline test to be successful learners of the low frequency cues did not make any predictions in the early stage 1. However, in stage 2, they appear to show some overgeneralization, with more looks at the overgeneralization competitor (red line) as well as at the target (green line). By stage 3, they seem to have retreated from making the wrong predictions and started looking more at the target label from 300ms.

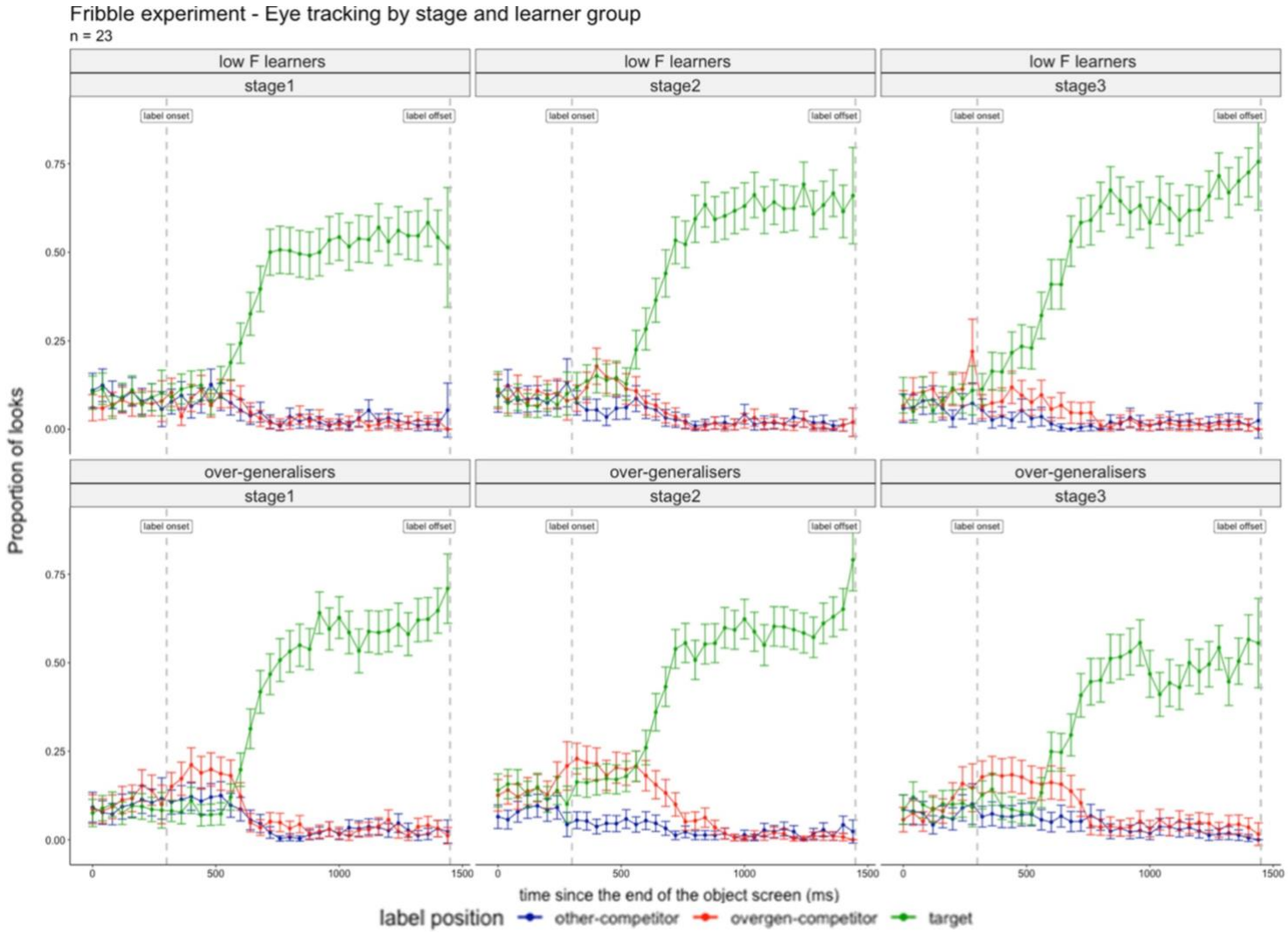


Figure 8. *Experiment 1:* Proportion of looks to each region of interest split by learner group and experiment stage (stage1 = trials 1-55; stage 2 = trials 56 to 110; stage 3 = trials 111 to 160). The label onset and offset are signaled at 300ms and 1450ms with the dotted vertical lines. The continuous data were binned every 40ms, with error bars marking 95% CI. Low frequency learners n = 10; over-generalizers n = 13.

3.3 Discussion

Experiment 1 examined on-line behaviors for discriminative learning by tracking adults' eye-movements in a replication of the object-label design of Ramscar et al.(2010). Starting with findings of the 4AFC offline test, we expected the participants to learn both the high and low frequency objects for the corresponding labels. With the input order of seeing objects first, learners could benefit from prediction errors and cue competition to dissociate the uninformative cues. However, while the offline test showed strong learning of the high frequency objects, as a group, participants did not generalize the low frequency cues significantly above chance. This is different from Ramscar et al. (2010), which showed strong learning effect in the low frequency condition (78% in their study; 41% in this experiment).

Turning to the eye-movement data, we looked at fixations on the label positions before and after label onset, with the goal of pinpointing when participants made predictions about the upcoming label for the given fribble. The on-line behaviors were checked through a pipeline analysis. Firstly, the baseline behavior was established in the control condition: participants looked at the target positions for control labels significantly more after the label onset and not before because the control labels were not predictable. This identification confirmed that eye tracking for the experimental design was functioning as expected. Thus, the analysis looking for predictive eye-movements was licensed. For the high frequency trials, a significant difference was found in fixations on the target position and each of the two other positions before label onset, indicating that participants were correctly predicting the target label based on high frequency cues. For the low frequency condition, participants looked more at the position for the target label than for the competitor classified as "other" (i.e. "wug" in figure 9), but they also looked more at the competitor categorized as "overgeneralization label" (i.e. "tob" in figure 9) than at the "other" competitor. Meanwhile, they did not look more at the target than the overgeneralization position. This pattern suggests that they were predicting the upcoming label based on both the discriminating cues (predicts the target) and the body type feature (predicts the overgeneralization label).

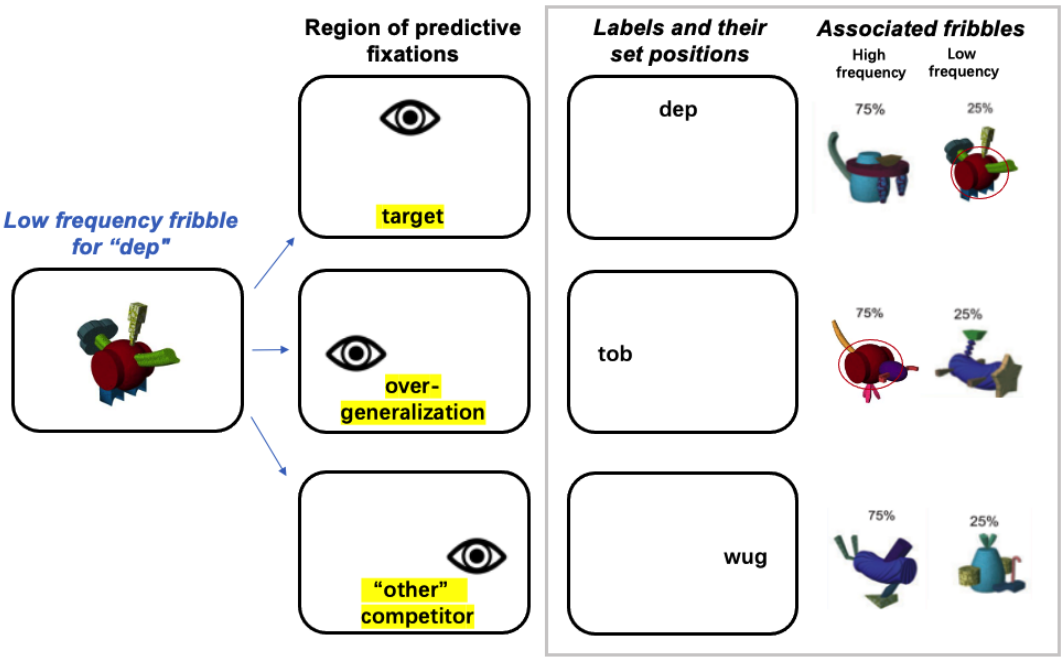


Figure 9. *Experiment 1:* The three competing regions of fixations following the low frequency fribble of label "dep". The set positions for each test label are shown as reference to indicate the label predicted by each region of fixations. The associated high and low frequency fribbles are listed to indicate the difference between "over-generalization" and the "other" competing regions. Note that for label "dep", the low frequency fribble shares the body type, a salient feature, (circled in red) with the high frequency fribble of "tob". This makes the fixations at the "tob" position an indication of overgeneralization of the salient but non-discriminating cue, the body type feature. Since fribbles of "wug" do not share features with the low frequency fribble of "dep", fixation at the position of "wug" is considered the "other" competitor.

These eye-tracking data are in line with the offline test which showed stronger learning of the high frequency items than low frequency items. The difficulty in generalizing with low frequency cues as seen in the anticipatory fixations for *both* the target *and* the overgeneralization labels was also reflected in the offline test with the low performance for low frequency items. It is possible that due to the limited learning time, the low frequency cues did not accumulate enough predictive values to be significantly discriminated from the uninformative features. On the other hand, there were considerable individual differences in the performances of offline test. The participants were then split into two learner groups according to their offline performances to investigate whether some were overgeneralizing and some were not, and how this

changed over the experiment as they learn from the prediction errors. Given the small data sample, the proportions of looks to each region of interest in the three stages were plotted for the two groups but no inferential statistics were conducted, making these observations exploratory and conclusions tentative. The plot showed that over-generalizers tended to make predictive looks to the overgeneralization positions even in the first learning stage, and this continued throughout the training phase. This may suggest that they were not learning from prediction errors. In contrast, the learners of the low frequency cues showed a more classic pattern of cue competition: they did not start making the predictions based on overgeneralizing cues until later – the middle stage of training. By the end, they reduced overgeneralizing and seemed to have discriminated the low frequency cue from the salient distractor.

In sum, the data showed that learners in these experiments *did* make predictions. This finding is in line with the account of discriminative learning, in that predictions allow learning to test out how accurate the proposed object-label mapping is and adjust the predictive value of it at the cost of other competing cues. However, most participants did not learn the low frequency cues in this experiment, contrasting with the previous experiment in Ramscar et al. (2010), and the plots of the different participant groups suggest that these participants did not retreat from overgeneralization. It would be interesting to know why the over-generalizers did not learn from their predicative errors even though they seemed to have made predictions before the learners did. This discussion will be continued in chapter 5.

To further establish these results, it is important to replicate them with a new sample and ultimately a larger one. However, rather than repeating with fribbles, Experiment 2 was conducted in a new domain. Note that the process of discriminative learning is supposed to be domain general. Repeating the previous experiment within this different domain could therefore help increase the validity of existing findings if results could be replicated. Instead of learning to associate labels with complex objects in a way that involved ignoring some part of the objects (i.e. body type of fribbles) and focusing on

others, Experiment 2 used a new paradigm where participants would learn simple objects, but learning their meanings involves ignoring other objects in a complex scene. This experimental paradigm is called “cross-situational word learning” (CL).

