Cross-situational word learning:

Interaction with linguistic domains and underlying mechanisms

Patricia Besonia Amillos

Thesis submitted for the award of Doctor of Philosophy

School of Education, Communication, and Language Sciences

May 2023

Abstract

Infants encounter words while having multiple objects in their view. The breadth of the learning space is the concern which several word learning theories aim to address. One proposed mechanism is cross situational word learning (CSWL) that states that word referent mapping takes place across multiple encounters.

There are several advantages to CSWL as a method for supporting early word learning. It requires no innate constraints, presents a continuity with phonetic development using statistical regularities in the input, and can account for the importance of a co-occurring cue that indicates reference: eye gaze. CSWL presents an opportunity to link three areas in language acquisition: word learning, phonetic development, and social cues.

There are different proposed approaches on how learners perform CSWL. The two accounts are the associative learning and the hypothesis testing, but the discussion is still ongoing as to which of the two better describes learning during CSWL experiments.

This thesis investigated if learners could simultaneously track phonetic information while learning words through CSWL in two internet-based studies with adults and children using a pre-test post-test design. It also examined how the tracking of co-occurrence frequencies during a CSWL task was influenced by the difficulty of the learning situation as influenced by trial spacing and social cues.

Study 1 found that adults showed sensitivity to the contrast after training but only in a referential-based task. The influence of task type on measuring discrimination abilities was further observed in Study 2 when adults were able to differentiate the contrast in an AX task but only when stimuli type and presentation are simplified. Children did not benefit from training despite being less susceptible to L1 influences. Study 3 found that social cues aided in locating the target faster although participants only began tracking a single referent later in the learning phase.

Acknowledgments

I would like to express my heartfelt gratitude to the following individuals who have given me the opportunity to undertake this research and have provided their unwavering support throughout.

First of all, I would like to thank my supervisors Dr. Barbara Höhle and Dr. Nick Riches for their guidance, knowledge, and expertise on this project. This also includes the patience and understanding they have extended to help me complete my PhD in the midst of a global pandemic. I feel grateful that Prof. Höhle has always provided a way forward despite the many uncertainties brought about by COVID and her advice has been a source of reassurance that I will finish. I also deeply appreciate Dr. Riches for always making time to discuss my ideas in person or on paper and also dedicating the time to support my stay in the UK. He has written countless queries, emails, and letters of support which has provided me a stable environment to keep working on my project. This project and all the opportunities that have come from it would not have been possible without you both.

I owe this opportunity of pursuing a PhD to the IDEALAB program and the people behind it – the directors, coordinators, and my fellow PhD students. I would like to sincerely thank Dr. Roelien Bastiaanse and Dr. Lyndsey Nickels for every word of encouragement that they have given me during my panels. It means a lot to know that my work reflects how grateful I am for being accepted into this program. I also owe considerable thanks to Dr. Christos Salis who has continuously advocated for the support I need to stay in Newcastle and to finish my project. My thanks also go out to Anja Papke, Alice Pomstra, and Dr. Leonie Lampe for every carefully written document that they have provided me. I am extremely fortunate to have received all of this support from IDEALAB to help me fulfill this goal.

This project was possible thanks to the help of the BabyLab of the University of Potsdam and the LingLab of Newcastle University for their support in stimuli creation, participant recruitment, and data collection. I would also like to thank Tom Fritzsche for his expertise in the design and analysis of eyetracking experiments. Thanks is also due to Prof. Kenny Smith for sharing his repository of images to be used in Study 3. I am also grateful for all the parents, children, and adults who have participated in these experiments. They have given my experiments their time and full attention despite the unusual words and tasks that came their way.

I would like to thank my colleagues at HESA and Jisc who have provided me opportunities to apply the skills I have learned during the PhD in the higher education sector. Thank you for being supportive as I balanced work with writing up my dissertation. I also sincerely thank my line manager Priyanka Petinou for being a wonderful mentor. I am humbled by the amount of faith she places in my abilities and her belief that I can one day help others as much as she has helped me.

I would also like to thank my friends scattered across the world. Thanks to my friends in Potsdam for being a piece of home, especially to Rowena who I have to thank for introducing me to such a brilliant group of people in the first place. I also would like to thank Simone and Daiani who have been there for every adventure since the EMCL and now as I followed them to the UK. I also feel grateful to Bea, Cesca, John, Kerwyn, Nadine, Nyx, and Nico for a decade of friendship that persists despite the 6500 miles between us. I would also like to express my thanks to Rachel for her daily encouragement that I can and will write this thesis. It really helped to hear that someone who writes for a living has placed that kind of belief in my own writing. A big thanks to Ellyn who has made space for every thought surrounding research, life, and life of a PhD student attempting to do research whether it is five sentences or five pages long.

I am also grateful for my friends Bruna, Pauline, and Mareike who I had the pleasure of sharing this experience with. I will always appreciate how Bruna makes the experience of doing a PhD less lonely with her empathy and kindness. Thank you to Pauline who is so thoughtful, caring, and positive and always convinces me that things will work out for the better. I have been grateful for every day of my friendship with Mareike. She has comforted me with every setback, helped me face every challenge, and cheered me on with every triumph. I could not have predicted that this PhD would give me three friends that I would treasure this dearly.

Lastly, I would like to thank my family who has been my biggest source of strength and inspiration throughout. I would like to thank Justin who never allowed my doubts to overshadow my desire to try. Thank you to my father, Pio, who has shown me how to strive for excellence and persevere in my work. Thank you to my mother, Susan, who is unwavering in her belief and support that I am capable of achieving my dreams. Thank you to my sister, Suzanne, who helped me realize my sense of purpose at a pivotal time in my life. It is with the greatest satisfaction that I dedicate this piece of work to you Dad, Mom, and Suzanne.

Table of Contents

List of Figures and Tables

Chapter 1. Introduction

This chapter begins with a description of the early development of word learning and the referential ambiguity problem. In most word learning contexts, children are faced with multiple possible referents when first encountering a novel word. How children overcome the overwhelming possibilities is one major focus of word learning literature (Golinkoff & Hirsch-Pasek, 2000). This chapter then introduces the prominent theories in the field, beginning with the proposal that innate constraints that guide word learning, followed by the counterargument that emphasizes the social nature of early linguistic interactions, and then that which is the focus of this dissertation: the statistical-based proposal of cross-situational word learning (CSWL). A description of CSWL, both as a mechanism of learning and an experimental paradigm are provided in section 1.2, as well as a historical overview of trends in CSWL research in the following section. In summary, the current research on CSWL has moved towards studying statistical-based word learning as it integrates with other linguistic domains, extending findings to L2 learning and other populations, and determining the underlying mechanism used to organize statistics that support word learning.

The studies in this dissertation focus on L2 phonetic learning in CSWL, the role of supportive social cues in CSWL, and the two major mechanisms proposed on how learners aggregate information during CSWL. Relevant literature and models that support these individual topics, both in language acquisition and L2 learning, are discussed in their own sections. This dissertation originally had a strong developmental focus in its inception, but the COVID pandemic and its resulting restrictions on recruitment made testing infants extremely challenging. All of the studies thus recruited adults and study 2 had a comparison of adults and 4-year-old children. Aside from changing the population being studied, the pandemic also required adapting the method to internet-based testing due to restrictions on in-person testing. This chapter discusses in section 1.7 the opportunity to develop methods for internet-based testing as a potential way to address the replication crisis and the overrepresentation of the WEIRD population in psychology and linguistic research. Finally, this chapter ends with an overview of the studies in the succeeding chapters.

The early linguistic development of infants is often measured in words. Infants are met with high praise when they begin to respond to their name. Years later parents will still remember their child's first word. On the other hand, a child who has not produced any words past their first birthday are monitored for developmental delays. These situations emphasize how strongly words are associated with an emerging mastery of language. The growth of an infant’s vocabulary is a tangible way of observing that they are on their way to becoming a competent communicator.

It might be easy to underestimate the process of acquiring words as basic given that it occurs at the beginning of linguistic development. Word learning seemingly only involves infants learning that a specific linguistic form corresponds to a specific object, supported by the finding that most early words are referential (for a review see Waxman et al., 2013). Before infants can even get to this point, they first need to develop segmentation skills which allow them to locate word forms in the continuous speech stream. They must also successfully filter out variations in the phonetic signal that do not create changes in meaning and differentiate forms with overlapping acoustic and articulatory features. Successfully identifying linguistically relevant regularities in the speech signal is fundamental to infants linking meanings to words.

Discovering the meaning of words is also not a straightforward process if one considers that the learning environment is rife with ambiguity. Quine (1960) illustrates this using a hypothetical situation where a linguist hears the expression “gavagai” from a speaker of an unknown language. The speaker points at a rabbit, but do they mean the entire animal, one of its body parts, or the fact that it is hopping away? This is something that cannot be immediately determined with the information available in this context. Literature on early word learning often uses this example as a parallel to the experience of infants attempting to discover the meaning of a word for the first time. When an adult looks at a round object while uttering “ball”, are they naming the object, its shape, its color, or a category of objects that bounce? Furthermore, one needs to recognize that the act of gazing at an object while producing a label communicates a referential relationship. During word learning, infants need to exercise a combination of phonetic, lexical, and social skills while contending with the breadth of the learning space.

The complexity of word learning becomes evident when all these factors are considered and aligns with its slow and effortful beginnings. This seems to change at around 18 months when the “vocabulary spurt” occurs, which is characterized by a rapid increase in the rate at which children acquire new words (Bloom, 1973; Goldfield & Reznick, 1990). The focus of this research is investigating how individuals combine phonetic, lexical, and social information specifically in ambiguous word learning situations in order to become efficient word learners.

1.1 Theories of Word Learning

One proposal is that infants possess innate internal biases that narrow the word learning space. These innate constraints or principles limit the range of permissible options when linking words to referents. A prominent principle in the existing literature is mutual exclusivity (Markman, 1989). According to this principle, learners assume that objects can only have one name. This influences word learning through guiding learners to assign a novel label to an object for which the label is unknown over another with a known label. Evidence of mutual exclusivity is found in infants at around 17 months who did not assign a second label to a familiar object (Markman et al., 2003; Halberda, 2003). This preference also persists in 4-year-olds (Jaswal & Hansen, 2006; Au & Glusman, 1990). Aside from mutual exclusivity, there also appears to be a bias towards assigning novel labels as a name for a whole object rather than its parts (Landau et al, 1988; Soja, 1991). There is also a finding that novel words tend to be interpreted as nouns rather than verbs or adjectives by 2 and 4-year-old children (Hall et al., 1993). There are several constraints proposed to support early language development which are beyond the scope of this current chapter but are explained in detail in Woodward (2000).

However, is constraining the learning space to the point where ambiguity is resolved possible? For example, infants may encounter multiple objects all of which they have not yet learned the label for in which case mutual exclusivity would be unhelpful. It also becomes a problem when infants need to learn the name of a part of an object in which the whole-object bias would lead them to the wrong conclusion. It seems impractical to have a constraint to address every potential ambiguous word learning situation since that would require knowledge of all the types of ambiguity one can encounter. It is more likely that innate constraints can serve as a guide on how to navigate the learning space rather than the main mechanism that initiates word learning. Furthermore, fully relying on innate constraints reduces the infant to a passive role. It does not acknowledge the need to actively analyze characteristics of the input to address the ambiguity being encountered.

Additionally, it appears that use of mutual exclusivity only becomes reliable in infants with larger vocabulary size (Kalashnikova et al., 2016). If infants consistent use of mutual exclusivity only develops alongside growth in receptive vocabulary, then this implies that mutual exclusivity is not a necessary precursor to building the lexicon. This is also supported by the finding that 14-month-old infants were unable to exploit the use of mutual exclusivity (Halberda, 2003). Bilingual infants at 17-22 months were also found unable to use mutual exclusivity when compared to monolingual infants of similar ages and vocabulary sizes (Houston-Price et al., 2010). Bilingual infants whose early linguistic experience is that one object can be referred to by multiple labels would create a different bias as compared to monolingual infants whose experiences align more with the predictions of mutual exclusivity. Mutual exclusivity does play an important role in word learning when infants have built a sufficiently large vocabulary to maximize this strategy (Markman et al., 2003), however it does not appear to be prerequisite to early lexical development nor does it align with the early linguistic experiences of bilingual infants.