Another motivation for Experiment 2 is that CL would make it possible to include Ramscar et al. (2010)’s manipulations of input order to see how the prediction making process differs across conditions in which discriminative learning is thought to occur (object-label) or not occur (label-object). This is an advantage over fribbles, since their design as a visually compact object is not suitable for eye tracking in terms of seeing where participants look at the features. In contrast, the separated individual objects as cues in CL would make it more possible to track eye-movements in the label-object condition. However, it only makes sense to investigate the label-object condition once the domain is at least established for the object-label condition (since DL theory makes the clearest predictions under the object-label order). Thus, Experiment 2 focused again on testing the object-label condition and aimed to replicate Experiment 1 in a new paradigm.

4. Experiment 2 – cross-situational learning

Yu & Smith (2007) introduced the CL paradigm as a way of explaining how correct word-referent mappings are made when there is referential ambiguity (i.e. when objects are encountered as part of a scene). They noted that learning word-referent mappings occurs through repetitive encounters across trials of learning, not at a single exposure. Although not generally discussed in terms of DL for language acquisition, the paradigm aligns with the theory that the uncertainty in environment could be reduced by using prediction errors to determine which parts of the scenes are *not* relevant for particular words. As CL proceeds, participants see varying contexts in which a target referent is presented together with other competitors. They could potentially start mapping by making predictions with possible referents and gradually adjust their predictive values with the co-occurrence information in each trial. Uncertainty reduction for learning words in context could thus be operated in the same way as was done for learning word categories in the previous experiment.

In the experiment of Yu & Smith (2007), adults participants were presented with a set of trials where they viewed multiple object pictures and heard a single spoken word, establishing the CL paradigm. No specific information about the word-referent mapping was provided but the words were always paired with their corresponding objects, although other distractors were also present. After the learning phase, a 4AFC test was used to see if participants could select the correct object for a given word from 4 options. Although there were three different manipulations in terms of how many objects and words were presented for each trial, participants could consistently choose the correct object above chance under all three conditions. The results suggested that adults can learn word meanings in the setting of CL. A second experiment in the same paper replicated the results with different numbers and repetitions of target words. Other studies such as Dautriche and Chemla (2014) and Zettersten et al. (2018) implemented the paradigm to investigate the effect of context and co-occurring frequency with the referent objects on word learning. Dautriche and Chemla (2014)

found faster learning of words with richer contextual information (e.g. coherent presentation of animal pictures or repetitive co-occurrence of tomato and knife). Zettersten et al. (2018) noticed learning of the co-occurring relationship between objects alongside mapping the objects with words. To date, there is one study that applied eye tracking on top of the CL paradigm – Roembke and McMurray (2016) looked at where participants were fixating in the visual scene and used this information to see how participants were processing for word learning, specifically whether they were actively testing hypothesis of possible meanings for a word compared with accumulating distributional regularities. Although the current experiment, a replication of Experiment 1 translated into a CL design, also used eye tracking to investigate word learning process. It was different from Roembke and McMurray (2016) because we were interested in seeing prediction making in the DL process and this would be demonstrated by the participants' saccades at where label would appear after a given visual scene.

Other than the CL paradigm in Experiment 2, the design was similar to Experiment 1 and Ramscar et al. (2020), however, one change given the potentially greater complexity of objects in scenes was that instead of having six visual categories and three labels, there were four visual categories and two labels. A similar simplification was also adopted in Vujović et al. (2021) which used two labels and four categories of fribbles and successfully replicated the results of better learning under the discriminative input order as in Ramscar et al. (2010). In addition, Experiment 2 was extended to two sessions to prolong the learning time as a result of the low learning effect obtained for piloting.

4.1 Methods

4.1.1 Participants

34 participants were recruited through Prolific (www.Prolific.co), and among colleagues and students in University of Oxford. All participants were proficient English speakers over 18 years of age, with normal or corrected to normal vision, and

none reported any history of neurological or learning impairments. They provided consent to participate and were paid for their time.

4.1.2 Materials and design

For the CL paradigm, we created varying scenes with unknown figures and asked the participants to learn two word labels ("wug" & "tob") and one control label ("bim"). The visual stimuli used to create the scenes were taken from the Novel Object and Unusual Name database which consists of 64 photos of real-life items that viewers cannot easily name (i.e. the Novel Object and Unusual Name Database; Horst & Hout, 2016). Each of "wug" and "tob" consistently followed one of two "discriminating" objects which occurred in a fixed position in the scene and was accompanied by four other objects (see figure 10). Similar to Experiment 1, the control label "bim" was still associated with the blue color cue and additional offline tests[3] were conducted after training.

[3] In addition to the AFC tests for generalization and pairing described in 4.1.3, a contingency judgement test was also conducted for Experiment 2 as in Experiment 1, where novel scenes were paired with the labels and were rated. The test was not reported in the current paper for the same reason as the one in Experiment 1. See footnote 1 (section 3.1.2).

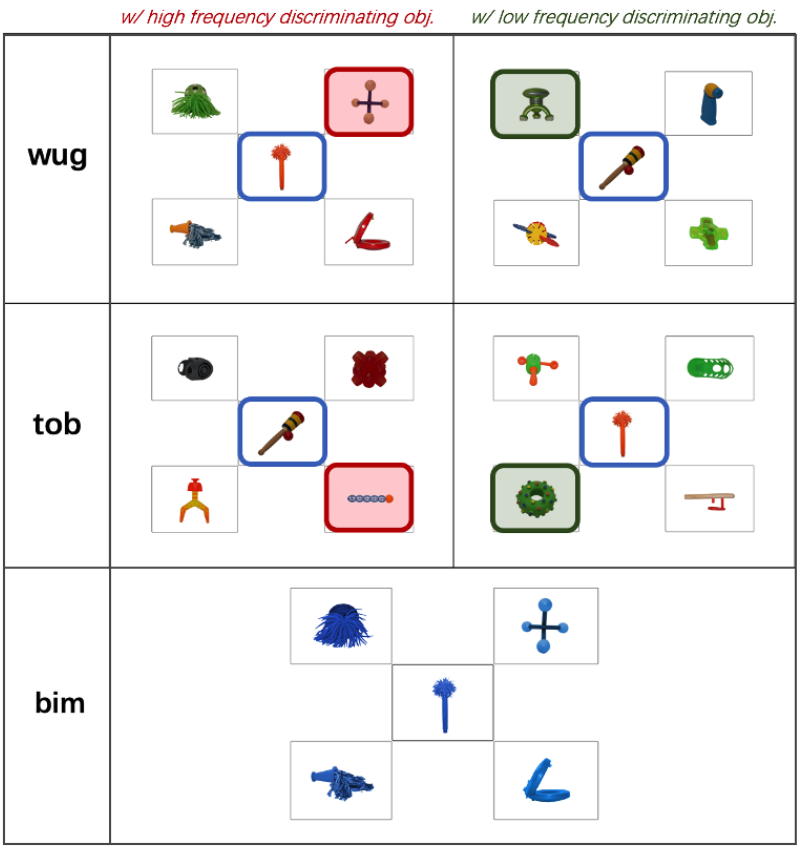


Figure 10. *Experiment 2:* Demonstration of the stimuli set. High frequency discriminating objects are marked in red, low frequency discriminating objects in green. The center objects circled in blue are two high saliency distractors used across the two label categories (colored circles are for presentation purposes only and were not viewed by the participants). The combinations of discriminating objects and high saliency distractors in association with the labels were fixed, whereas the other distractor objects vary randomly across trials.

4.1.3 Procedure

Experiment 2 was covered under the same ethical approval as Experiment 1 (see 3.1.4) and followed a similar procedure (figure 11). In addition to a 3AFC test using novel scenes which was equivalent to the 4AFC test with the fribbles, an additional 3AFC pairing task was included to test learning of individual objects in isolation as cues to the word labels.[4] The CL paradigm was piloted twice (N=7; N=20) to test the functioning of eye-tracker and the length of training needed. It was found that longer training was necessary and the final experiment contained repetition in a second session

[4] This test was not possible with fribbles because the discriminating features of fribbles could not be tested in isolation.

to provide more exposure. Participants were therefore asked to come back to the experiment after 12 hours once they had completed with the first session. The two offline tests were included for both sessions to test learning progress in the generalization and pairing scores. The two-part experiment as a whole took around one hour to complete.

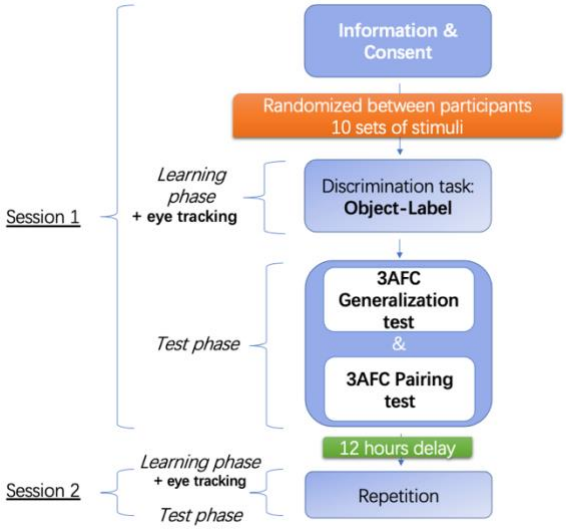


Figure 11. Procedure of Experiment 2.

Learning. There were 144 trials in total for each session, specifically 96 trials for the two high frequency objects for each word label, 32 trials for the low frequency objects, and 16 trials for the controls (table 2). Each session started with the calibration (as in Experiment 1) and participants' goal was to discover the reference objects that paired with the two novel words. The trials in each session were separated into 5 blocks with breaks and recalibration in between.

task	control	High frequency	Low frequency	Total for session 1	Total for session 2
learning	16	96	32	144	144
3AFC Generalization	8	20	20	48	48
3AFC Pairing	5	4	4	17	17

Table 2. *Experiment 2:* Trial arrangement. 4 more trials were included in 3AFC pairing to test central objects in isolation, which added up to 17 total trials in both sessions.

In each trial, participants were initially presented with a scene at 1000ms and then were shown the corresponding label at either the left or right side of the screen at 3500ms (figure 12). As in Experiment 1, the control label can again appear in either location. 10 lists of stimuli sets were constructed to randomize label locations, object locations and the object used for each participants. The discriminating objects, salient distractors and word labels were kept consistent throughout the two sessions for each participants, including the offline tests as well.

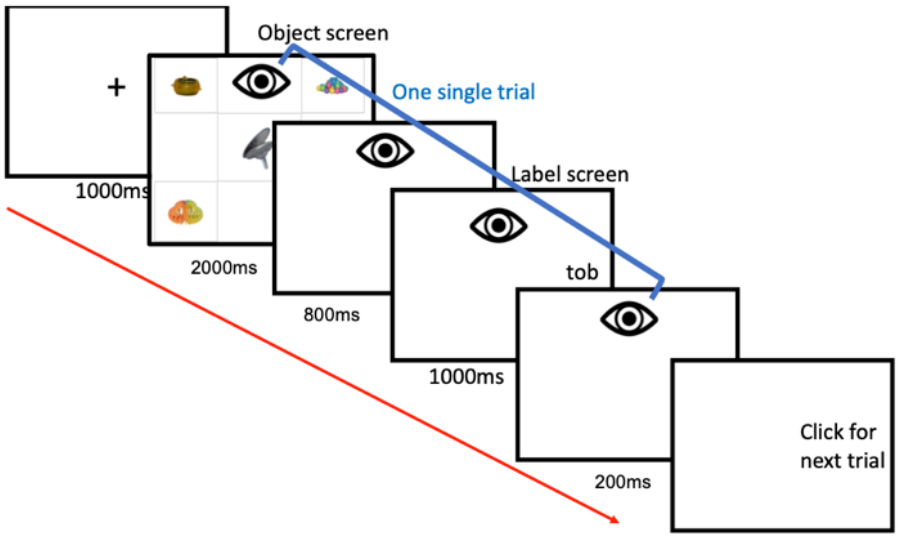


Figure 12. *Experiment 2:* Learning phase set up and timelines. The time frames where eye-movements were tracked are marked with the eye symbol. The duration of each screen is marked below.

3AFC generalization. This test was implemented to test participants' ability to generalize the object-label pairing they learned to a new scene. The novel scenes used as stimuli were in the same fixed combinations of discriminating and salient distractor, but with new combinations of the random objects. For each scene, participants had to choose among three labels to decide which word fitted the scene the best (figure 13, left). The test consisted of 48 trials in each session, with 8 trials for the controls and 20 trials for each frequency condition (table 2). Each trial had 3.5 seconds response time and participants were alerted from the start.

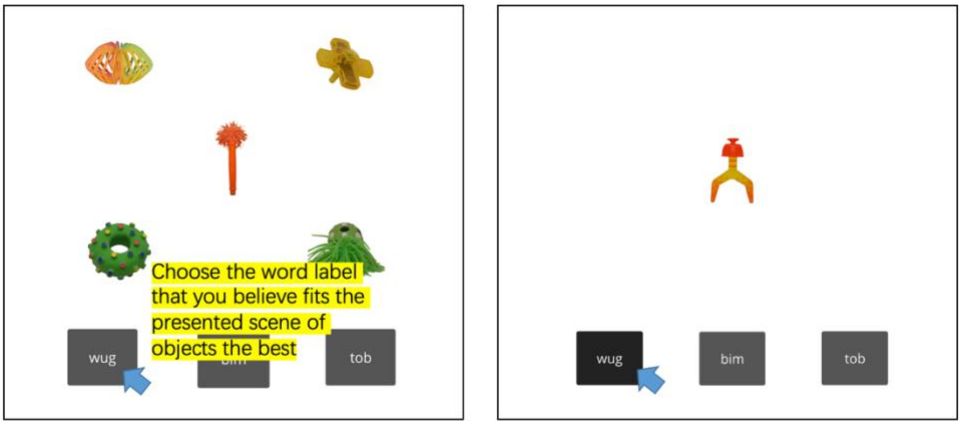


Figure 13. *Experiment 2:* Interfaces for the test phase tasks.

3AFC pairing. The pairing test was different from the generalization test in that only the target object was included, positioned in the center of the screen (figure 13, right). Participants were again asked to choose a best fit among three labels. The test included trials for the 5 blue control objects, two high and low frequency objects for each test label, and the central salient distractors. Altogether there were 34 trials for the two sessions (table 2). Each trial was timed for a response time of 3.5 seconds.

4.2 Results

4.2.1 3AFC generalization

The data analysis for experiment 2 was done in R (R Core Team, 2020) and started with

the offline tests to obtain a general understanding of participants' learning. Data filtering removed three participants who did not complete both sessions of the two-part experiment. Further inspection of accuracy for the control condition in 3AFC generalization test (i.e. trials where the correct answer should be "bim") revealed that another 9 participants scored less than or equal to 80% correct. As with data cleaning in the fribble experiment, these low performance participants were excluded as well, leaving 22 participants valid for data analysis.

As the 4AFC test in Experiment 1, test trials with "bim" as the target were removed. For the remaining test trials (where "wug" and "tob" are targets), since the number of trials in which participants picked the control label "bim" made up only 0.7% of the total trials, these trials were removed so that chance level of performance was 50% instead of 33%, allowing a direct presentation of the test performances. The timed-out trials that comprised 2% of the remaining test trials were also excluded to ensure data quality.

It can be seen from a violin plot with included boxplot visualization (figure 14) that performance in the high frequency condition was higher than that for low frequency items in both sessions. This was confirmed statistically by using a logistic mixed effects model to examine the effect of frequency and session on accuracy. Frequency was a fixed factor and separate fixed effect slopes for session were fit for each of the high and low frequency conditions. Participant was also included as a random factor as well as random by participant slopes for within participant effects (i.e. frequency, session and the interaction of frequency by session). As predicted, there was a significant main effect of frequency (beta = 2.54, SE = 0.64, $p < .001$), showing that participants demonstrated overall higher performances in high frequency trials (88% correct) than in low frequency trials (50% correct).

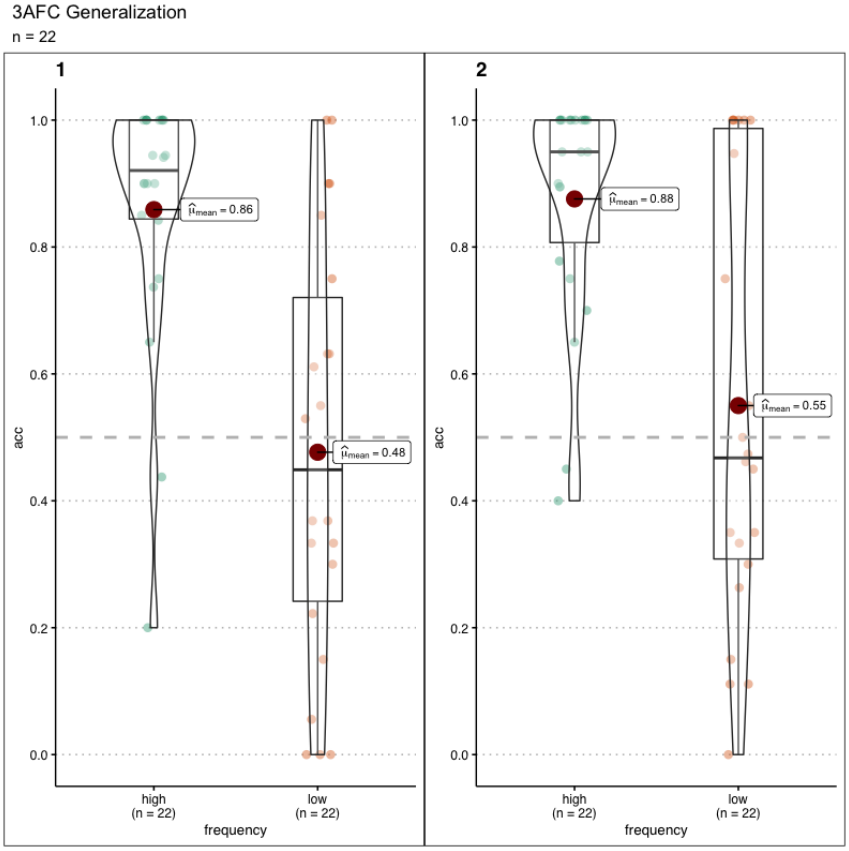


Figure 14. *Experiment 2:* Proportion of correct trials in the 3AFC generalization test split by frequency and session. Points show the mean accuracy by participant, with the kernel probability density demonstrated by the violins. The boxes mark the interquartile range around the median of accuracy for each frequency conditions which are marked with the black lines. The grey dotted lines show chance level of 50% since control trials and trials where participants chose the control item were removed.

There was no significant positive slope for session for high frequency (beta = 0.52, SE = 0.61, p = .40), but the general high performances suggested a ceiling effect. For the low frequency trials, effect of session was significant (beta = 0.98, SE = 0.48, p < .05). Two further models were run over the low frequency trials for each session to see if there was a significant intercept, which would indicate that performance was above chance, however no significant result was found for either sessions (beta = - 0.19, SE = 0.50, p = .70; beta = 0.78, SE = 0.62, p = .20).

4.2.2 3AFC pairing

Due to technical issues, the data from the 3AFC pairing test was not recorded for one

participant so that this analysis was conducted on the data collected from 21 participants. The same participants as for the generalization test were removed as were control trials where the answer should be the label "bim". For the remaining trials, participants wrongly picked the label "bim" instead of "wug" or "tob" in 2% of the total trials – these were also removed so that the performance chance level was 50%. Moreover, another 0.9% of test trials were removed due to participants lack of response. The data are presented in Figure 15.

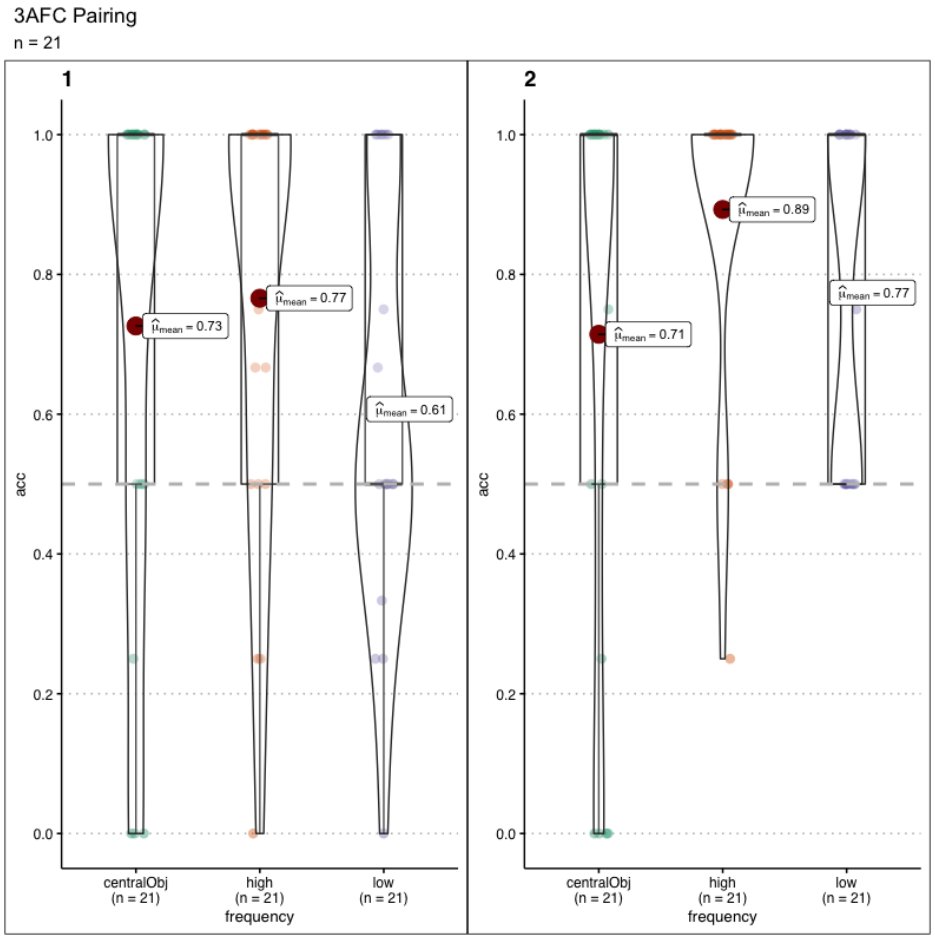


Figure 15. *Experiment 2:* Proportion of correct trials in the 3AFC pairing test split by frequency and session. Points show the mean accuracy by participant, with the kernel probability density demonstrated by the violins. The boxes mark the interquartile range around the median of accuracy for each frequency conditions which are presented as the black lines. The grey dotted lines show chance level of 50%. Data from 21 participants were plotted.