Lastly, as Nelson (1988) suggests in her work evaluating the constraints approach, Quine's hypothetical situation might not actually reflect what the infants experience in word learning. It assumes that infants intend to guess objects that adults refer to with words. Although it is very difficult to ascertain the intent of infants at this stage, research suggests that caregivers are the ones who make efforts to adjust the content of their language to match the current focus of an infants' attention (Tamis-LeMonda et al., 2013). In a wide scale study of 190 mother-infant pairs from diverse cultural backgrounds, mothers were more likely to use referential language in response to their 14-month-olds' object exploration. Utterances like this which are both semantically and temporally contingent to the focus of the child's attention are found to predict vocabulary development at 18 months (McGillion et al., 2013). Given this, it is more likely that caregivers provide supportive labeling during optimal opportunities when infants are attending to an object that emphasizes word-object links. This then lends a supportive naturalistic context for CSWL to take place that arises from the social understanding of reference between caregiver and child.

A different approach places the social and communicative aspects of these caregiver-child interactions as the foundation of word learning (Akhtar & Tomasello, 1998; Akhtar & Tomasello, 2000). It does acknowledge that innate constraints, associative learning, attention, and memory all contribute to successful word learning, but all act in support of social pragmatic processes. Tomasello (2000) strongly argues against the constraints theory as not only does it ignore which events children might consider salient, but also some of the constraints could also create hindrances to their learning of other early word types. For example, Tomasello states that the whole object constraint would make it difficult for children to learn verbs, prepositions, or adjectives. The principle of extendibility, which assumes that one word can be extended to other perceptually similar objects, would create problems for proper nouns which only specifically refer to one object.

Tomasello continues to argue that a comprehensive word learning theory based on constraints would have require knowledge in advance of every type of referent an infant can encounter, then specify a corresponding constraint that would aid its acquisition. He instead proposes a more general mechanism that could support word learning through the infant's ability to read the cues from the speaker such as eye gaze and pointing which are highly referential. Infants' word learning is not just socially scaffolded, but also socially motivated.

The social-pragmatic account argues that word learning is not a mere result of temporally contiguous occurrences of labels and objects, but of the infant's understanding of adults' intention to communicate and the desire to engage in meaningful communication as well. This argument does not necessarily contradict CSWL, as temporally contiguous correspondences between labels and objects are the object of focus of shared interactions between infant and caregiver. If anything, it further emphasizes the relevance of CSWL experiments including social cues.

Baldwin et al. (1996) observed strong intention reading skills in 18 to 20-month-old infants. Infants were able to link a novel label to an object they were exploring when the speaker producing the label was visible in front of them but not when they were outside of their view. Tomasello & Akhtar (1995) also observed that 2-year-olds could use referential intent to link a label to an object even when both were not simultaneously available in the learning context. This ability to capitalize on referential intent for language learning develops when infants become able to coordinate their attention between a speaker and a third entity, an object or event, relevant to the current interaction. This is known as joint attention and allows infants to follow the subject of a speaker's gaze or gestures.

A link between infants' joint attention abilities and word learning has been observed in several studies (Tomasello & Todd, 1983; Tomasello & Farrar, 1986; Mundy & Newell, 2007). Joint attention was measured through coding interactions between parents and infants where linguistic input contained references to objects or events that were the focus of shared attention. These were then correlated with measures of vocabulary from 12 to 24 months. With joint attention, infants can direct their attention to the object or event at the center of the interaction which would also be the likely focus of linguistic input.

However, like the constraints approach many of the studies that support the use of social cues and referential intent have been conducted in toddlers. There are also findings that infants younger than two years cannot reliably use eye gaze to link object and referent when conflicting cues are present. Hollich et al. (2000) presented 12-month-olds, 19-month-olds, and 24-month-olds with a novel word and two possible referents: an interesting toy and a boring toy. The interesting toy was more perceptually salient and confirmed to be so when infants presented longer looking times to it compared to the other toy. Experimenters then introduced conflict by making the boring toy the focus of a speaker's gaze while labeling. The results showed that the two younger groups chose to look at the perceptually salient toy when hearing the label over the boring one even if it was the focus of the speaker's gaze. A second experiment was performed where objects of equal salience were presented, but infants at 12-13 months still did not successfully link the label with to the object the speaker was looking at. It was only when more explicit cues such as the speaker touching and handling the object while labeling did the 12-month-olds show evidence of learning.

Recent studies have also found that successful word learning can still proceed both in typical and atypical development without joint attention (Scofield & Behrend, 2011; Akhtar & Gernsbacher, 2007). The dyadic interactions where joint attention is observed also appear to be non-existent in other non-Western societies, yet children from these cultures are not assumed to be generally delayed in word learning development (Akhtar, 2005). There is still substantial evidence that supports the links between social skills and word learning (Carpenter et al., 1998; Baldwin & Moses, 2001), but the claim that they form the foundation that gets word learning off the ground is challenged by these findings.

A third account proposes that word learning is supported by a mechanism that is simpler than both of what the previous frameworks have suggested. Associative learning occurs when two perceptual cues are strongly associated with each other where the presence of the first one will increase attention to the other (Rescorla & Wagner, 1972). This type of learning begins as a slow and effortful trial-and-error process where infants link words and labels that occur contingently in learning situations. The associative learning account argues that the speed of word learning accelerates when learners observe the perceptual similarities of objects that are assigned a certain label (Smith, 2000). It is this pattern that objects with shared perceptual similarities are often heard with a certain label that scaffolds productive word learning. That is when children hear the word "car" and see its referent multiple times, they eventually learn to generalize that the label is for “car-shaped” objects instead of other properties such as color or material.

This shape bias is different from the biases stated in the constraints theory which are assumed to be innate (Markman, 1992). Rather, this shape bias is gradually built as a result of associative learning. Landau et al. (1988) tested this bias in groups of 2-year-olds, 3-year-olds, and adults. Participants were shown an object of a specific shape labeled with a pseudoword. They then proceeded to show the participants objects of varying size and material to the original but retained the shape of the original object and another that differed in shape but retained its other properties. All three groups of participants were more likely to agree that objects that were similar in shape to the original object could be labeled with the pseudoword, while they were more likely to disagree for the objects that differed in shape. This is presented by the authors as evidence that children and adults used shape to determine if unfamiliar objects can be referred to by a novel word.

The advantage of the associative learning approach is its simplicity as it does not presuppose any pre-existing knowledge from infants, but rather explains how knowledge emerges from observing patterns in the input. Consistent co-occurrences between objects’ perceptual properties and the labels they are given emerge from experience and support word learning. Associative learning has been observed as early as 3-6 months in other domains of development such as motor learning (Tripathi, et al., 2019) and space-pitch perception (Dolscheid, et al, 2014). However, this simplicity is also the main criticism against it. The proponents of the constraints-based approach might argue that the problem-space is too large for associative learning to even begin. Those who advocate for the social-pragmatic approach can argue that it reduces learning to pairs of temporally and spatially tied stimuli when there is strong evidence for a social component in infants’ early cognitive and linguistic development (Gerson et al., 2017; Carpenter et al., 1998).

Furthermore, the evidence for a shape bias being a primary cue for early word learning seems artificial because of the way this was tested in the lab. Experiments purposefully use objects of unfamiliar and arguably less functional shapes which do not serve a function in the context that they are introduced. Consequently, the only remarkable thing about these objects would be their perceptual properties and a frequently heard linguistic label. This is not how infants experience objects in naturalistic settings. The objects that are the center of child-caregiver actions have some function in shared activities, such as for play, bathing, feeding, that allow the child to experience much more of its properties than just its shape.

However, the simplification of the word learning process is an argument that can be used for any of the prior approaches as each of the proposed mechanisms for word learning are tested in isolation. This is a by-product of the control necessary for experimental conditions, but it strays further from the reality that multiple cues are present in the environment and that they interact with each other in learning. While it is important to identify the primary mechanism which allows word learning to proceed, it is also equally important to investigate how multiple cues interact with each to contribute to a successful word learner.

1.2 Cross-Situational Word Learning

The word learning theories discussed have attempted to address two main issues: 1) what is/are the mechanism(s) that allows word learning to proceed and 2) how do learners cope with ambiguity in word learning situations. So far, the theories discussed have implied that the second issue is directly addressed by the solution they offer to the first concern. The constraints-based approach assumes an innate set of assumptions about word-object relationships that eventually narrow down the learning space for children. The social-pragmatic approach proposes that the infant attunes to the social cues given by their communication partner that allows them to understand referential intent and thus identify the object which is the focus of attention and linguistic input. The associative learning approach, on the other hand, states that learners can detect patterns of objects and words occurring together and slowly learn the shape bias which allows them to infer in future learning situations that words often refer to objects of the same shape.

One thing in common among these theories is that they seek to resolve ambiguity in the immediate moment that the infant first encounters a novel word and a set of referents. There is less emphasis on the possibility that another opportunity to learn the correct word-referent mapping might arise in the future. If this is the case, then learners would not require an immediate answer to the ambiguity problem and can use future opportunities to gather more information on the correct mapping. Word learning research with children around 2 years of age shows that repeated exposures were necessary for them to learn and retain new words (Horst, 2013; Mather & Plunkett, 2009; Schwab & Lew-Williams; 2016).

Furthermore, infants at 6 months old showed longer looking times for infant-directed speech with enhanced repetition produced by the same speaker (McRoberts et al; 2009). This could be considered as infants becoming more sensitive to verbal repetition in the speech input. Shortly thereafter at 9 months, infants then exhibit the ability to segment continuous speech into words based on probabilistic properties of syllable structure found in the ambient language (Jusczyk et al., 1999; Thiessen & Saffran, 2003). Such close timing between the sensitivity to repetition and speech segmentation suggest the importance sensitivity to statistical regularities in the language input in beginning to discover word forms.

Cross-situational word learning (CSWL) is a proposed mechanism for how children overcome the two issues mentioned at the beginning of this section. CSWL is said to be is similar to the statistical mechanism used in segmenting continuous speech in that the learner detects statistical regularities between two streams of information: the linguistic input containing the word forms and the visual input from the observable events in the environment (Smith & Yu, 2008). While single learning instances might be initially uninformative as to what the correct mappings are, the referential ambiguity is resolved through continuing to compile co-occurrences between words and objects with multiple learning opportunities. By combining which object co-occurs most frequently with a word across learning situations, the learner thus arrives at a conclusion about the correct pairing. CSWL proposes that co-occurrence frequencies derived from multiple exposures to a word and its referent is how word learning begins.

In the lab, CSWL as an experimental paradigm is implemented by presenting multiple objects and multiple labels in a single trial to simulate referential ambiguity. The lack of correspondence between the position of objects on the screen and the order of mention make it impossible to decide based on a single trial alone which the correct referent is. Learners are provided multiple learning trials to provide them with a distribution of regularly co-occurring words and objects.

Figure 1.1 is an example of a trial in a CSWL experiment. A learner is shown multiple objects at once on a screen and will hear a corresponding number of labels (Trial 1). However, there is no relationship between the order of naming and the location of objects on the screen. In most cases, the assumption would be that the objects are named left to right. However, the location of objects and order of mention in Trial 2 would soon disprove that. The lack of correspondence between order of mention and location of objects persists throughout the training trials and is essential in maintaining within-trial ambiguity. To establish any consistent word-referent mapping, learners would be required to consolidate information across multiple trials in order to uncover the accurate co-occurrence frequencies. By compiling co-occurrence frequencies from Trials 1 to 3, the learner could conclude the following word-object relationships in Figure 1.2 based on co-occurrence frequencies. A key assumption of CSWL both as a mechanism and as an experimental paradigm is that co-occurrence frequencies would consistently be the highest for the correct word-referent mapping.