Figure 15 shows a consistent higher accuracy in high frequency trials when compared to the performance in low frequency trials. Meanwhile, performances for both high and

low frequency trials seemed to have improved in session 2. Note that the graph shows the data for both items for the two frequency condition and the central objects which were highly salient in the learning phase. Since the core hypotheses of the offline tests were not interested in the central objects, we removed these trials from the 3AFC pairing data set and investigated the same questions as for the 3AFC generalization.

Statistically, the performance accuracy for 3AFC pairing was also analyzed with a logistic mixed effect model, with frequency as the fixed factor and separate slope fit for the fixed effect of session for each frequency condition. The random effect structure included intercepts for participants and a by participant slope for frequency (the slopes for session and interaction between slope and session were excluded to allow the model to converge). There was a significant effect of frequency (beta = 2.44, SE = 0.10, $p < .05$), which suggested a consistent better performance in high frequency trials (82% correct) than low frequency trials (68% correct) in line with the 3AFC generalization test. The results confirmed that participants improved across sessions for pairing high frequency items, which was demonstrated by a significant positive slope for high frequency trials (beta = 1.69, SE = 0.58, $p < .05$). For low frequency trials, there was a marginally significant slope for session (beta = 0.66, SE = 0.35, $p = .06$). Additional models were run with data for low frequency items in each session to see whether the intercept was above chance. These showed that participants did perform significantly above chance for the low frequency items at session 2 (session 1: beta = 0.52, SE = 0.27, $p = .052$; session 2: beta = 1.32, SE = 0.41, $p < .05$).

4.2.3 Eye-tracking data

The analysis of eye-tracking data for CL was guided by the same inspective pipeline as for Experiment 1 (section 3.2.2).

a) "Did the behavior in control condition meet default expectations?"

The participants should look at the region where the control label is located on the

screen *after* the label appears.

b) "Did they predict the upcoming label on the basis of the high frequency cues?"

For high frequency trials, participants should look at the target position more than the competitor position around the time frame *before* label appears.

c) *"Did they learn the low frequency cues?"*

For low frequency trials, participants should show the same fixation difference towards the target position as for the high frequency trials. They may also overgeneralize the salient central distractors as a cue for the following label and thus fail to look more at the target position.

Data was collected from the object offset until 200ms after label offset, which included the label onset at 800ms [5]. A very small number of trials were excluded because the screens were shown for too long (likely due to an error on the participants' computer). However, the data removed was less than 3% of the total data and did not affect the critical time window.

Data filtering was the same as for the offline tests. Three participants were removed due to incompletion of the full experiment and 9 more were excluded due to low performance for the control condition in 3AFC generalization test. Since there was another participant whose eye-tracking data was not successfully recorded, there were in total 21 subjects left for further analysis.

Figure 16 shows the time course of fixations on the target and competitor regions collected for each trial in the learning phase. Starting with the graphs for the control condition, the expected baseline pattern whereby participants clearly looked at the

[5] As in Experiment 1, the eye-tracking data before the object offset was excluded since it was not related to the predictions tested for this study. The object offset originally at 300ms was then treated as 0ms for the eye-tracking time line.

target more than at the competitor once the label had appeared seemed absent in both sessions – the lines for proportion of looks on the target region (red) were overlapping with that for the competitor region (green). To follow the pipeline analysis, the observation was checked statistically to decide whether further predictions would be investigated.

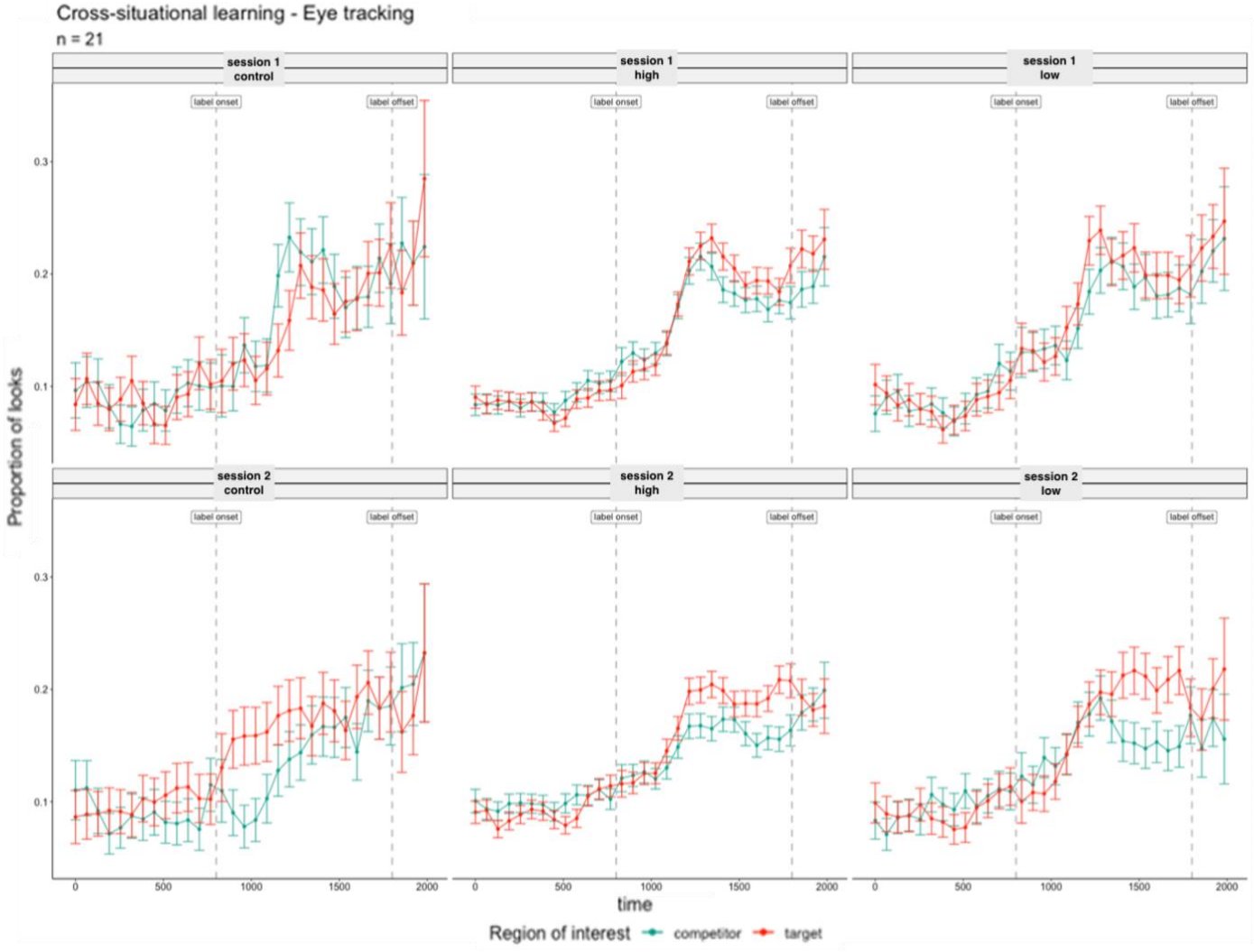


Figure 16. *Experiment 2:* Proportion of looks to each region of interest split by frequency and session. The label onset and offset are signaled at 800ms and 1800s with the dotted vertical lines. The continuous data were binned every 64ms, with error bars marking 95% CI.

For each session, the number of looks at the target region versus at the competitor in a control trial was extracted for two key windows: trial onset (0ms) to post label onset (1000ms, allowing 200ms after label onset at 800ms for the programming of an eye-movement) and post label onset (1000ms) to trial offset (2000ms). We are interested in whether there are a greater proportion of looks to the target than the competitor in each time window (a difference score). To meet the normality assumptions of our inferential statistics, the empirical logit transformation following Barr (2008) was applied to this difference score. Starting with session 1, one-sample t-tests showed that this difference in looks between competitor and target was not significant (i.e. not above 0 which is chance) either before or after label onset (before: t (20) = -0.41, p = .70; after: t (20) = -1.04, p = .31). The differences in looks at each label position for the two time windows was also compared using a paired t-test and was not significant either (t (20) = 0.61 , p = .55). Moving on to session 2, the same t-test results were also not significant (before label onset: t (20) = -0.02, p = .98; after: t (20) = 1.30, p = .21; before vs. after label onset: t (20) = -1.1 , p = .28). Together with the results of session 1, the lack of fixations towards the target region at label onset for control trials suggests that the participants were not processing the stimuli as expected (they were not looking at the label even when it was on the screen). Visual inspection does not suggest they were looking more at the target than the competitor region for the two frequency conditions either. Therefore, no further investigation was conducted on the trials of high and low frequency conditions.

4.3 Discussion

Experiment 2 aimed to show on-line predictive behaviors as evidence for DL in a CL paradigm. Starting with the offline tests, the 3AFC generalization test was a conceptual replication of the offline test in Experiment 1 (and Ramscar et al., 2010), thus a better learning performance in the high frequency trials was expected and the performance for the low frequency trials should be at least above chance. However, the current generalization test obtained similar results to Experiment 1 in finding the frequency difference, but the generalization performances for the low-frequency cues

were not above chance in both sessions, although there were again individual differences. Additionally, a significant effect of session was found here, which indicates that participants might have been learning the low frequency cues gradually over sessions, suggesting some decrease in overgeneralization.

The 3AFC pairing test was conducted to examine discrimination of the three labels paired with individual referents separately. While the previous test looked at learning of the relationships between the discriminative objects and the distractor object combined in one scene (equivalent to the compact fribble image in Experiment 1), 3AFC pairing looked at learning of the discriminative object in isolation. Sharing the same predictions as the other offline test, the task replicated the better performance for the high frequency trials and here found improvement in performances for both frequency conditions, although the improvement in pairing was only marginally significant for the low frequency objects. However, there was evidence of learning in the low frequency condition in that accuracy was above chance in session 2. Thus, it appears that participants do show some learning of the discriminative object-label mapping with more exposure. This preliminary evidence could support the extending of DL mechanism to this new domain of CL. The difference between the two tests also suggests that when the distractors are present, the difficulty of the generalization task might have counterbalanced the present effect of learning. Particularly, the predictive values for low frequency objects were not high enough to let participants ignore the high frequency central objects. Moreover, in line with findings in Experiment 1, the performance for low frequency trials in both 3AFC generalization test and session 2 of the 3AFC pairing test demonstrated a lot of individual variations, and more so than their high frequency counterparts. These observations are considered further in the general discussion in chapter 5.