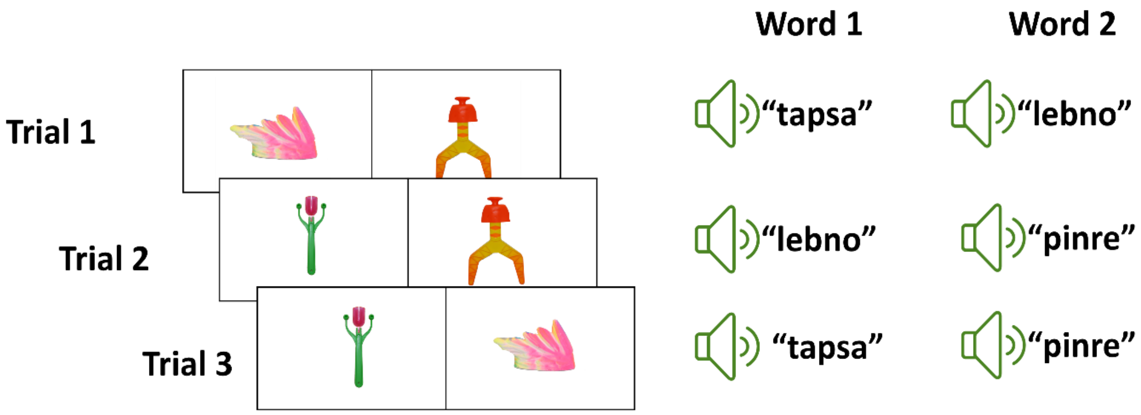


Figure 1.1

An example series of trials used in lab based CSWL experiments

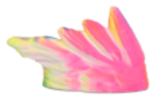
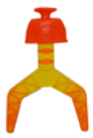
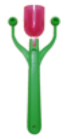
			
tapsa	**2**	**1**	**1**
lebno	**1**	**2**	**1**
pinre	**1**	**1**	**2**

Figure 1.2

Co-occurrence frequencies of objects and labels presented in Figure 1.1

1.3 Overview of Cross-Situational Word Learning Experiments

The mechanism of cross situational word learning was first suggested by Pinker (1989) as a way to limit the space of observed events when determining the meanings of verbs. He proposed that the child could encode a list of properties, such as the direction of the event, location of an object, causation, a set of properties of a theme or actor among others and add this to the tentative definition of the verb. Upon the next encounter, the child would discard any of these encoded features that contradict the current learning situation. The child must be able to use information across situations to identify elements of meaning that are consistent and reject any that are not in order to limit the learning space.

Gleitman (1990) builds upon this proposal by Pinker and suggests that a probabilistic rather than absolute approach is likely taken by the learner as the correspondence of linguistic stimuli with observable events in the world is bound to be imperfect. It would be too extreme if the child rejects a verb-meaning relationship on the basis of a smaller proportion of mismatches than correct correspondences in the co-occurrence of verbs with scenes. She also notes another potential concern is that this process quickly becomes more complicated as the set of properties encoded by the child continues to grow across multiple situations.

It was Siskind's (1996) computational work that provided the first proof of concept that cross-situational lexical acquisition is plausible for a learner. He constructed the task of cross-situational word learning as a formal mathematical problem where a learner takes the set of all possible meanings in each encounter with the word and finds the intersection among sets. The result would be the consistent meaning that appears across all learning situations. Computational simulations show that the algorithm could successfully map words to meanings in the face of referential ambiguity and noise in the input, two of the major concerns in early word learning.

Seminal work by Yu and Smith transformed this formal simulation into an experimental paradigm applied to nouns. This formed the basis for future lab based CSWL studies. The first experiment performed by Yu & Smith (2007) recruited adults to learn 18 novel word-referent pairs using only co-occurrence frequencies in a CSWL paradigm similar to what was described in the previous section of this chapter. The authors manipulated the referential ambiguity of the learning situation by varying the number of words and objects presented within a trial where participants either encountered 2, 3, or 4 words at a time. Participants performed above chance across all three conditions and remembered more word-object pairs with decreasing referential ambiguity.

In the same paper by Yu & Smith (2007), a second experiment was run where the total number of word-object pairs to be learned and the number of exposures for each pair was manipulated with referential ambiguity held constant. Participants performed better in the condition with more word pairs to be learned. The authors conclude that in a closed system of word-referent pairs, a larger data set may have been more beneficial as more evidence is available on which co-occurrences are systematic as opposed to spurious. Taken together, the results of these initial two experiments show that adults could learn word-referent mappings with only co-occurrence frequencies after only six minutes of training.

In Smith and Yu (2008), the CSWL paradigm was tested in 12 and 14-month-old infants. A simplified version of the design in their earlier work was used. Infants were presented 6 novel word-object pairs appearing two at a time within a trial. The direction of infants' eye gaze was coded as a measure of learning. Results show that both groups were looking longer at the target compared to the distractor for 4 out of the 6 words. The authors conclude that infants are also able to derive multiple word-referent mappings through co-occurrence frequencies alone.

Both experimental and computational investigations on CSWL have advanced equally since then. As the studies in this dissertation are behavioral, the following paragraphs give a larger emphasis on the experimental evidence available. There were some initial studies that sought to determine the complexity of the learning space in CSWL experiments under which learners can still succeed. These studies manipulated the number of words and referents to be learned, competition from alternative referents, co-occurrence frequencies, and referential uncertainty (Kachergis et al., 2009; Yurovsky et al., 2013; Suanda et al., 2014; Suanda & Namy, 2012; Smith et al., 2011; Bunce & Scott, 2017).

There have also been experiments that extended using CSWL to train 2.5-year-old children to learn verbs. While children were able to learn a set of intransitive verbs above chance, only the group of children with vocabulary scores above the median for the group were able to learn transitive verbs (Scott & Fisher, 2012). Although individual linguistic abilities come into play, this provides some evidence that it is possible to determine verb meanings through CSWL experiments.

Several studies have focused on testing how semantic, syntactic, and social information support learning in CSWL experiments. In terms of contextual effects of learning, it was found that adults learn faster when consistent groupings of objects are shown across learning trials regardless of whether it is a semantically coherent group (e.g., animals) or an unrelated set of objects that just repeatedly appear together (Dautriche & Chemla, 2014). This has relevant implications in naturalistic word learning settings as objects rarely appear in isolation. Objects tend to occur together in contexts whether they are semantically related (e.g., a banana and an apple) or not (e.g., a door and fork in the kitchen). This was supported by Chen and Yu's (2015) experiment which used pictures of real objects that were presented in semantically related or unrelated groups. These familiar objects were paired with novel labels in a CSWL experiment with adults. They found that showing semantically related objects during learning resulted in better performance than presenting groups of unrelated objects together. Findings by Gangwani et al. (2010) have also found that adults are also able to learn superordinate names for categories of novel objects while simultaneously learning one-to-one mappings of novel words and objects. It appears that, at least in adults, the learners are also able to encode information about the situational context and semantic relationships during CSWL.

Work on syntactic learning in CSWL has focused on how determiners (Monaghan et al., 2015) and sentence-level constraints (Koehne & Crocker, 2015) interact with CSWL. Determiners reliably precede nouns in English (Monaghan & Mattock, 2012) and can assist infants to find the referent of a noun much quicker (Kedar et al., 2006; Kedar et al., 2017; Dye et al., 2018). In Monaghan et al. (2015), adults were tested on their ability to use novel words that served functions similar to determiners to aid their learning of novel words that either labeled objects or actions during CSWL. A set of monosyllabic non-words were created to serve as function words that would indicate the grammatical class of words that followed it. One set of novel bisyllabic words served as nouns which labeled geometric shapes and was preceded by a specific function word. The other set of novel bisyllabic words served to label motions (e.g., bouncing, growing, shaking) and were preceded by another monosyllabic word. During the learning phase, participants were shown two geometric shapes each performing a different motion. Two pairs of monosyllabic-bisyllabic words pairs were played, one referring to the name of the shape and the other the motion being shown. Participants were able to learn both novel words for the shapes and motions, which indicates the ability to use the monosyllabic word before it to determine whether the label was for the object or action. This finding is able to address how learners can learn words of different classes in multi-word utterances in CSWL experiments. This question is relevant to development since the input that infants are exposed to is likely continuous speech rather than isolated labels for objects.

Most of the work integrating social cues during CSWL has been computational. However, results of formal simulations and results of an experiment by MacDonald et al., (2017) have so far aligned. Computational studies that used annotated videos of parent-child interactions found that understanding the speaker’s referential intent (Frank et al., 2009) and attention to several social cues such as eye gaze (Frank et al., 2013) improved outcomes of models in determining word-object mappings. This successful understanding of reference appears to be essential for CSWL to occur. When adults were given an alternative instruction not related to discovering word-referent mappings during a CSWL task, they were unable to learn the novel word-referent relationships despite being exposed to systematic co-occurrence frequencies (Wang & Minz, 2018). Adults also take into consideration eye gaze as a reliable cue for reference (MacDonald et al., 2017). This last study will be discussed in later sections as it informs the design of study 3, but altogether these results can be taken as preliminary evidence that understanding cues that indicate a speaker's referential intent can assist learners during CSWL.

The ability to encode phonetic information while tracking co-occurrence frequencies in CSWL experiments has also been studied using both native and non-native contrasts. Phonetic development is closely tied to early word learning development as these two areas lead to mutual improvements in the other (Werker & Curtin, 2005), a relationship that will be fully discussed later in this chapter. In terms of encoding detailed phonetic information during CSWL, research in both adults and 12- to 20-month-old infants is available (Escudero et al., 2016a; Escudero et al., 2016b). Both studies have found that both groups of learners can learn minimal pair words containing native contrasts during CSWL, with minimal pair words being more challenging than non-minimal pair words for both adults and infants.

Recent work by Tuninetti et al. (2020) expanded on these initial findings by using minimal pair words that contained a non-native vowel contrast using CSWL. Adult native speakers of Australian English were trained with minimal pair words that contained Dutch vowel contrasts for one group, while another group of participants was trained on Brazilian Portuguese vowel contrasts. In both experiments, learners successfully learned word-referent mappings despite them containing non-native vowel contrasts. However, participants showed more difficulties learning words containing more confusable contrasts. This study will be discussed in more detail in studies 1 and 2 as it is directly related to the questions they investigate. Overall, it appears that learners can encode fine phonetic detail during CSWL experiments. However, it appears that the degree of similarity between word forms to be learned directly impact performance in CSWL.

The field has also expanded by testing bilinguals and those with developmental or acquired language disorders. The majority of the studies comparing monolinguals and bilinguals during CSWL found that although both groups learn equally well in mapping 1:1 word-referent relationships, bilinguals perform better than monolinguals when two novel labels are used to refer to one object (Escudero et al., 2016c; Poepsel & Weiss; 2014; Benitez et al., 2016). The CSWL abilities of persons with aphasia (Peñaloza et al., 2017), children with autism (Hartley et al., 2020), and children with DLD (McGregor et al, 2022) have also been studied and while they do retain the ability to learn using this paradigm, their results were qualitatively different from controls. Persons with aphasia had poorer outcomes and showed slower learning compared to both healthy young adults and an older control group and their performance was modulated by the severity of their aphasia. School-aged-children with autism, on the other hand, were equally as accurate as language-matched controls in identifying referents, but were significantly slower (Hartley et al., 2020). This is in contrast to findings in 7-year-old children with DLD where they showed poorer performance both on measures of learning and retention, which was attributed to weaknesses in their initial encoding (McGregor et al, 2022). These findings suggest that the outcomes of CSWL vary accordingly based on individual linguistic experiences.

Lastly, there is a longstanding discussion on how exactly learners collect and organize co-occurrence frequencies during CSWL. As Gleitman (1990) surmised from the beginning, CSWL quickly grows in complexity once the set of words to be learned the list of possible referents continues to grow. There must be a way that this information is handled to avoid overwhelming the learner and two major accounts have been proposed. These accounts are the associative learning account and the hypothesis testing account. Both are thoroughly explained in section 1.6 of this chapter and further discussed in the context of study 3 where associative learning is reintroduced specifically in the CSWL context. While there are studies that favor one approach over the other, there also exists an argument that these two mechanisms are one and the same (Yu & Ballard, 2007).