The analysis for eye-tracking data followed the pipeline order and control data were analyzed first. However, unlike in Experiment 1, participants did *not* show the expected default pattern of eye-movements in either session – they did not clearly

look more at the label once it appeared on the scene. Therefore, no more analysis was continued as any behavior pattern observed for the trials in frequency conditions could be spurious. It was thus not possible to see if participants conducted predictive eye-movements in training. Also recalling that the CL paradigm was originally adopted for applying eye tracking in both an object-label and label-object order to investigate DL, the paradigm at present is unfortunately not suitable for this extension until positive results for baseline behaviors can be found in this current object-label condition.

Why did participants not clearly look at the label position even once it appears on the screen? One possible answer is that the web-based eye-tracker is malfunctioning due to the use of an updated version of Webgazer.js in Experiment 2. Alternatively, the time frame between object offset and label onset (800ms + 200ms) might not be enough for participants to switch their fixations to the label, due to the fact that the previous visual scene was more complex under the CL paradigm. The participant could also need longer than 200ms to program their eye-movement after seeing a more complicated presentation of objects. In addition, when comparing the time frame setting with previous experiments, it is possible the label screen was proportionally too short: they were 1000ms for the current experiment compared with 1150ms in Experiment 1 (and in Ramscar et al., 2010). This design was settled on the basis of pilot results in order to be able to include more trials within the total learning time, but the experiment may have lost the ability to see the effect in eye-movements by making those changes regarding how long each screen lasts. However, we cannot determine whether the absence of default behavior is due to any particular reason such as the use of different version of Webgazer, changing the timeline of presentation, or a combination of the reasons above.

Since the lack of eye-tracking evidence could be due to various problems with the paradigm, the key finding of this study is in the offline data. Participants did learn the labels from cross-situational exposures, at least to the extent that they could

generalize when there was no salient non-discriminating object present (i.e. 3AFC pairing). Given that we saw clear evidence of prediction making with eye-tracking data in Experiment 1, the fact that no similar evidence was shown in the new paradigm is likely a reflection of problems in our eye-tracking design rather than evidence against the DL theory. Moreover, the problems discussed above suggest that the new paradigm would be unsuitable for exploration of the label-object condition as it currently stands, hence this was not attempted for this thesis despite being part of the original motivation for this experiment.

5. General Discussion

The current study investigated DL which is a constructivist theory that may provide an alternative account to the nativist perspective of language acquisition. The theory suggests that the discriminative process of language learning may be driven by prediction making (section 2.3) and Ramscar et al. (2010) had provided indirect evidence of this process. Therefore, this study applied eye tracking to the design of Ramscar et al. (2010) and aimed at finding direct evidence of prediction making to support the DL mechanism in word learning tasks. Recalling that the nativist perspective highlighted the role of negative evidence in environmental input to constrain language development, DL theory can provide a mechanism for obtaining negative evidence: if learners can make prediction errors when pairing environmental cues with their outcomes, they can then incrementally adjust the predictive strengths of the candidate cues (Ramscar & Yarlett, 2007). The study followed this approach and expected to see on-line evidence of participants predicting linguistic outcomes from environmental cues while learning in an object-label setting of input specifically designed to capture for cue competition.

The theory was tested in two experiments. The first one was a direct replication of the discriminative condition in Ramscar et al. (2010), which included a learning phase where participants see complex objects followed by a label (with eye tracking now incorporated), and an offline test of generalization. The second one adapted Experiment 1 in a cross-situational learning paradigm to check consistency of previous findings across domains. Learning phase with eye tracking and two offline tests were conducted, namely a generalization AFC test and a pairing test. In offline tests of both experiments, we expected to see greater learning of discriminative cues in the high than low frequency conditions, but the low frequency condition should be at least above chance given the previous findings for learning of low frequency items in object-label ordering (i.e. Ramscar et al., 2010). The results of offline tests in both experiments indicated learning of high frequency cues with high level performance

of generalization, but the low frequency condition only achieved above-chance accuracy in the second session for 3AFC pairing task in Experiment 2. Since no comparing label-object condition was investigated in this study, we could not compare the two learning conditions and check if low frequency items were still generalized better after the discriminative order training. However, the offline results can be traced back to the learning phase performances with eye-tracking data. We hypothesized that predictive behaviors would be seen with fixations towards the target label position before the label appeared. The observation for Experiment 1 was consistent with the 4AFC generalization test and also showed direct evidence of prediction: The target positions were clearly predicted (i.e. prior to the label appearing) for high frequency trials. The low frequency condition showed prediction making for both target and over-generalization labels when comparing to the other competitor, but the target label was not distinguished from the over-generalization labels. These eye-tracking results demonstrate that the lack of accurate generalization with low frequency objects was not because the participants failed to learn anything about these objects in training or that they could not make predictions. In fact, they predicted two candidates from the low frequency objects, but they could not distinguish between them to perform accurate generalization. Consequently, exploratory analyses were carried out to categorize participants based on their performances in the generalization test, separating the learners who gained above chance accuracy for the low frequency trials and the over-generalizers who did not, and inspected their eye-tracking data across three stages of learning. The analysis suggested there were differences in how each type of participants make predictions over time – some *were* predicting the target position over the competitor by the end of training, and subsequently used these cues in the offline test generalization. Although the analysis is tentative given the small sample, these individual differences are potentially interesting in light of differences in L2 learning and so will be discussed in section 5.1.2.

Despite having obtained similar offline data in Experiment 2, findings of predictive

looking were not replicated because default looking behaviors could not be established in the control trials. The problems in the design implementing a new cross-situational learning paradigm for the experiment were discussed in section 4.3. Therefore, this chapter will focus on interpreting the learning performance for the low frequency objects in Experiment 1, and how this ties in to real time prediction. This chapter will first try dissecting the group performance by exploring the individual differences in learning. Then some implications of the findings will be highlighted, and finally the limitations of the study will be discussed.

5.1 Individual differences

5.1.1 The current study

A distinctive finding for the learning phase of the current study was that, when encountering high frequency fribbles, participants made accurate predictions for the upcoming labels, and when encountering low frequency fribbles, they made predictions based on both the low frequency informative cues *and* the salient uninformative cues. This finding in the eye-tracking data can directly confirm prediction making and supports a key assumption for the use of discriminative learning mechanism in a major aspect of language acquisition – mapping of words and their referents (Quine, 1960). However, the expected learning effect in the low frequency trials was not seen in the offline tests. There was a lot of individual variance in both experiments, though, and visual inspection of the eye-tracking in Experiment 1 suggesting that some learners may have used cue competition to stop making predictions with the overgeneralized cues, while others did not. Comparing low-frequency-learners (who showed above chance accuracy for the low frequency trials in the offline test) and the over-generalizers (who did not), we saw different patterns of prediction making during the course of learning. Although no statistical analysis was conducted due to the small sample size for each group, figure 8 suggested that over-generalizers made predictions based on the uninformative overgeneralization cues from the first stage of the training phase, whereas the learners only started making predictions at stage 2. Particularly, learners make predictions with both

overgeneralization cues and discriminative cues at stage 2, but by stage 3, they stopped predicting with the over-generalization cues and sticked with the discriminative cues only. In comparison, the over-generalizers continued to use the overgeneralization cues till the end of stage 3. These observations suggest that the different ways how participants engage with cues in the environment could make a difference in their learning. A general implication could be that language learners should not aim at finding single assumptions for a word's meaning when learning non-native languages. Instead, they should keep their mind open even though the words might remain ambiguous for a long time. In fact, lexical ambiguity is inherent to languages and more than 80% of frequency word forms can refer to multiple meanings (Rodd, 2020). The ability demonstrated by the learners of low frequency cues is essential to word learning in real language contexts. Thus, it will be meaningful to find out why some participants could not discriminate the low frequency informative cues as the others did.

According to the theory of DL, uninformative cues should be "unlearned" because the prediction errors will reduce the strength of the weights with the proposed labels. Since predictive values are adjusted at the cost of other cues, the uninformative cues are weakened when incorrect predictions are made based on these cues, and the other label associated with the discriminating cue is encountered instead. Following this procedure, the over-generalizers should encounter prediction errors from beginning of the learning process as they started making predictions with uninformative cues very early. These prediction errors should have weakened the salient uninformative cues for the over-generalizers more when comparing to the learners who started prediction at a later stage. The over-generalizers should have then unlearned the uninformative cues faster because they have made more adjustment to the predicative values, but this was not the case in Experiment 1. It seems that over-generalizers did not learn from their incorrect predictions even though they started predicting before the learners did. This unexpected finding could be possibly explained with potential problems in processing low frequency cues as a competitor for the salient

uninformative cues: Since the low frequency cues were discriminating for a certain word label, they should have gained positive values every time the participants make predictions with them. However, this is based on the fact they do receive equal opportunities as a candidate of discriminating cues in over-generalizers' learning process. Considering that every other aspect of the experimental setting were replicated directly from Ramscar et al. (2010), including the trial arrangement for the two frequency conditions, the only difference in the current study was that we included more trials for the control condition that implemented the single blue color as the cue for a word label (Ramscar et al., 2010: 30 trials in total; Experiment 1: 40). This control condition of label-object pair could have created bias among participants and made them believed that salient features like color is a typical informative cue. In this way, the predictive values of salient features might be adjusted with different weights when comparing to the other cues, which explains their reluctance to discriminate against the over-generalization labels.

Additionally, in real-life circumstances, it is usually a combination of variant features that discriminate a concept behind the label. Colors are common salient features overlapping across objects and thus are not always discriminating if uttered in single terms such as "yellow!". Colors can also be misleading for object identification since they are often partially informative in the case like "a yellow rugby shaped fruit" (i.e. lemon). Despite being not as reliable, they may still be able to receive overt attention from us when perceiving an object (Frey et al., 2008). This preference towards certain cues, developed either naturally or through the constant involvement of cue competition in the past experiences, could have made them weigh more than the other cues and therefore harder to unlearn.

Despite the possibilities listed above, the explanation for the different learning performances of participants is currently unclear. It is also possible that if the experiment went on longer, the over-generalizers could have retrieved from the salient cues. However, regardless of the interpretation, the study identified distinctive groups

with consistent differences throughout the on-line training and the offline tests. Given the findings, it will be interesting to consider whether these might relate to individual differences in adult L2 learning, which is discussed in the next section.

5.1.2 Adult L2 learning

L2 learning has long been found more difficult for adult learners with considerable individual variation compared with the more uniform competency reached in L1 acquisition (Johnson & Newport, 1989; Hartshorne et al., 2018). To account for this difference, many researchers have adopted a nativist point of view that when adults passed a critical period for setting universal grammar for L2, they could never learn the language as well as their L1 (Bley-Vroman, 1989; DeKeyser, 2000). However, from the perspective taken in this study, the competency difference could also result from potential problems some adults have in conducting DL. Logically, one possibility is that they might have problems making predictions compared with L1 speakers, meaning that the learning process does not provide enough value adjustment for the cues to map with words or other linguistic units. This account can be supported by the studies reviewed in section 2.4 which found that L2 learners made less predictions than the native speakers (e.g. Grüter et al., 2014; Mitsugi and MacWhinney, 2016). However, the current study speaks against this account because Experiment 1 has found clear evidence of prediction making for adult participants. Although in more complex languages, L2 learners could be making less predictions, this study provides another possibility to explain the individual difference – they might not learn from the predictions they made, just like the over-generalizers in Experiment 1. In this way, individual differences in learning presented by this study could support DL mechanism for language acquisition because the discriminative process allowed learners to have their own pace of learning. Moreover, unlike relying solely on a critical period of learning as the nativists do, the DL theory can potentially provide a more nuanced explanation embedded in the learning process for the performance difference in adult L2 learning, though, these results and interpretation are speculative at present due to the small sample and no inferential statistics could

be done for the individual difference analysis. This limitation is discussed in 5.2.3.