This dissertation continues the work around CSWL in several ways. Studies 1 and 2 investigate if words containing novel phonetic information can be learned during CSWL. These studies differ from Tuninetti et al. (2020) in that it 1) uses a non-native consonant contrast that is heard as the same phoneme in the L1 by naïve listeners; 2) tests learning of the contrast at the perceptual and lexical levels; and 3) compares performances of 4-year-olds and adults. Study 3 incorporates social cues in CSWL to further explore if a reliable cue of a speaker's referential intention aids in reducing the referential ambiguity during CSWL and thus possibly affecting the mechanism that learners employ. Although co-occurrence frequencies are sufficient to determine word-referent mappings, there are other reliable cues present during learning that the learner could choose to integrate. This study aims to investigate how learners observed patterns of recalling word-referent relationships potentially adapt to the demands of the current environment. The following sections discuss literature related to these topics.

1.4 The Link between Phonetic Learning and Word Learning

As Smith and Yu (2008) have discussed in their seminal work with CSWL with infants, the statistical learning involved in CSWL is reminiscent of infants using sequential probabilities to segment continuous speech (Safran et al., 1996) as both involve extracting statistical patterns present in the input. The main difference is that while speech segmentation involves finding sequential regularities in one modality, CSWL requires computing co-occurrence frequencies across two modalities: words and referents. The first link between phonetic learning and CSWL is the shared mechanism of statistical learning.

1.4.1 Statistical learning in speech perception development

Extracting words from continuous speech is challenging as speech does not possess invariable acoustic cues to mark word boundaries. This is the benefit of statistical information as it is rich and readily available in the input. Earlier in this chapter, speech segmentation was mentioned as an instance where infants exploit statistical regularities to break into word learning. For example, 7-9-month-old infants were to prefer statistical information over conflicting stress cues, showing that infants can determine word boundaries independent of their knowledge about the stress patterns in their native language (Thiessen & Saffran, 2003). Saffran et al. (1996) also found that 8-month-olds were able to leverage 2-minutes of statistical learning to segment continuous speech at word boundaries rather than within words.

Yurovsky et al.'s work (2012) investigated the connection between statistical speech segmentation and CSWL. In this study, a corpora of child-parent naturalistic free play tasks were analyzed and two predominant regularities in the input were found. The first was that the label always occurred in the final position of the utterance and the second was that it was preceded by an article. These properties were used to generate an artificial language to which adult participants were exposed to different combinations of these regularities. One group was exposed to both the regularities mentioned earlier (full language condition). Another group was only exposed to an article consistently appearing before a label, but the position of the label was in the middle of the utterance instead of the end (onset only condition). One condition retained only the final position of the labels but did not include an article (position only). The last group were exposed to stimuli that had neither the regularities of the article or the word position (control condition). This group was exposed to the same set of words but without the regularities described above. These utterances were presented in a CSWL experiment where participants heard multiword utterances labeling novel objects. They were then tested on their ability to perform speech segmentation (i.e., identify the words in the artificial language) and their knowledge of word-object mappings.

The results of this study found that participants were successful both in segmenting speech and linking words to referents in the full language condition where both regularities were presented. Participants who were only exposed to the position only condition showed reduced performance in word segmentation but comparable performance in object-label recognition. On the other hand, participants who were exposed to the onset only condition had poorer performance in object-label recognition. Finally, when neither of the regularities were available participants performed poorly in both measures of segmentation and learning word-object relationships. These findings show that extracting regularities for speech segmentation and CSWL is not only possible but also produces better outcomes.

Another area in speech perception development where statistical learning has shown a potentially significant role is in the formation of phonetic categories. Before the end of their first year, infants already show a pattern of speech perception similar to adults who share their native language (Werker & Tees, 1984). Infants' sensitivity to sound contrasts that do not signify a change in meaning declines at around 10 months. This is known as the perceptual narrowing window (Maurer & Werker, 2014).

The formation of these native categories is proposed to emerge from infants' sensitivity to distributional patterns of sounds. Maye et al. (2002) exposed 6–8-month-olds to either a unimodal distribution or bimodal distribution of voiced vs. voiceless unaspirated stop consonants. In the unimodal condition, the stimuli the occurred the most was a token that appeared in the middle of the voicing continuum. In contrast the bimodal condition presented stimuli near the endpoints of the continuum most frequently. Looking times indicated that infants exposed to the bimodal distribution recognized the sounds in the contrast as different, while those exposed to the unimodal distribution did not show discrimination of the voicing contrast.

Word learning places an additional demand that strains speech perception abilities, wherein infants find it difficult to distinguish a contrast in a word learning task that they could discriminate in a pure perception task (Stager & Werker, 1997). However statistical information regarding the distribution of the sounds in a lexical context aids infants so that this is overcome. Thiessen (2007) tested 15–16-month-old English-speaking infants on a minimal pair /t/ and /d/ contrast. When infants were provided additional trials that presented the /t/ and /d/ contrast occurring in different non-minimal pair words, they were able to detect a change in the label assigned to an associated word that required them to detect the contrast. The occurrence of the two phonemes in different lexical and referential contexts heightened sensitivity to the phonemic differences and allowed them to be used in word learning.

There is also a finding that unique referential contexts can also guide phonetic attunement. Yeung & Werker (2009) exposed 9-month-old infants to the voiced Hindi dental-retroflex stop contrast using minimal pair words where each word was presented with a unique object. Looking times from the infants showed that they were able to discriminate the two sounds after being shown consistent pairings with distinct objects. This work was expanded upon in Yeung et al. (2014) to determine whether it was the referential relationship between the words and objects that influenced infants' ability to perceive the contrast as different. In this second experiment with 9-month-old infants but using a Cantonese tonal contrast, an additional three trials were presented before the target trials for one group of infants. The additional trials showed familiar objects being labeled (i.e., banana, car, keys). These trials served to emphasize the referential relationship between the novel objects and labels that followed. Results showed that infants shown the referential trials were able to discriminate the contrast, but this effect is modulated by vocabulary size. It was only infants who had receptive vocabulary scores higher than the group median on the MacArthur Communicative Development Inventories (CDI) who were discriminating the contrast. Collectively these two studies offer some evidence that lexical and referential information also contribute to formation of phonetic categories, which add another link between speech perception and word learning development.

Infants have experience using statistical information to learn patterns in the speech stream which prepares them to parse words and learn their meanings. Statistical learning also supports the formation of phonetic categories, which reorganizes infants speech perception abilities to be more sensitive to contrasts that lead to differences in meaning in their L1. Patterns of consistent acoustic and articulatory features at the phonetic level can be further assembled into consistent word forms, which is the next level of information that becomes available to infants after the initial step of speech segmentation.

1.4.2 Processing Rich Information from Multi-dimensional Interactive Representations (PRIMIR)

The Processing Rich Information from Multi-dimensional Interactive Representations (PRIMIR) (Werker & Curtin, 2005) is a conceptual framework that unifies speech perception and word learning development. Although there are prerequisite speech perception skills to word learning, the latter also drives developments in the former hence further linking the two processes. The framework is comprised of the initial biases that direct infants' attention to specific features of the signal, the resulting representations from the rich information in the speech signal, and the different levels at which these representations are organized. Furthermore, it is explicit in its proposal that statistical learning is the underlying mechanism that allows infants to sift through the input. Therefore, this framework is highly relevant to research in CSWL.

The framework proposes that infants possess innate biases that act as filters for the speech signal. For example, infants show a preference for speech over non-speech from birth (Vouloumanos & Weker, 2007) and they also show a preference for infant-directed speech (Cooper & Aslin, 1990). These biases paired with statistical learning allow the learner to better detect regularities across different levels of representation. At the phonetic level, groupings of acoustic and articulatory features form categories which are reflected through categorical perception (Liberman et al., 1957). There are also features in speech that provide information regarding a speaker's age, gender, and affect which are called by the model as indexical information.

From the phonetic level of representations, these sound units could be grouped together into word forms that also present persisting statistical regularities. At this level, word forms are specifically not tied to any meaning. Statistical learning is also key in discovering these word forms from the speech stream. Studies have shown that infants can better segment novel words if they are preceded by highly frequent words in their natural language such as their name (Bortfeld et al.,2003) or determiners (Höhle & Wessenborn, 2000; Shi et al., 2006). At this level, the infant stores word forms as distinct exemplars. It is only when they form associative links to word meaning and specific forms that they can recognize which features are relevant to the meaning for a form.

These representations can be combined and interpreted across multiple different dimensions referred to as planes. On the general perceptual plane, infants only pay attention to linguistically salient information such as sound categories in speech. This representation can be organized into the next levels. On the word form plane, patterns of coherent phonetic and indexical information are associatively linked to object knowledge. The rich repetition in infant directed speech is crucial in strengthening these links and emphasizing which phonetic variation is consistent across situations. Eventually the repetitions will create enough semantic and phonetic overlap in the exemplars for phonemes to emerge. Phonemes, as abstract units that distinguish meaning, can create a more efficient listener and word learning by guiding them towards linguistically meaningful contrasts. Representations at this level are more abstract and do not carry indexical or allophonic detail. The larger the lexicon grows, the stronger the representations at the phonemic level become. It is in this way that word learning also leads to developments in phoneme categories. Consequently, increased attunement only to acoustic properties that change linguistic meaning also reduces attention to uninformative variation when detecting word forms.

1.4.3 Learning novel phonetic information during CSWL

The sections above have emphasized the shared statistical underpinnings of speech perception and cross situational word learning. This shared mechanism creates a smooth transition between the two linguistic areas as learners can continue to rely on an effective strategy and continuously extend their search to regularities across different hierarchies in the input. The first area discussed was how the prerequisite skill of speech segmentation relies on statistical regularities and the findings of Yurovsky et al. (2012) show that adult learners can simultaneously track information regarding word boundaries and word-object relationships. Second, it was discussed that phonetic learning also relies on distributional learning to establish distinct sound categories. As the infant carries on exploiting statistical regularities in the input to distinguish meaningful from random phonetic variation, it can also extend this strategy to finding co-occurrence between words and referents. It has also been touched upon how phonetic learning and word learning are involved in what can be described as a mutualistic relationship where each area contributes to further developments in the other. Phonetic learning increases the learner’s sensitivity to variability in speech that distinguishes meanings in the L1. Word learning, on the other hand, provides a feedback mechanism by reinforcing these categories when learners encounter sounds in distinct lexical and referential contexts. This then motivates the investigation as to whether learners can simultaneously learn novel phonetic information while also tracking co-occurrence frequencies between words and objects.

Previous work with CSWL and phonetic processing compared 12-20-months-olds' abilities to learn both minimal and non-minimal pair words during CSWL (Escudero et al., 2016b). They were able to fixate longer on the target above chance for both word types, which shows that infants show some ability to encode fine phonetic detail during CSWL. So far, the youngest age that CSWL experiments have been performed with is 12 months old. Infants at this age would already be past the perceptual narrowing window and would already be aided by phonetic categories patterned after their native language. Infants younger than this would still only be beginning to find phonetically consistent word forms alongside mapping these to potential referents. To simulate the task of learning novel phonetic information alongside novel word-referent mappings in infants at 12 months, studies 1 and 2 made use of a non-native contrast in novel words. Although not directly equivalent to first language acquisition, the use of an L2 contrast with learners past the perceptual narrowing window would require them to reorganize their perception to reflect that a set of acoustic and articulatory features regularly create a difference in meaning which is the goal when forming L1 categories.

However due to the challenges in infant recruitment due to the COVID pandemic, the studies in this dissertation have explored this question in adults and four-year-old children instead. The inclusion of children as a comparison group is motivated by the finding that better L2 learning outcomes are closely tied to age (Piske et al., 2001; Flege, 1999) which may present differences in how adults and children are able to learn a non-native contrast in CSWL. These groups have more linguistic experience in their L1 that would influence their L2 learning, which is why a model of L2 speech perception in learning is discussed in the next section as it would be more relevant to the population this dissertation tested.