In sum, there were individual variation in participants' performance that we could not fully account for, and it will be interesting to see in future work whether the same participants who over-generalize in this type of experiment also find natural language learning difficult in an L2 context. However, the current findings contribute to the current literature by demonstrating that adults clearly can and do make predictions about upcoming words even after a small amount of exposure to a novel language. With that being said, L2 learners could try embracing the ambiguity of word meanings at the start of their learning process to prevent facing the potential problems in learning from prediction errors, which may occur if they are stuck with an overgeneralized cue.

5.2 Limitations

5.2.1 The artificial language setting

The current study employed an artificial language, and the use of this methodology has both benefits and limitations. In two experiments, a set of pseudo words (i.e. "wug", "tob", "dep" and "bim") was used as printed stimuli throughout the learning phases. This setting seemed to match with the "caricature" image summarized by Saffran (2009), which depicts a typical study of statistical language learning to be with a simple computation performed over simple elements in a simplified language. For studies like these, the potential relevance of such findings for real-life scenarios may seem far-fetched. However, the current study deliberately replicated the design of Ramscar et al. (2010). The artificial setting of using ambiguous words with high and low frequency meanings was constructed to allow ambiguity in the learning process so that cue competition could be investigated. The use of printed rather than acoustic word stimuli, while unnatural for spoken language acquisition, was essential to capture the internal learning process using eye tracking, so that "predictions" could have physical locations on a screen. With spoken words, we would not be able to tap the learners' mind by tracking their eye-movements. Moreover, Nixon (2020) and

Vujović et al. (2020) both used acoustic stimuli to investigate discriminative learning mechanism and found highly similar results to Ramscar et al. (2010), suggesting that the core learning is the same across modalities.

To sum up, the artificial language setting was necessary for the current study, but we must nevertheless be cautious in concluding that the results can be generalized for learning natural languages. It will be important to show with empirical evidence that the predictions of the theory can also extend to longer term, natural language learning. On this pathway of exploring learning mechanisms of language acquisition, studies with artificial language provide stepping stones to go from establishment of theories towards application in natural contexts.

5.2.2 Online eye tracking

Another salient limitation of this study is that no sensible eye-tracking data was obtained for experiment 2. While this might be due to the timing arrangement for the cross-situational learning paradigm, we must acknowledge that a new eye-tracking method was used. Webgazer as a javaScript algorithm embedded in the Gorilla platform is a brand-new technology and the current thesis is (to my knowledge) the first language learning study to make use of it. It detects facial information through web cameras and operate the predictions of real-time eye positions on computer browsers. This created a lot of space for confounding factors such as the individual difference in devices and environment (which also leads to participant exclusions). However, the biggest concern is still the precision of tracking. The traditional way conducting eye-tracking experiments in lab is to use equipment which processes the reflected infrared light from participants' eyes. Although both web-based and traditional eye-trackers translate eye-movement data from the relative positions of eyes and pupils, studies have shown differences in accuracy. Semmelmann and Weigelt (2018) compared in-lab and at-home eye tracking with the exact algorithm implemented in the current study with three classic behavioral tasks, fixation, pursuit, and free viewing. The online data in both the fixation and pursuit tasks was less

accurate than the data in lab. In addition, Madsen et al. (2021) applied the two eye-tracking methods to study performance in online education and demonstrated significant difference in gaze position accuracy for the lab and home settings. However, rather than aiming for precise depiction of fixations, the current study was interested in identifying predictive saccades towards certain parts of the screen. Both Semmelmann and Weigelt (2017) and Madsen et al. (2021) found that indication for this kind of patterns could be comparable for the two settings. The patterns seen in Experiment 1 are indeed another proof that web-based eye-tracker can be used to determine saccades making in behavioral tasks. Therefore, it is unlikely the absence of default looking behaviors in Experiment 2 was due to the lack of precision in the tracking of eye-movements. Despite that the participant-run online experiments are less controlled than the laboratory ones, with careful instructions and appropriate design, online eye tracking can lift boundaries and open up new possibilities for studies of language and cognitive science.

5.2.3 Sample size

Due to the sample size being too small after splitting the participants into two learner groups based on their performance in the generalization tests (n=10 learners of low frequency cues; n=13 over-generalizers), no inferential statistics was conducted for the individual variance observed in low frequency trials in Experiment 1. Any observation regarding the relationship between performance in the offline tests and prediction making during training would need further confirmation in a larger sample. However, despite that both the study and participant recruitment were conducted online, it took longer than expected to get as many participants as we would have liked for this experiment, because the eye-tracking algorithm had certain visual requirements on both the subjects and devices. Even after the data were successfully collected, it required a lot of cleaning to proceed to statistical analysis. Therefore, it was not possible to get a much larger sample for the current thesis, but with the experiences gained from the data collection process, the limitation could be managed in future studies.

6. Conclusion

The study conducted two experiments to investigate the DL process of language acquisition with two learning paradigms. Other than replicating the original design of the object-label order for input from Ramscar et al., (2010), the study applied eye tracking to the training sessions in both experiments to obtain evidence for the predictive behaviors. According to the DL account, learners get access to negative evidence in the input without innate constraints because they adjust the values of potential cues for the linguistic outcome by making prediction errors. Thus, during the training sessions, we aimed to observe predictive saccades towards the target label positions before it appears to directly corroborate DL with evidence of prediction making.

Although Experiment 2 did not generate meaningful eye-tracking data, the predictive behaviors were observed for both low and high frequency conditions in Experiment 1. Particularly, the participants produced default fixations for the control condition and made predictive saccades based on high frequency cues. As for the low frequency trials, they made predictions based on both informative cues and salient distracting cues (no significant difference between fixations towards the target label for the low frequency cues and the over-generalization label). These observations confirmed that learners make predictions in word learning as hypothesized by the DL theory and provided further support to the input-driven, domain-general account of constructivists.

The learning performance was also measured in offline tests which allowed direct comparison to the offline results in previous studies. However, the results did not align with Ramscar et al., (2010) as there was less correct generalization with low frequency items and participants only achieved an above-chance accuracy for pairing low frequency cues in the second session of Experiment 2. An attempt was made to link this divergence with the evidence of individual differences in low frequency trials.

By separating the participants into two learners' groups based on their performances in the generalization test in Experiment 1, the eye-movements data were analyzed separately across three stages. Although no statistical analysis was conducted due to small sample size, the plots showed distinctive predictive behaviors between the two participants groups: those who gained above-chance accuracy generalizing low frequency cues in the offline tests waited till later in the training to make predictions and could eventually unlearn the uninformative cues. The over-generalizers who did not gain above-chance accuracy started overgeneralization at the start of the training session and failed to retreat from this overgeneralization throughout the whole training process. These difference in prediction making indicated that some people could be too attracted by the salient cues like colors to overcome and deduct their predictive strength during cue competition – the implicit DL mechanism could be blocked for learners who are too focused on the task of mapping label outcomes with certain cues. Although speculative, this finding can potentially provide a different account to the poorer learning performance of L2 speakers when compared to their learning in L1: some L2 learners could have difficulty learning from the incorrect predictions they made similarly too as the over-generalizers in Experiment 1.

Note that this study has several limitations regarding the design of alternative learning paradigm, the sample size of participants, the artificial language setting and the eye-tracking using web camera. However, these problems also provided valuable directions for future research. For example, future studies investigating on-line DL with complex stimuli should make sure there is enough intermission between cue offset and label onset so that learners can fully process the available information. It will also be interesting to move from artificial languages and explore if predictive behaviors and individual differences of performances are present in real L2 learning. Further exploration of the different learning across individual learners, and why some could not learn from prediction errors, might ultimately lead to greater understanding of discriminative language learning, and could also possibly extend to the domain of cognitive processing and language disorders.

Overall, it is a positive direction for the field of developmental science that theories are being tested in a precise way to provide clearer, more detailed and falsifiable account for the process of language acquisition. By reflecting on existing findings and extending previous designs with new experimental methods, this study provided on-line evidence for prediction making during the discriminative process of word learning. However, contrary to reaching a conclusion, this study is only a first step probing into the real-time behaviors during language learning. The observed findings also served to inspire new predictions about “what” and “how” language learning is or could be, which await future studies to testify.

7. References

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N. and Evershed, J. (2018). *Gorilla in our Midst: An online behavioral experiment builder* bioRxiv, 438242 doi: https://doi.org/10.1101/438242

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

Barr, D. J. (2008). Analyzing 'visual world'eyetracking data using multilevel logistic regression. *Journal of memory and language, 59*(4), 457-474.

Bates, E. (1979). *The emergence of symbols: Cognition and communication in infancy (Language, thought and culture)*. New York; London: Academic Press.

Bley-Vroman, R. (1989). "What is the logical problem of foreign language learning?" in *Linguistic perspectives on second language acquisition*. eds. Gass S. and Schachter J. (Cambridge: Cambridge University Press), 41–68.

Brill, E., Magerman, D., Marcus, M., & Santorini, B. (1990, October). Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the 5th Jerusalem Conference on Information Technology, 1990.'Next Decade in Information Technology'* (pp. 380-389). IEEE.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 11-53). New York: Wiley.

Canfield, R. L., & Haith, M. M. (1991). Young infants' visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology*, *27*(2), 198.

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

Chomsky, N. (1959). A review of BF Skinner's Verbal behavior. *Language, 35* (1), 26–58.

Chomsky, N. (1980) On Cognitive Structures and their Development: A reply to Piaget. In *M. Piattelli-Palmarini, ed. Language and Learning: The debate between Jean Piaget and Noam Chomsky.* Harvard University Press.

Chomsky, N. (1995). Language and nature. *Mind, 104*(413), 1-61.

Chipere, N. (2001). Variations in native speaker competence: Implications for first-language teaching. *Language Awareness*, *10*(2-3), 107-124.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157-205.

Cipriani, E. (2015). The generative grammar between philosophy and science. *European Journal of Literature and Linguistics*, (4), 12-16.

Crain, S., Thornton, R., & Murasugi, K. (2009). Capturing the evasive passive. *Language Acquisition, 16*(2), 123-133.

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.

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 892.

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in second language acquisition*, *22*(4), 499-533.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117-1121.

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917-930.

Duffy, S., Huttenlocher, J., & Crawford, L. E. (2006). Children use categories to maximize accuracy in estimation. *Developmental Science, 9*(6), 597-603.

Dąbrowska, 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, *58*(4), 931-951.

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it?. *Frontiers in psychology, 6*, 852.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. *The handbook of second language acquisition*, *14*, 63.

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*, *48*(1), 71-99.

Fenson, L. (2007). *MacArthur-Bates communicative development inventories*. Baltimore, MD: Paul H. Brookes Publishing Company.