1.4.4 Second Language Linguistic Perception (L2LP) Revised Model

A model that can help explain the learning of non-native contrasts is the Second Language Linguistic Perception (L2LP) revised. This is a model that describes non-native speech perception in naïve L2 learners as well as the developmental trajectory of their perception abilities to accommodate new categories that did not exist in their L1 (van Leussen & Escudero, 2015). This model describes how speech information from the L2 is organized across multiple representations. It also proposes a set of predictions regarding how L2 learners would perceive non-native sound contrasts that they have yet to acquire and how the necessary adaptations to their native sound categories to result in successful discrimination of non-native contrasts. The assumption of this model is that the beginning state of L2 learning is the end state of L1 learning or that learners use their L1 categories as templates for sounds that they encounter in the L2. This model was chosen as a framework for this dissertation as it explicitly acknowledges that it is compatible with developmental models such as the PRIMIR as both share the assumption that perceptual learning is meaning driven. That is, learners attempt to improve their ability to perceive differences in non-native contrasts in order to create lexical representations in their L2 vocabulary. The explicit link between phonetic learning and word learning is of key importance in both these models and aligns with what studies 1 and 2 aim to investigate by phonetic learning during CSWL.

In this section, only the levels of representation in the L2LP model will be discussed. Its predictions about the results of L2 speech perception in learners of varying experience are covered more in depth in the next chapter. The goal for now is to demonstrate how the PRIMIR and L2LP also organize representations in a similar way with clear links from acoustic stimuli to the mental lexicon.

The first level of representation in the L2LP is the acoustic level where speech enters the peripheral auditory system. The acoustic characteristics are then carried over to the phonetic level where features consistent with the learner's L1 are retained, but with context-specific allophonic detail. These first two levels are what comprise the pre-lexical levels of speech perception. What follows the phonetic level is the phonemic level where only sound units that denote changes in meaning are stored. This level also contains the canonical forms of words and morphemes. Finally, the consistent word forms are linked to their possible meanings at the lexical level. The latter two levels comprise the lexical stages of speech perception. Word recognition is modeled as a step-by-step process across the order these four levels were discussed with links established between adjacent levels.

However, the revised version of the L2LP acknowledges that this path is not the only way towards word recognition. The units of representation and links may be fixed, but the strength of the connections between levels can vary. There is therefore competition for the optimal path for word recognition through opting for the one with the least number of weak links. The strength of the link between the acoustic level and phonetic level is directly inherited from the L1. In the revised L2LP model, the link between the phonetic and phonemic level is not one-to-one and the connections are what gradually shift through the course of L2 learning. Finally since it is unlikely that meanings from the L1 can be directly transferred into the L2 vocabulary, the links to the phonemic level are dependent to the connections present in the earlier levels. Another feature of the revised model is that it allows for parallel processing from initially only modelling sequential processing. The architecture of the L2LP bears similarities to the PRIMIR and emphasizes the direct link between phonetic learning and word learning even in the L2.

1.5. Combining Social Cues with Cross-Situational Word Learning

While in lab-based studies learners can map words onto referents through pure CSWL, there is evidence through computational models that cross-situational information paired with social cues improves learning outcomes (Yu & Ballard, 2007; Frank et al., 2009; Lazaridou et al., 2016). Computational evidence also tells us that co-occurrence frequencies alone may be insufficient in naturalistic word learning settings. Frank et al. (2013) analyzed a set of videos showing mothers and their infants in object-centered play. In this very limited learning space of 6 possible referents, they found that co-occurrence probabilities between the words and toys were truly ambiguous. Although a label was indeed more likely to occur with its corresponding toy, it was not a large enough difference for the computational learner to distinguish it from the other toys in the set. It even appears that a link to a referent cannot be established for 30% of a parent's utterance to a child, because the referent was absent when a label was spoken (Kyger, 2013; Harris et al., 1984). Successful word learning might be a combination of sensitivity to co-occurrence frequencies, social cues, and multiple opportunities for learning.

Statistical patterns in the input can reveal word-referent relationships, but social cues are ostensive and often reliable cues to word meanings that are frequent in infant-directed speech. Speaker's eye gaze, in particular, has been shown to be widely used by infants to learn new words in varied contexts (for review of evidence see Baldwin & Moses, 2001). Eye gaze can be particularly useful in an ambiguous word learning situation as it can help the learner locate what a speaker is referring to. For example, 18–20-month-old infants assign a label to a novel object when a speaker visibly looks at it while labeling but not when this referential cue was absent (Baldwin et al. 1996).

Caregivers often label perceptually present objects that are also the focus of the child's attention (Harris et al., 1983; Tomasello & Farrar, 1986; Kyger, 2013). The focus of caregivers and children's attention is often determined through the direction of their eye gaze, which was found to be a powerful predictor of later language development (Brooks & Meltzoff, 2008). Infants continue to develop this sensitivity to eye gaze throughout the first year of life (see Meltzoff & Brooks, 2009 for a review). A systematic review also found strong evidence that infants are more likely to learn word-referent mappings supported by eye gaze cues by modulating their attention (Çetinçelik et al., 2021).

There is a concern as to whether it is really the gaze that infants follow or merely the head movement of the speaker which is a larger and more salient movement (Butterworth & Jarrett, 1991). Brooks and Meltzoff (2002) isolated the effects of eye gaze through their Gaze Following: Eyes Open/Closed paradigm. Two identical objects were placed on either side of a speaker facing an infant. The speaker turned their head to one side either with their eyes opened or eyes closed. At 10 months old, infants looked significantly longer to the side where the speaker turned only if their eyes were open. Younger infants at 9-month-olds had equal looking times regardless of whether eyes were open or closed. At 10 months, objects specifically looked at by a speaker gain salience and are subject to longer looking times. The authors argue that this could pave the way for labeling of the object by the caregiver which could support word-referent mapping. The authors did a follow-up longitudinal study and found that infants who showed higher gaze following behavior in the open eyes condition had significantly larger vocabulary growth (Brooks & Meltzoff, 2008).

It has not been tested if young learners can make use of the speaker's gaze to resolve inconclusive cross-situational statistics. When infants encounter referents with equal co-occurrence frequencies, will a social cue be useful in this decision? Adults were found to be highly sensitive to speaker gaze in CSWL, immediately settling on a referent for a novel word that an on-screen speaker looked at (MacDonald et al., 2017). In domain-general learning, 9-month-old infants were able to make use of a speaker’s gaze in order to guide them towards a consistent stream of statistical information. These provide some evidence that social cues can successfully guide statistical learning in children.

However, study 3 has only tested the role of social cues in adults but still contributes to the discussion of how a reliable cue for reference could impact the learner’s tracking of co-occurrences. Previous CSWL studies have accounted for the ambiguity problem by manipulating the number of objects on the screen to reflect that the naturalistic learning space also contains a sizeable number of potential referents. This neglects the other aspect of having such a rich naturalistic learning environment which is the presence of multiple overlapping cues that could aid word learning. Eye gaze could also be used to modify referential ambiguity in CSWL. The next section discusses how referential ambiguity relates to affect how learners compile co-occurrence frequencies in CSWL, which leads to how these two issues were investigated in study 3.

1.6. Mechanisms of Cross Situational Word Learning

The underlying mechanisms of how statistics are accrued and processed in CSWL are still debated. The first, often referred to as Associative Learning, states that learners keep track of multiple possible referents for each target word, aggregating co-occurrences, and selecting the most likely pairing based on this information. In contrast to this is the Hypothesis Testing account which maintains that a learner selects a hypothesized meaning for a word and confirms or rejects this hypothesis on the next encounter. If rejected, the learner reformulates the hypothesis and subjects it to further testing. So far, behavioral data, online measures, and computational models have supported one view or the other or even a combination of both.

Although the initial investigation of Smith & Yu (2008) with infants did not comment on the underlying mechanism of CSWL, their replication of this study in 2011 supports an associative learning model (ALM) as a viable explanation of CSWL. Using an eye tracker, they collected moment-by-moment eye movement data and used this to form a model that combined data across all trials. In Suanda et al. (2014), children 5-7 years of age showed an effect of frequency in their incorrect selection of foils. The selection of foils that more frequently co-occurred with targets indicate that more than one possible referent was being considered at the time, lending evidence to the ALM. However, Trueswell et al. (2013) criticize the findings of studies like Smith & Yu (2008) due to the fact that learning is only measured after the entire training period. In those studies, there is no way to measure if learners are aggregating statistics across trials or have arrived at the correct conclusion through another method such as hypothesis testing. The hypothesis-testing approach (HT), also called propose-but-verify in some papers, was tested in adults and 2 and 3-year-old children (Woodard et al., 2016; Trueswell et al., 2013). These experiments required participants to make trial-by-trial selections on their hypothesized referents for pseudo-words. One piece of evidence for HT comes from the learner's incorrect responses, wherein only participants who answer correctly on the preceding trial had a better chance of scoring correctly on the next trial (Trueswell et al., 2013). Meanwhile, the children did not show preference for distractors that merely co-occurred with the target showing no evidence of recall from previous trials (Woodard et al., 2016).

As pointed out by Aussems and Vogt (2020), it seems that studies that used forced-choice tasks in training were more likely to find evidence for the HT approach (Woodard et al.; Trueswell et al., 2013) and those employing more passive, vision-based paradigms find evidence for AL (Suanda et al., 2014; Smith & Yu, 2008). Apart from the task type, it seems that the difficulty of the learning situation also affects the findings on strategies that learners employed. Smith et al. (2011) found that apart from referential ambiguity, interleaved learning trials for a target word are more challenging than consecutive exposures to a single word. Thus, in interleaved learning situations and those with high referential ambiguity learners were observed to employ HT. Otherwise when exposures for a word are consecutively presented and referential ambiguity is low, learners use a full-eliminative approach to CSWL via AL. Aussems and Vogt (2020) confirm that consecutive exposures are much easier for learners. Through inputting gaze behaviors in an algorithm, their findings supported a "conservative" CSWL mechanism wherein learners only select a proposed referent when enough statistics have been aggregated. These studies support the notion that a continuity rather than a duality in the underlying mechanisms of CSWL that depends on the learning contexts at hand.

Infants were also seen to be affected by trial spacing similar to adults. As an example, the younger group of 16-month-old infants in Vlach and Johnson's (2013) study were unable to learn the words if the trials are interleaved. These studies implicate memory demands in these findings. However, a study with school-aged children showed no differences in learning outcomes between the two modes of presentation (Hu et al., 2017). These findings appear to suggest that the performance of adults and infants in CSWL are more affected by trial spacing than of children.

1.7. Adapting Internet-Based Testing as a Result of COVID-19 Restrictions

The beginning of this project coincided with that of the COVID-19 pandemic which made in-person data collection nearly impossible. The experiments performed in this study would traditionally be lab-based and online testing is seen as a less than ideal alternative. However, implementing this study online would add to the evidence available on how online studies can be viable for the future of linguistic research. Online testing has its strengths in its cost efficiency and accessibility that would potentially address two major problems affecting the field: the replication crisis and the bias towards studying Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations (Arnett, 2008; Singh, 2022; Kidd & Garcia, 2022), which are the populations accessible where majority of research is being conducted. While this study itself recruits from populations that would be considered WEIRD, its goal is to provide insights on how reliable experimental data can be collected through internet-based testing and contribute to resources for other researchers looking to recruit remotely.

Implementing experiments online requires less resources for both researchers and participants which make scaling up studies achievable. For example, a study on the critical period hypothesis in second language learning introduced as an online quiz that guesses a participant’s dialect of English recruited over 680,000 unique participants (Hartshorne et al., 2018). Researchers can conduct multiple testing sessions simultaneously compared to inviting participants one-by-one to the lab. Participants also do not have to spend time and money traveling to the lab and can perform experiments at their most convenient time. Recruitment is also facilitated by being able to share a link to the experiment through social media or even dedicated crowdsourcing platforms such as Amazon Mechanical Turk or Prolific. The ease of automating steps from the beginning to the end of data collection leads to the scalability of studies.