Frey, H. P., Honey, C., & König, P. (2008). What's color got to do with it? The influence of color on visual attention in different categories. *Journal of Vision, 8*(14), 6-6.

Friederici, A. D., & Wessels, J. M. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & psychophysics, 54*(3), 287-295.

Gallistel, C. R. (2002). Frequency, contingency and the information processing theory of conditioning. *Frequency processing and cognition*, 153-171.

Gold, E. M. (1967). Language identification in the limit. *Information and control, 10*(5), 447-474.

Gómez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in cognitive sciences, 4*(5), 178-186.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-436

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological science*, *18*(3), 254-260.

Grüter, T., Rohde, H., & Schafer, A. (2014). The role of discourse-level expectations in non-native speakers' referential choices. In *Proceedings of the annual Boston university conference on Language Development*.

Grüter, T., Zhu, Y. A., & Jackson, C. N. (2021). Forcing prediction increases priming and adaptation in second language production. *Prediction in Second Language Processing and Learning, 12,* 207.

Hallett, P. (1986). Eye movements((and human visual perception)). *Handbook of perception and human performance., 1*, 10-1.

Han, Z. (2013). Forty years later: Updating the fossilization hypothesis. *Language Teaching*, *46*(2), 133-171.

Hare, M., McRae, K., & Elman, J. (2004). Admitting that admitting verb sense into corpus analyses makes sense. *Language and Cognitive Processes*, *19*(2), 181-224.

Hartshorne, J. K., Tenenbaum, J. B., and Pinker, S. (2018). A critical period for second language acquisition: evidence from 2/3 million English speakers. *Cognition, 177*, 263–277. doi: 10.1016/j.cognition.2018.04.007

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.

Indefrey, P., & Goebel, R. (1993). The learning of weak noun declension in German-children vs artificial network models. In *5th Annual conference of the of the Cognitive Science Society* (pp. 575-580). Erlbaum.

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(3), 235-247.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive psychology*, *21*(1), 60-99.

Johnson, E. K., & Seidl, A. H. (2009). At 11 months, prosody still outranks statistics. Developmental science, 12(1), 131-141.

Johnson, S. P., Amso, D., & Slemmer, J. A. (2003). Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences*, *100*(18), 10568-10573.

Juliano, C., & Tanenhaus, M. K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the 15th annual conference of the Cognitive Science Society* (pp. 593-598). Erlbaum Hillsdale, NJ.

Jusczyk, P. W., Friederici, A. D., Wessels, J. M., Svenkerud, V. Y., & Jusczyk, A. M. (1993). Infants′ sensitivity to the sound patterns of native language words. *Journal of memory and language*, *32*(3), 402-420.

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.

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & psychophysics*, *61*(8), 1465-1476.

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.

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

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of child language*, *12*(2), 271-295.

Madsen, J., Julio, S. U., Gucik, P. J., Steinberg, R., & Parra, L. C. (2021). Synchronized eye movements predict test scores in online video education. *Proceedings of the National Academy of Sciences, 118*(5).

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53-85.

Martí, A. M., Taulé, M., Kovatchev, V. & Salamó, M. (2019). DISCOver: DIStributional approach based on syntactic dependencies for discovering COnstructions. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2018-0028

Martin, C. D., Thierry, G., Kuipers, J. R., Boutonnet, B., Foucart, A., & Costa, A. (2013). Bilinguals reading in their second language do not predict upcoming words as native readers do. *Journal of Memory and Language*, *69*(4), 574-588.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive science*, *26*(4), 393-424.

Mitsugi, S., & Macwhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism: Language and Cognition*, *19*(1), 19-35.

Newmeyer, Frederick J. (2008). Universals in syntax. *Linguistic Review, 25*(1-2), 35-82.

Nixon, J. S. (2020). Of mice and men: Speech sound acquisition as discriminative learning from prediction error, not just statistical tracking. *Cognition*, *197*, 104081.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, p.3839--3845.

Pelucchi, B., Hay, J. F., & Saffran, J. R. (2009). Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, *113*(2), 244-247.

Peters, A. M. (1977). Language learning strategies: does the whole equal the sum of the parts?. *Language*, 560-573.

Piaget, J. (1999). *Play, dreams and imitation in childhood* (Developmental psychology; 25). London: Routledge.

Popper, K. R. (1959). *The Logic of Scientific Discovery*. University Press.

Quine, W. (1960). *Word and object* (Studies in communication). Cambridge, Mass.: Massachusetts Institute of Technology.

R Core Team (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org/.

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., & 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.

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

Reynolds, G. S. (1961). Behavioral contrast. *Journal of the Experimental Analysis of Behavior*, *4*(1), 57.

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.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 2: psychological and biological models* (pp. 216-271).

Hyams, N. (1987). The theory of parameters and syntactic development. In *Parameter setting* (pp. 1-22). Springer, Dordrecht.

Rodd, J. M. (2020). Settling into semantic space: An ambiguity-focused account of word-meaning access. *Perspectives on Psychological Science, 15*(2), 411-427.

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. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition*, *81*(2), 149-169.

Saffran, J. R. (2009). What is statistical learning, and what statistical learning is not. *Neuroconstructivism*, 180-95.

Sampson, G. (2005). *The 'Language Instinct' Debate: Revised Edition*. A&C Black.

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods, 50*(2), 451-465.

Skinner, B. F. (1938). *The behavior of organisms: an experimental analysis.* Appleton-Century.

Skinner, B. F. (1957). *Verbal behavior.* New York: Appleton-Century Crofts.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive psychology*, *49*(3), 238-299.

Smith, N. (1999). Psychological reality. In *Chomsky: Ideas and Ideals* (pp. 93-135). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139163897.004

Street, J. A., & Dąbrowska, E. (2010). More individual differences in language attainment: How much do adult native speakers of English know about passives and quantifiers?. *Lingua*, *120*(8), 2080-2094.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive psychology*, *50*(1), 86-132.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive psychology*, *79*, 1-39.

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language learning and development*, *3*(1), 1-42.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*(4), i.

Treffers-Daller, J., & Calude, A. (2015). The role of statistical learning in the acquisition of motion event construal in a second language. *International Journal of Bilingual Education and Bilingualism*, *18*(5), 602-623.

Treffers-Daller, J., & Ziyan, X. (2016). 7. Can Classroom Learners Use Statistical Learning? A New Perspective on Motion Event Construal in a Second Language. In *Crosslinguistic influence in second language acquisition* (pp. 121-146). Multilingual Matters.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental psychology: Learning, memory, and Cognition*, *19*(3), 528.

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

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning. *Journal of Memory and Language*, *119*, 104231.

Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature*, *412*(6842), 43-48.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147-162.

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., 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., 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.

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*(1), 1-14.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015.

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

Zettersten, M., Wojcik, E., Benitez, V. L., & Saffran, J. (2018). The company objects keep: Linking referents together during cross-situational word learning. *Journal of memory and language*, *99*, 62-73.

8. Appendices

Appendix 1. CUREC approval.

SOCIAL SCIENCES & HUMANITIES INTERDIVISIONAL RESEARCH ETHICS COMMITTEE

Research Services, University of Oxford, Wellington Square, Oxford OX1 2JD
Tel: +44(0)1865 616576 Fax: +44(0)1865 280467
ethics@socsci.ox.ac.uk

6 January 2021

Dr Elizabeth Wonnacott
Department of Education

Dear Elizabeth

Research Ethics Approval (CUREC 2)
Ref No: R73484/RE001
Title: Language learning in adults and children

The above application has been considered on behalf of the Social Sciences and Humanities Interdivisional Research Ethics Committee (IDREC) in accordance with the procedures laid down by the University for ethical approval of all research involving human participants.

I am pleased to inform you that, on the basis of the information provided to the IDREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly approval has been granted.

Should there be any subsequent changes to the project that raise ethical issues not covered in the original application you should submit details to the IDREC for consideration:

https://researchsupport.admin.ox.ac.uk/governance/ethics/apply/sshidrec#collapse394916

Please note that your study may be selected for review by the SSH IDREC during an annual audit. You may also be required to submit a brief annual progress report on each anniversary of study approval, until the study is completed.

Yours sincerely,



Jennifer Blaikie
Research Ethics Manager

cc: Jackie Bridges

Appendix 2. *Experiment 1:* Demonstration of experimental interface – Information and consent sheet.

Department of Education, University of Oxford, 15 Norham Gardens, OX2 6PY

INFORMATION SHEET

Alien Learning Game

The Alien Learning Game has been designed by researchers in the Department of Education at the University of Oxford to investigate how do we learn a new language. Please read the following information before you decide whether you wish to take part.

Who?

The Alien Learning Game have been designed by Dr Elizabeth Wonnacott and her collaborators in the Department of Education at the University of Oxford. The game is designed to be informative for researchers exploring language learning. The present research is funded by the Leverhulme Trust. The study has been reviewed and approved by the University of Oxford's Central University Ethics Committee for research involving human participants (Ref No: R73484/RE001)

Why?

We have designed this game to better understand language learning. We compare different learning methods to see how well you learn novel words. This will help researchers identify better ways to teach new languages in a variety of settings.

Who can take part?

We would like adults aged between 18 and 35 years to participate. In order to participate, it is important that:

- You are a native speaker of English, or you understand English like a native.
- You don't have any history of neurological and language impairment.
- You don't wear spectacles/goggles and aren't affected by strabismus.
- Your computer needs to have access to a webcam

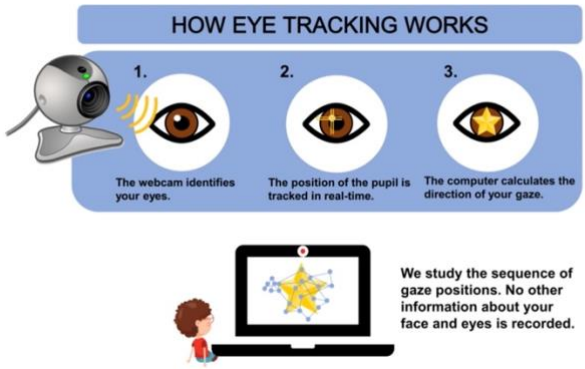
Please let us know if you have any concerns.

What will happen?

In this online game you'll be presented with images of novel objects, and be told their names. Your task is simply to pay attention to them for the duration of the experiment (around 30 min) and trying to remember their names. You'll be asked to recall the objects' name and physical appearance at some point. Sometimes you'll be asked to familiarize with the tasks and to read carefully their instructions.

We record eye movements during the game using the camera built into your laptop. We record where you are looking in terms of spatial co-ordinates (no photos or video are recorded). The eye-tracking procedure is non-invasive and will not cause any harm or discomfort.

We're going to infer where you're looking during the experiment through your webcam. Rest assured that the data we record are simply spatial coordinates of your eyes' gaze, and no information about your face and your eyes is recorded.



Will my taking part in this project be kept confidential?

Your answers will be completely anonymous. All the information that we collect about you during the course of the research are related ONLY to your performance in this experiment. We do not have access to your identity at any point. This information will be kept strictly confidential and you will not be able to be personally identified in any ensuing reports or publications. We make sure that this is the case by assigning you a novel random ID number that is not linked to any of your existing IDs present on Prolific or in Gorilla, moreover we do not ask your age or any other personal information that might connect you to your novel ID.

Potential risks?