Access to a larger pool of participants is relevant to the replication crisis because low statistical power is one of its causes (Stanley et al., 2018). Replicability, or the ability to reproduce the findings of scientific studies, has proven to have low rates particularly in psychology (Open Science Collaboration, 2015). As the field of linguistics moves towards empirical methods and quantitative approaches, this discussion becomes relevant here as well (Berez-Kroeker, 2018; Sönning & Werner, 2021). Failed replications of linguistic phenomena found the original studies to be underpowered (Nieuwland et al., 2018, Harrington Stack et al., 2018). Statistical power is the probability of rejecting a false null hypothesis but is also related to the likelihood of a statistically significant finding to be a true effect (Button et al., 2013). Statistical power increases as sample size increases (Serdar et al., 2021), which is where the scalability of online testing would be beneficial. The availability of resources for data collection is a very real and practical limitation which can be potentially addressed by the ease afforded by online testing.

Not only does online testing allow access to a larger pool of participants but could also pave the way for more diverse recruitment in linguistic research. Behavioral sciences have mainly relied on recruiting undergraduate students in universities as they are close to campus and are considered a homogenous population. Furthermore, an analysis of empirical articles found in six APA journals showed that 98% of first authors in these papers were from the US, Europe, or other English-speaking countries (UK, Canada, Australia, and New Zealand) and 95% of samples come from the same background as well (Arnett, 2008). Both the researchers and samples alike belong to WEIRD societies, who make up the majority in behavioral science research despite comprising the minority of the world's population. The last point in particular is concerning as researchers are trying to draw conclusions about humans without acknowledging that there are issues with the generalizability of the findings (Henrich et al., 2010).

This problem also exists in language acquisition where the majority of articles published in four major journals of the field were on English and Indo-European languages (Kidd & Garcia, 2022). The WEIRD population is well-represented in linguistic research despite not accounting for the majority of language speakers. This creates a problem since these findings are often taken as the norm and languages from under-represented groups are seen as variations of this and is a conclusion based on the sociocultural context of the field rather than a representative sample of languages (Singh, 2022).

Online experiments and crowdsourcing platforms could address these issues by making participating in research more accessible to individuals outside WEIRD backgrounds. Crowdsourcing platforms allow users with access to a mobile device and an internet connection to register. Studies could be advertised to individuals outside of higher education or do not live in proximity to a university or research lab. This would also allow researchers easy access to a larger group of language speakers who may not be local in their geographical area. Researchers themselves outside of the WEIRD context could benefit from online experiments as they make use of more accessible and cost-effective technologies. Desktops and laptops for home use running experiments on web browsers are within 100ms accuracy of recording reaction times (Anwyl-Irvine et al., 2020b). Platforms can also make use of a webcam as an eye tracker. With growing digital access in emerging economies (Pew Research Center, 2019), developing online testing methodologies could benefit local researchers who do not have the same access to resources as those in the West.

1.8. Overview of Experimental Chapters

Study 1 recruited adults to test whether novel phonetic information can be tracked during CSWL. This experiment made use of a non-native Hindi contrast. This chapter consists of two experiments. The first measured sound discrimination performance through an AXB task and subsequent use of contrasts in a lexical context in a word recognition task. The pattern of results in the first experiment required consideration that the AXB task is considered to be heavier in working memory demands compared to other sound discrimination tasks, so a second experiment using an AX task was implemented with a new group of adults.

Study 2 explored the same questions as the previous chapter but recruited 4-year-olds and a new group of adults to compare their performance with. Non-native speech perception is strongly tied to age compared to other linguistic areas, and it was hypothesized that children might present an advantage over adults. Namely, that children were expected to show learning in both sound discrimination and word recognition tasks. The sound discrimination task was also adapted to be more child friendly. The supervised testing sessions held with children over video calls paved way to use screen recordings to capture preferential looking data, which served as an implicit measure of learning for children.

Study 3 recruited adults for a webcam-based eye tracking experiment to compare how the difficulty of the learning situation affects the way learners aggregate cross-situational information. Two factors were manipulated, trial spacing and the presence of eye gaze, to adjust the demands of the task. Trial spacing reflects that there can be delays between exposures to words in naturalistic settings, with a larger delay making the learning situation more difficult. On the other hand, eye gaze is a reliable and frequent cue in naturalistic settings that can reduce ambiguity significantly and can encourage the learner to decide on a referent sooner.

Chapter 2. Adults Ability to Perceive the Non-Native Hindi Dental-Retroflex Contrast following CSWL-based Training

Non-native speakers may encounter speech sounds that are phonemically distinct in their second language but not in their first. Some common examples are the /r/ and /l/ contrast for Japanese learners of English (Miyawaki et al., 1975) and the Hindi dental-retroflex contrast for English speakers (Werker et al. ,1981). English speakers would have difficulty distinguishing minimal pairs /d̪al/ (lentil) and /ɖal/ (branch) in Hindi, while Japanese speakers might find rock and lock which are different words in English quite similar. Inaccurate perception also impacts how L2 learners pronounce words containing these sounds as there is evidence that it partly contributes to inaccurate production (Flege, 1995; Flege, 1992). These non-target pronunciations lead native speakers to perceive an L2 learner's speech as having a foreign accent. An L2 learner's first language has a large influence on how well they perceive non-native speech sounds, but this influence can be outweighed through targeted training of non-native speech perception (Pisoni et al.,1982; Tees & Werker, 1984, Fuhrmeister & Myers, 2017; Fuhrmeister et al., 2020, Hayes-Harb, 2007). However, there still remains much to be investigated regarding which components of training are effective in assisting adult learners to differentiate non-native sound contrasts.

2.1 Theories of Non-Native Speech Perception

The influence of a learner's native language on non-native speech perception is underscored by three prominent theories. Native phoneme categories are well established due to adults' extensive experience with the phonological system of their native language. How accurately a non-native sound is perceived is in turn dependent on its relationship with these native phoneme categories. The Speech Learning Model (SLM) (Flege, 1992) predicts that the larger the phonetic dissimilarity between an L1 sound category and an L2 sound, the easier it will be for a learner to differentiate them and establish a separate category for the L2 sound. The SLM, however, focuses on individual phonemes and does not elaborate further on how a pair of sounds forming a non-native contrast would be perceived. The Perceptual Assimilation Model (PAM) (Best, 1995; Best & Tyler, 2001) addresses this by providing explicit predictions regarding how pairs of non-native contrasts would be perceived by beginner L2 learners.

2.1.1 The Perceptual Assimilation Model (PAM)

The PAM states that listeners will compare a non-native sound in terms of articulatory-gestural properties to their existing native phoneme categories. If the sound is similar enough to a category in their native language, the sound is assimilated or perceived as belonging to that category. If the discrepancy in articulatory-gestural properties is large enough to be perceived, listeners would not assimilate the sound and will consider it as belonging to a separate category. There is evidence that listeners can still assimilate a sound into a native category despite detecting some discrepancies, but the sound is recognized as less native-like (Polka, 1995). Thus, it is important to note that assimilation is not absolute or an "all-or-none" phenomenon.

There are four different patterns of assimilation for non-native contrasts in the PAM. The first is known as the Two-Category (TC) type (Figure 2.1) where each member of the non-native contrast is assimilated into distinct native phoneme categories. This "one-is-to-one" pattern is predicted to be the easiest to perceive and produces near ceiling performance in listeners.

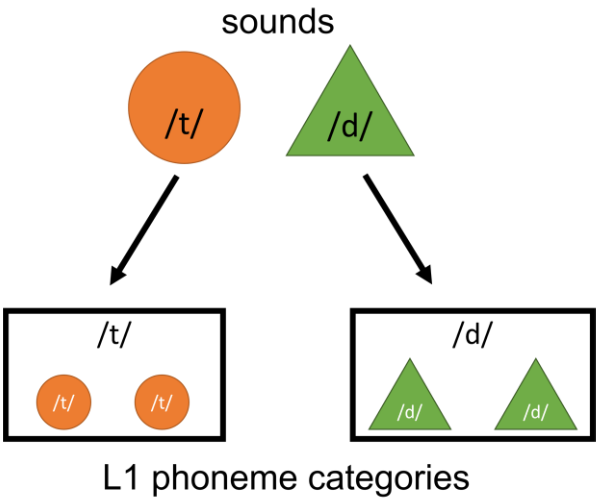


Figure 2.1

Two-category assimilation where two sounds fit into distinct L1 phoneme categories

The next two patterns can be considered "many-is-to-one" relationships since two members of a contrast are assimilated into a single native phoneme category. There is, however, an important distinction between these patterns. The Single Category (SC) type is when both sounds in a non-native contrast assimilate equally well or equally poorly into a single native phoneme category (Figure 2.2a). In contrast to this, the Category-Goodness (CG) type of assimilation is when one member of the contrast assimilates better than the other (Figure 2.2b). This difference creates distinct patterns in perception. The SC type is predicted to have the poorest outcomes in discrimination, as both members of the contrast are basically equivalent in their assimilation into a single L1 category.

The CG type, however, can still produce moderate to good discrimination scores despite both members being categorized into one native phoneme category. As assimilation is not absolute, it is possible to differentiate sounds based on them being good versus poor exemplars of the native phoneme. However, the performance will not be as good as for contrasts in the TC type. There are two varieties of the CG pattern. The first version is expected to produce very good results, which is when one sound is categorized as a good exemplar and the other a poor exemplar. This is known as the uncategorized-categorized pattern. The next variety can produce discrimination performances along a broader spectrum, which is the uncategorized-uncategorized assimilation. This is when neither of the sounds can be perfectly mapped onto an L1 category and discrimination would depend on the similarity of the two sounds with each other and to the L1 model.

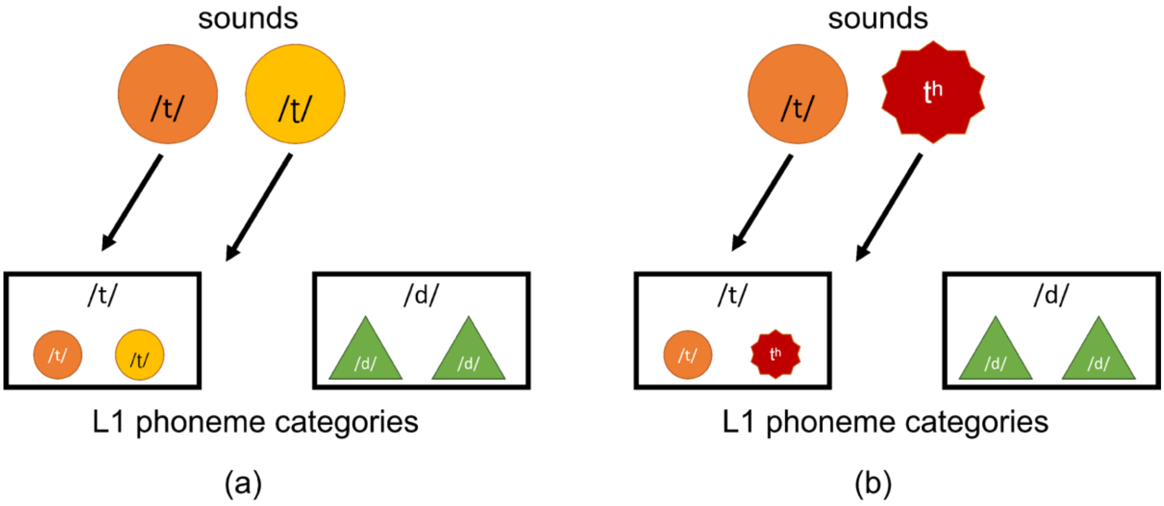


Figure 2.2

Single-category assimilation where two different sounds assimilate equally well into the L1 phoneme category (a) and category-goodness assimilation where both sounds are assimilated but one is considered a poor exemplar (b)

Lastly, there also exists the Non-Assimilable (NA) type wherein neither member of the non-native contrast is assimilated into any of the L1 phoneme categories (Figure 2.3). This pattern can still produce moderate to good discrimination performances but for reasons different from the TC and CG types. One explanation is that contrasts considered as NA are perceived as non-speech sounds and thus are not discriminated in reference to the L1. In both the SLM and PAM, the L1 phoneme categories exert their influence on whether a non-native sound is discriminable or even considered as speech.

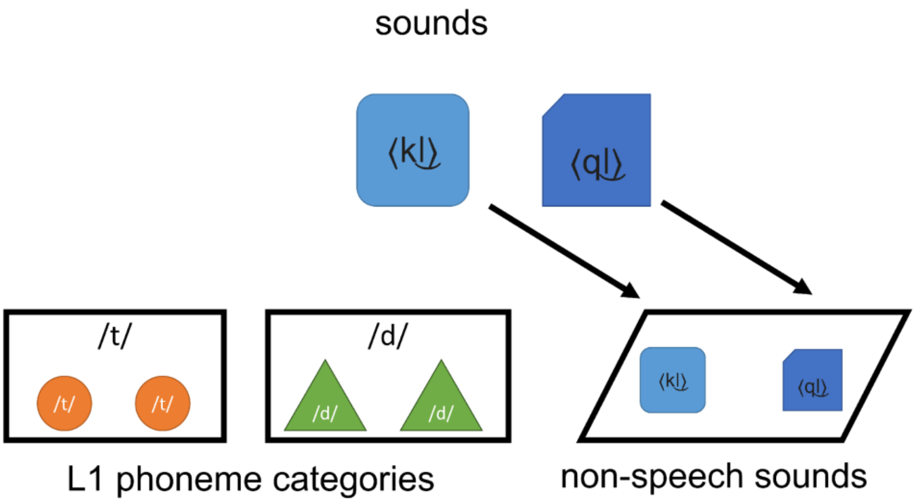


Figure 2.3

Non-Assimilable pattern for sounds which cannot be assimilated into any of the L1 categories

From these assimilation patterns, we can infer that all non-native contrasts are not equally difficult. Some can even be perceived without any specific training. A study testing discrimination of Zulu clicks and other click languages showed that monolingual-English speakers without prior exposure to these sounds performed at near ceiling accuracy on an AXB sound discrimination task (Best et al., 1988). Participants in this experiment also responded that they were sure of their responses on more than 70% of the trials. According to the authors, this performance is due to the Zulu clicks being too distinct to be assimilated into any of the existing phoneme categories in English and were therefore discriminated as non-speech sounds. This performance would indicate an NA type of assimilation.

2.1.2. The Second Language Linguistic Perception Revised (L2LP) model

Previously introduced in the preceding chapter, the Second Language Linguistic Perception (L2LP) model revised (van Leussen & Escudero, 2015) differs from the PAM in that it provides predictions on how non-native speech perception changes from a naive to advanced L2 speaker. Most patterns of perception in the PAM have direct equivalents in the L2LP and the latter further specifies the adjustments required in the learner's phonetic categories in order to perceive the novel contrasts.

The first case we discussed in the PAM was the Two-Category assimilation where two L2 sounds are distinct enough to be assimilated into separate L1 categories (Figure 1). In the L2LP, this would be called a "similar" scenario. Perceiving the contrast simply requires copying the L1 categories and adjusting the boundaries to match those of the L2. Both models predict this pattern to be the easiest.

There are two “many-to-one” assimilation patterns discussed in the L2LP. One of these is the single-category assimilation in the PAM. This would be known as the “new” scenario in the L2LP where more than one L2 sound closely maps onto a single L1 category. This would be a difficult contrast to perceive in both models. Thus, learners must either create a new L2 category or split the existing L1 category.

The second many-to-one pattern is the category-goodness variety. These are known as “subset” scenarios in the PAM. This is either when one sound is perceived as a better exemplar of an L1 category than the other (categorized-uncategorized) or both do not directly fit into any L1 category but are poor exemplars in differing degrees (uncategorized-uncategorized). While the PAM still predicts good discrimination outcomes to be possible with the CG patterns, the L2LP predicts potential issues beyond discrimination if the L2 learner does not adapt their categories. That is, it may lead to spurious contrasts at the lexical level which can hinder the attainment of full L2 proficiency.

2.1.3 Non-native speech perception abilities in adults

Even if some assimilation into L1 categories did occur, listeners without prior training can still perceive differences on non-native contrasts. Polka (1995) tested a group of native American English speakers on an AXB task and a keyword identification and rating task. The participants were asked to differentiate between tense and lax pairs of high front and back rounded German vowels, which are only considered as allophonic variants and not phonologically contrastive in English. Based on the AXB task scores, their discrimination abilities were above chance even without training. The reason for this performance is said to be reflected in the type of assimilation pattern the participants displayed. Evidence was found for a CG type of assimilation in the identification and rating tasks. In these tasks, participants were provided English keywords which each represented a vowel category in English. After participants heard a word containing a German vowel, they were asked to assign it to the keyword containing an English vowel they thought the word with the German vowel was most alike and to rate the quality of their match from a five-point scale ranging from “poor” to “very good”. There was a preference to map the German tense vowels /u/ and /y/ to a single English category. More importantly, the vowel /u/ received higher ratings and was thus considered to be a “good” exemplar in contrast to /y/ which was deemed to be “less good” due to its lower scores. In accordance with the PAM and L2LP, the AXB task performance and disparity in similarity scores imply that the two non-native sounds were discriminated based on their differing degrees of assimilation to a native phoneme category. It appears that an inherent ability to discriminate non-native speech sounds remains when a perceivable difference can be maintained even in the L1.

Although discrimination outcomes for the SC or Subset patterns are predicted to be poor, naïve listeners do show some evidence of within-category discrimination. Pisoni et al. (1982) asked native English speakers to discriminate among synthetic tokens of the voiced /b/, voiceless aspirated /p/, and voiceless unaspirated stop /p^h/ in an AXB task without any prior training or feedback. Although it is unsurprising that these listeners can identify and discriminate tokens of /b/ and /p/ above chance, the majority were also able to consistently differentiate the third category /p^h/ which is not phonologically contrastive in English. Without any attempts to enhance sensitivity nor redirect attention to the contrast, they concluded that their participants must be using the psychophysical characteristics of the stimuli. They arranged a follow-up experiment that trained additional participants for an hour on four consecutive days to investigate if structured training could further improve participants' discrimination skills. Key components of the training were a familiarization phase and an identification task paired with immediate feedback. As early as the second day, participants were already showing more consistent performance on identifying three distinct sound categories compared to participants who did not receive any training. The authors concluded within-category speech perception abilities can indeed be enhanced with further training.

2.2 Non-Native Speech Perception Training Studies

One such training method employed in experiments is to allow listeners to engage in self-monitored listening tasks. In these experiments, participants are allowed to choose from a pool of stimuli that contains different tokens of the contrasts to be learned and thus determine the selection and sequence of sounds that they hear. They are expected to compare the tokens and discover which categories exist among the stimuli. Researchers only determine the number of sound tokens participants are allowed to select within the training session. Once this number is met, participants then proceed to the test phase. Tees and Werker (1984) made use of this method to teach native English speakers to discriminate either the Hindi dental and retroflex stops or the voiced and unvoiced breathy dental stops. In training, participants listened to exemplars from one phoneme category and signaled when they wanted to hear exemplars from a contrasting category. There was a total of 300 training trials and after every 50 signaled changes they were given a discrimination test of 10 trials. Discrimination was tested using a category-change task, where participants should press a button if they detect a change in the category of the sound being played. If they were successful on these 10 trials, they were given 20 more discrimination trials. Authors found that this type of training can produce improved discrimination outcomes immediately and after a delay, but the outcomes are directly related to the difficulty of the contrasts themselves. Namely, they found that participants assigned to the voiced-unvoiced contrast condition outperformed those learning the dental-retroflex contrast. Voicing is a contrastive feature in English which would have led to a two-category pattern of assimilation leading to better discrimination outcomes. The retroflex, on the other hand, does not create a meaningful contrast in English and would be assimilated into a single L1 category thus leading to poorer discrimination outcomes.

In comparison, category-based and identification-based training methods allow listeners less control as experimenters already predetermine the order in which stimuli are presented during the experiment. Category-based discrimination training requires listeners to listen to two tokens and decide if they are the same or different sounds at the phonetic level. Alternatively, identification training exposes participants to one token at a time and asks them to decide on which category the sound belongs to. Participants are told which sounds from which language they need to discriminate at the beginning of the experiment. In cases where participants already have previous experience with the target language, such as those in Flege (1995) and Bradlow et al. (1999), they are provided the standard written orthography of the sounds to represent the categories. However, familiarization can also be done by allowing them to listen to tokens representing each of the categories. This was what Strange & Dittman (1984) did before implementing a category-based discrimination training in Japanese speakers learning /r/ and /l/. After the familiarization, participants had to press a button to indicate if they thought the two sounds were the same or different. Feedback was provided by a light flashing over the correct button. While they did show improvements on discriminating the synthetic stimuli used in training, there was limited carry over to a new word contexts or natural speech. Identification training of the same contrast produced improvements in discrimination in the immediate post-test and three months post-training (Bradlow et al., 1999). Flege (1995) directly compared the category-based discrimination and identification directly to see if one method was more effective than the other. They found no significant differences in the improvements in discrimination related to each training method.

Another variety of training that has been effective in enhancing perception of non-native contrasts is word learning. Studies have shown that lexical knowledge has an impact on phoneme discrimination abilities (Mora, 2005; Feldman, 2013; Freeman et al., 2021;). Mora (2005) tested English as a second language learners who spoke Spanish/Catalan as their native languages on their ability to discriminate L2 contrasts embedded in real words than non-words. Overall, the learners were significantly more accurate in perceiving the contrasts in known real words than in non-words. The authors explained this finding by explaining that the non-words mixed in with the real words influenced participants to adopt a phonetic approach to discrimination over a phonemic one, since non-words are not tied to meaning. This would allow them to focus on the acoustic features of the stimuli which reduces the assimilation influences from the L1. Adults can also be trained with simple lexical information without ties to semantics to categorize non-native sounds as distinct from each other. Feldman et al. (2013) found that adults who heard two non-contrastive vowels in different words were more likely to categorize the vowels as different than if they had heard them appear in the same words. Based on this, learners consider lexical information surrounding sounds to aid in discrimination.

Minimal pairs are a lexical context used in training studies to emphasize the change in meaning signified by a single sound contrast. In two separate studies, adults were trained to discriminate the Hindi voiced dental-retroflex contrast using this context. Minimal pairs containing either the Hindi dental or retroflex voiced stop were paired with unique visual stimuli (Fuhrmeister & Myers, 2017; Fuhrmeister et al., 2020). Participants were first familiarized by showing each of the objects paired with their corresponding word five times. Afterwards, they heard a word and were shown two images and were asked to select the one that matches. They were given feedback on whether they were correct or incorrect for each trial with a total of 200 training trials. There were baseline measures of discrimination abilities before the training using an AX task, as well as measures immediately after training. An identification post-test, which was similar to the training phase but with only 50 trials and without feedback, was also administered. Participants showed improvements in discrimination and the identification task from the baseline assessment to the post-test. This finding provides further support for the phenomenon whereby minimal pairs provide additional support in the detection of non-native contrasts.

Minimal pairs have also been effective in teaching discrimination of non-native phonotactic sequences. Davidson et al. (2007) trained participants to perceive the differences between a phototactically possible initial CVC sequence in English (e.g., /zəgamo/) versus an impossible initial CC sequence (e.g., /zgamo/). Akin to the assimilation that happens at the phoneme level, listeners also assimilated these phonotactically impossible CC sequences into permissible CC sequences. In addition, the authors also examined whether simply embedding the contrast in a word and assigning it a referent is enough to improve discrimination performance or whether it is the use of minimal pairs which emphasizes the contrastive nature of the sounds. They then designed an additional experiment using non-minimal pair words that either contained possible or impossible CC sequences. Both training conditions produced improvements in discrimination for the syllable sequences. However, the improvements of the non-minimal pair condition were limited to tokens produced by the same speaker in training, whereas participants who were trained in the minimal pair condition exhibited improvements that were generalized to tokens produced by a novel speaker.