There are no potential risks in taking part in this game. The eye-tracking procedure is non-invasive and will not cause any harm or discomfort. The knowledge we gain from the study will help us understand how we learn novel words in another language.

What happens to the data collected?

We record only your performance in this experiment in the form of button presses, mouse clicks and eye movements coordinates. The Gorilla and Prolific platforms are the data controller with respect to your personal data and, as such, will determine how your personal data is used. Please see their privacy notice here and here. Gorilla shares with the researchers involved in this study part of the University of Oxford only fully anonymised data where a random generated key has been assigned automatically at the beginning of the experiment. We do not have access to your personal data at any time, and therefore it's impossible to link your data to your person.

Data about your performance will be shared with Dr Elizabeth Wonnacott and Dr Eva Viviani (University of Oxford) as Principal Investigator and Research Fellow assigned to the project, and with Dr Michael Ramscar (University of Tubingen) as collaborator to this project. We would also like your permission to use your anonymised data in future studies, and to share data with other researchers (e.g. in online databases such as Online Science Framework) and publish it on scientific journals. There is no personal information that could identify you. Responsible members of the University of Oxford and funders may be given access to data for monitoring and/or audit of the study to ensure we are complying with guidelines, or as otherwise required by law

Do I have to take part?

It's up to you to decide whether or not to take part—your participation is entirely voluntary! Your refusal to agree won't result in any penalties or prejudice and you'll be still get paid. However, if you do agree in taking part to our experiment, you'll be asked to declare your decision digitally by completing a consent form (at the bottom of the page). Remember that you can always opt-out at any time without giving explanations, and if you do decide to withdraw from the study you'll be asked what you wish to happen to the data you have provided to that point.

Any questions?

If you would like to discuss the research with someone or have any questions please contact us at **elizabeth.wonnacott@education.ox.ac.uk.**

Additional Information

In the unlikely event you have a concern about any aspect of the project, please contact us via email: **elizabeth.wonnacott@education.ox.ac.uk.** We will do our best to acknowledge your concern within 10 working days and give you an indication of how we intend to deal with it. If you remain unhappy or wish to make a formal complaint, please contact the chair of the Central University Research Ethics Committee at the University of Oxford (using the contact details below) who will seek to resolve the matter as soon as possible. **Chair, Social Sciences and Humanities Inter-Divisional Research Ethics Committee; Email:ethics@socsci.ox.ac.uk; Address: Research Services, University of Oxford, Wellington Square, Oxford, OX1 2JD.**

If you agree to participate please check the box below to get started

☐ I confirm that I have read the information available in this information sheet and I agree to take part

Next

Appendix 3. *Experiment 1:* Coding of two learner groups for low frequency condition.

	match[a]	mismatch[b]	contScore[c]	labPicScore[d]	subjLearnerType[e]
1	-69.8	55.4	-125.2	0.166666667	over-generalizer
2	-87	70.46666667	-157.4666667	0	over-generalizer
3	-0.2	-0.066666667	-0.133333333	0.1	over-generalizer
4	-44.13333333	64.73333333	-108.8666667	0.133333333	over-generalizer
5	4.4	8.933333333	-4.533333333	0.210526316	over-generalizer
6	-47.4	72.8	-120.2	0.041666667	over-generalizer
7	9.466666667	9.6	-0.133333333	0.105263158	over-generalizer
8	-19.2	29.73333333	-48.93333333	0.086956522	over-generalizer
9	-8.2	26.06666667	-34.26666667	0.473684211	over-generalizer
10	-43.26666667	64.6	-107.8666667	0.15	over-generalizer
11	-25.06666667	-2.8	-22.26666667	0.277777778	over-generalizer
12	-8.2	45.66666667	-53.86666667	0.2	over-generalizer
13	-19.66666667	62.33333333	-82	0	over-generalizer
14	19.93333333	-43.53333333	63.46666667	0.352941176	ambig[f]
15	7.133333333	-4.533333333	11.66666667	NA	ambig
16	-41.53333333	-34.53333333	-7	0.590909091	ambig
17	10.73333333	-28	38.73333333	0.5	ambig
18	4	0	4	0.1	ambig
19	6.066666667	-20.4	26.46666667	0.391304348	ambig
20	86.66666667	-99.93333333	186.6	0.842105263	learnerLowF[g]
21	56.73333333	-83.06666667	139.8	0.941176471	learnerLowF
22	99	-99.2	198.2	1	learnerLowF
23	31	-40.13333333	71.13333333	0.608695652	learnerLowF
24	87.06666667	-91.06666667	178.1333333	0.863636364	learnerLowF
25	99.6	-99.73333333	199.3333333	1	learnerLowF
26	99.73333333	-100	199.7333333	1	learnerLowF
27	49	-53.2	102.2	0.956521739	learnerLowF
28	23	-55.2	78.2	0.722222222	learnerLowF
29	29.2	-30.86666667	60.06666667	0.65	learnerLowF

a. Ratings in the contingency judgement test for trials of matched fribble and label pairs – should receive positive ratings.

b. Same as **a** but for mismatched fribble and label pairs – should receive negative ratings.

c. Scores of the contingency judgement tests.

d. Scores of 4AFC generalization test.

e. Coding of learner groups.

f. Participants with inconsistent scores of the two tests.

g. Learners of the low frequency cues.

Appendix 4. *Experiment 2:* Demonstration of experimental interface – Information and consent sheet.

Department of Education, University of Oxford, 15 Norham Gardens, OX2 6PY

LANGUAGE LEARNING IN ADULTS AND CHILDREN

Research Ethics Manager: Jennifer Blaikie

Researchers: Ms Haoyu Zhou, Dr Eva Viviani (Researcher), Dr. Elizabeth Wonnacott (Principal Researcher)

Department of Education, University of Oxford, 15 Norham Gardens, OX2 6PY .
haoyu.zhou@jesus.ox.ac.uk, eva.viviani@education.ox.ac.uk,
elizabeth.wonnacott@education.ox.ac.uk Tel. +44 07768092180

This study has been approved by the Social Sciences and Humanities Interdivisional Research Ethics Committee (IDREC) at University of Oxford: Project Ref No: R73484/RE001

What is our study about?

In this study, participants play a language learning game and learn an artificial language which has been designed by the experiments to mimic the properties of real language.

We have designed this game to better understand language learning. This will help researchers identify better ways to teach new languages in a variety of settings.

Why have I been chosen?

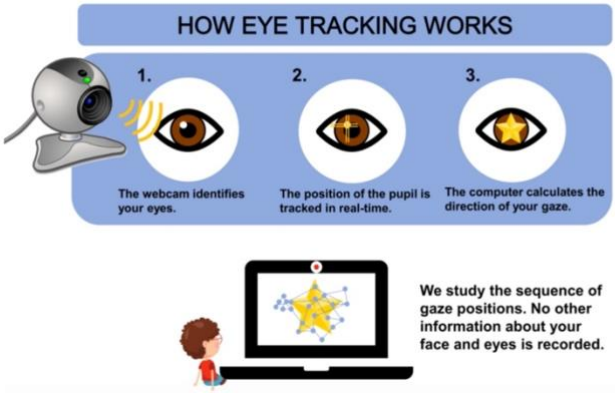
We are seeking to recruit native-like English speakers aged between 18 and 45 years to participate.

You should be having normal or corrected to normal vision and have no history of neurological and language impairment. It is also important that you are not affected by strabismus and **do not wear spectacles/goggles** during the experiment.

Your computer also needs to have access to a webcam for the eye tracking aspect of the study.

Eyetracking and data protection

This language learning game involves recording eye movements during the game using the camera built into your laptop, or an external web camera. We record where you are looking in terms of spatial coordinates **(no photos or video are recorded)**. The eye tracking procedure is **non-invasive** and will **not cause any harm or discomfort**.



What will I be doing ?

In this online game, you'll be presented with images of novel objects, and be told their names. Your task is simply to pay attention to them for the duration of the experiment (**around [30/45] min**) and trying to remember their names. You'll be asked to recall the objects' name and physical appearance at some point. Sometimes you'll be asked to familiarize with the tasks and to read carefully their instructions. [We're going to infer where you're looking during the experiment through your webcam. Rest assured that the data we record are simply **spatial coordinates** of your eyes' gaze, and **no information about your face and your eyes is recorded**.]

We will remind you of each task as you get to it and explain exactly what to do. We will also ask you to answer a few more questions giving details about the languages that you speak and any training you have had. This is because these factors have all been shown to have an effect on a person's ability to learn a language. If you feel unable to answer any of the questions for any reason, then please leave them blank.

Potential risks ?

There are **no** foreseeable discomforts, disadvantages and risks associated to this experiment. **(The eye-tracking procedure is non-invasive and will not cause any harm or discomfort.)** However, if you feel or believe there are any unexpected discomforts, disadvantage and risk which arises during the research, you should immediately contact us via email.

What happens to the data collected?

We record only your performance in this experiment in the form of button presses and mouse clicks. The Gorilla and Prolific platforms are the data controller with respect to your personal data and, as such, will determine how your personal data is used. Please see their privacy notice here and here. Gorilla shares with the researchers involved in this study part of the University of Oxford only fully anonymised data where a random generated key has been assigned automatically at the beginning of the experiment. We do not have access to your personal data at any time, and therefore it's impossible to link your data to your person.

Data about your performance will be shared with Dr Elizabeth Wonnacott and Dr Eva Viviani (University of Oxford) as Principal Investigator and Research Fellow assigned to the project, and with Haoyu Zhou (University of Oxford) as collaborator to this project. We would also like your permission to use your anonymized data in future studies, and to share data with other researchers (e.g. in online databases such as Online Science Framework) and publish it in scientific journals. There is no personal information that could identify you. Responsible members of the University of Oxford and funders may be given access to data for monitoring and/or audit of the study to ensure we are complying with guidelines, or as otherwise required by law.

Do I have to take part?

It's up to you to decide whether or not to take part—your participation is **entirely voluntary**!

Your refusal to agree won't result in any penalties or prejudice and you'll be still get paid. However, If you do agree in taking part in our experiment, you'll be asked to declare your decision digitally by **completing a consent form** (at the bottom). Remember that you can always opt-out at any time without giving explanations, and if you do decide to withdraw from the study you'll be asked what you wish to happen to the data you have provided to that point.

Questions and additional information?

If you have any queries or concern about any aspect of this study please contact: **elizabeth.wonnacott@education.ox.ac.uk**

We will do our best to answer your query within 10 working days to give you an indication of how it will be dealt with.

If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible: **Chair, Social Sciences & Humanities Interdivisional Research Ethics Committee; Email: ethics@socsci.ox.ac.uk; Address: Research Services, University of Oxford, Wellington Square, Oxford OX1 2JD**

If you agree to participate please check the box below to get started :)

☐ ***I confirm that I have read the information available in this information sheet and I agree to take part***

A few more questions on settings to ensure data quality:

In what kind of lighting will you be doing the experiment?

☐ Artificial lighting
☐ Natural/day lighting (preferred set-up)

Where will your laptop/computer be placed during the experiment?

☐ Lap
☐ Desk/Table (preferred set-up)

Will you be wearing prescription glasses during this experiment?

☐ Yes
☐ No

Will you be wearing contact lenses during this experiment?

☐ Yes
☐ No

Next