The training methods we have discussed so far rely on two things. The first of which is access to metalinguistic knowledge about the contrast they are learning. Participants are given information about how the contrasts they are hearing differ from each other or are provided the exact number of sound categories they are expected to learn. While information like this can certainly be present in structured L2 learning situations, often sounds in the L2 are introduced to adults simultaneously through vocabulary.

The second component is the use of immediate feedback to shift learners' attention to the relevant differences with the goal of forming the correct categories. Feedback has a large influence on what features adults attend to during learning. It can cause adults to shift their attention to other phonetic features, as well as increase the number of features they pay attention to and adjust the weight each feature is given when categorizing phonemes (Goudbeek et al., 2007; Harmon et al., 2019; McCandliss et al., 2002). Adult learners also favor using feedback over other conflicting forms of information when learning a novel sound contrast (Kabakoff et al., 2020).

However, feedback is not always present in all settings where a second language is encountered. Whether in class or in a conversation, learners are likely to receive more deliberate and explicit feedback on their production than their perception of individual sounds. This is because listeners can more easily notice and possibly correct production difficulties whereas perception difficulties are much harder to detect. Therefore, there must exist some mechanism that can allow learners to train their perception in situations where both metalinguistic information and feedback are unavailable.

2.3 Distributional Learning

One such mechanism that has been shown to enable phonemic categorization without feedback is distributional learning. Maye & Gerken (2000) found that adults can exploit the distributional properties of the input to categorize sounds. Although there is considerable variability in speakers' productions, the most frequently heard tokens for one phoneme category would fall into a single cluster along the continuum (i.e., a unimodal distribution). If there exist two contrastive phoneme categories in the language, the tokens for each of these sounds would fall on two different clusters. Despite some overlap, they would be represented distinctly on different ends of the continuum forming a bimodal distribution.

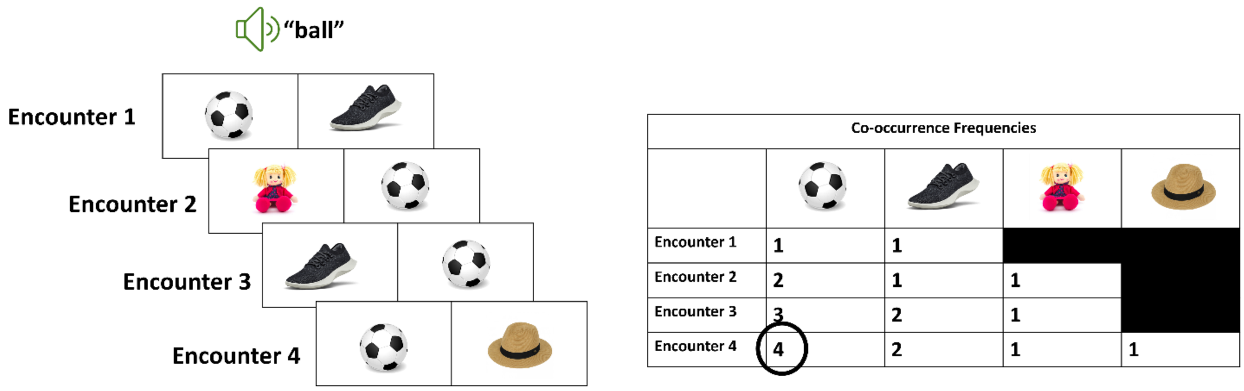
Maye and Gerken (2000) tested distributional learning by presenting native English speakers on a non-native sound contrast. They synthesized a continuum of naturally produced syllables that varied in equal steps in their VOT. The syllables from this continuum were either presented in a bimodal or a unimodal distribution to the participants. In the unimodal distribution a sound from the middle of the continuum was presented with the highest frequency, providing evidence that the acoustic differences between tokens were not contrastive. On the other hand, the bimodal distribution presented the two sounds close to the edges of the distribution most frequently providing evidence for the opposite. Since the bimodal group received distributional evidence that the two sounds were contrastive, they were more accurate on the same-different discrimination task than the unimodal group. This suggests that learners engaging in distributional learning use the most frequently occurring sound tokens as the potential nucleus of the sound category, while the less frequently occurring tokens are considered acoustically similar and therefore assigned as members of the category. Without feedback and knowledge about the phoneme categories, adult learners can extract phonetic category information from statistical patterns in the speech input. The success of distributional learning in training discrimination abilities has been replicated by other studies (Escudero & Williams, 2014; Kabakoff et al. 2020).

Findings by Hayes-Harb (2007) claim that word-based training procedures were superior to distributional learning in training discrimination between allophonic variants of the word-initial /g/. They found that participants who were trained using words paired with unique referents performed better in an AX task than those exposed to a bimodal distribution.

However, none of their participants were able to display knowledge of the contrast in a lexical based task where they had to match words containing the novel contrast to pictures. The previously discussed studies that used minimal pairs have also provided feedback during the training phase (Fuhrmeister et al., 2020, Fuhrmeister & Myers, 2017, Davidson, 2007). It is worth asking how much of the improvement in speech perception can be attributed to training sounds in a lexical learning context or specifically to the feedback provided during training. While it is in the best interest of learners to be provided with as much helpful information as possible, it is in the best interest of research to determine the exact contributions of each element in the training methods administered.

2.4 Cross-Situational Word Learning

In the current study, we attempted to incorporate the learning of a non-native contrast in a word learning paradigm that relies on distributional information from the learning scenario and does not provide any explicit feedback. Cross-situational word learning (CSWL) is a proposed mechanism of word learning where naïve word learners can extract word meanings despite the ambiguity that surrounds them (Yu & Smith, 2007; Smith & Yu, 2008). CSWL posits that the statistical patterns in the input itself can be sufficient for learners to link words to referents. Despite a learner being unable to determine a word's meaning in one specific encounter, they can store a list of potential referents with which the word occurred over time. This stored information is referred to as "co-occurrence frequencies" or how frequently a word occurs with a specific referent. This use of statistical patterns found in the input is a similarity between CSWL and distributional learning. Learners will continue to amass co-occurrence frequencies from ambiguous encounters, and thereby accumulate more information on the word's meaning. Eventually, one object will have a sufficiently high co-occurrence frequency compared to others which will help confirm it as the correct referent. Put into distributional learning terms, the correct referent for a word will thus appear as the single highest peak in a distribution of all other referents that appeared with it (Figure 2.4).



Co-occurrence Frequencies				
Encounter 1	1	1		
Encounter 2	2	1	1	
Encounter 3	3	2	1	
Encounter 4	4	2	1	1

Figure 2.4.

How learners can extract information from ambiguous single learning instances

Bearing this in mind, one can begin to compare the similarities between CSWL and distributional learning. For one, frequency information present in the input is what guides the learner towards the categories and meanings. Initial experiments demonstrated that adults were successful in CSWL after only six minutes of exposure to ambiguous learning trials (Yu & Smith, 2007). Feedback is no longer needed to guide learning as learners have already extracted the co-occurrence frequencies from the input. Another similarity is that learners need to attend to multiple learning trials to form an impression of the linguistic features of the learning space. They are not provided any metalinguistic information beforehand such as the number of sound categories that exist or an ostensive naming context. Instead, they must systematically analyze the input to discover regularities without explicit instruction to guide them. Thus, CSWL can be said to rely on implicit learning strategies (Kachergis et al., 2014). Using CSWL to train a non-native contrast can allow us to examine how much adults benefit from lexical information while relying on an implicit learning strategy and without the confounding effects of feedback.

A study where CSWL is used to train a non-native contrast can expand the current knowledge on the limits of this type of learning. The effects of non-native perception on CSWL have been explored in two sets of experiments by Tuninetti et al. (2020). Two groups of monolingual Australian English native speakers were recruited to learn words containing vowel-contrasts in Dutch and Brazilian Portuguese. In this study, the goal was to determine if perceptual difficulties in the L2 would prevent CSWL in the L2. Both studies used a similar experimental design but varied the linguistic stimuli. The task was for participants to learn words containing easy and difficult vowel pairings during CSWL.

For example, the first study using Dutch vowels used words with /i/ and /a/ as an easy contrast. The participants were expected to perform two-category assimilation in the PAM and a similar scenario in the L2LP since both map onto two different vowel categories in the Australian English as well. An example of a difficult contrast was /ɪ/ - /ʏ/. The first vowel, /ɪ/, could be mapped onto either /ɪ/ or /ɛ/. Meanwhile, the second vowel, /ʏ/, could be mapped onto either /ʊ/ or /ʉː/. The majority of the difficult contrasts they used followed a "subset" pattern where pairs can still be possibly discriminated based on how well they fit into a category (i.e., one can be considered a good exemplar and another a poor exemplar). The vowels were embedded in 12 CVC non-words and were paired with unique line drawings. A total of 72 trials were presented during the learning phase with participants encountering two words and two objects at a time. In the testing phase, participants heard a word and were asked to choose between two objects on the screen. The test phase made participants choose between objects that formed an easy minimal pair, difficult minimal pair, or non-minimal pair words.

The results of these experiments showed that participants in both groups were above chance in learning words with easy and difficult minimal pair types, but the latter were found to be more difficult based on lower accuracy scores. It appears that difficulty perceiving an L2 contrast does affect success in learning words during CSWL, but there is evidence that adults can also use co-occurrence frequencies to learn words containing non-native sounds in L2 learning.

2.5 The Current Study

Our study's contribution is testing whether phonetic information following a single-category assimilation or new scenario, the most difficult of the perception patterns according to the PAM and L2LP, can be learned during CSWL. This assimilation pattern is difficult since both the dental and the retroflex voiceless stop used maps onto one category in German and would require listeners to create a new category to accommodate one of the sounds for successful discrimination.

Specifically, we aimed to test if it is possible to extract non-native sound contrasts from words whose meanings are still ambiguous. Current evidence suggests that detailed phonetic encoding can occur during CSWL whether with native (Escudero et al., 2016a) or non-native sounds (Tuninetti et al., 2020) as long as the difference is perceivable. These studies also show that difficulty in perceiving the contrast affects outcomes in CSWL. No studies have yet used CSWL to train learners to discriminate a non-native contrast while tracking co-occurrence frequencies.

In this study, we implemented a pre-test post-test design to train native German speakers in discriminating the Hindi voiceless dental-retroflex stop contrast. This distinction does not exist in German. First, we aimed to test if being exposed to minimal pairs in a CSWL experiment would lead to improvements in the ability to discriminate between the dental and retroflex. Although this contrast is considered one of the more challenging ones to train (Weker, 1984), assigning each member of the contrast to a unique referent was found to be effective in training adults to discriminate this contrast (Fuhrmeister & Myers, 2017; Fuhrmeister et. al, 2020; Davidson, 2007). These studies, however, relied on external feedback to further reinforce the contrast. Nevertheless, adult learners were found to be successful in using distributional information to categorize two sounds into distinct categories (Maye & Gerken, 2000). The CSWL training paradigm used in this experiment provides both components: a unique referential context for each member of the contrast that consequently occur in distinct statistical distributions. Therefore, it is hypothesized that adults would display an increase in their post-test sound discrimination scores following exposure to a novel contrast in a CSWL training paradigm.

Our second question aims to see if learners can accurately retain these word-referent mappings in an identification task. While participants in Fuhrmeister & Myers (2017) and Fuhrmeister et. al (2020) showed above chance performance in the discrimination task as well as an identification task, those in Hayes-Harb (2007) only showed above chance performance for the discrimination task. To account for the discrepancy between the discrimination versus identification tasks, the author proposed that perhaps there is an intermediate phase between learning a non-native contrast and being able to apply it at the lexical level (Hayes-Harb, 2007). Given this discrepancy in performance, we implemented two separate tasks to test discrimination at purely the phonetic level and another at the lexical level.

However, the words used in the identification task in Hayes-Harb (2007) were a new set of words that contained the contrast, different from those used in training. This not only requires identification but tests the ability to generalize as well. In contrast, the Fuhrmeister studies only tested participants in the trained context. In the present study, the identification task only includes words that were trained during CSWL since it would require being assigned a referent. This is more similar to the context in the Furhmeister studies. Thus, it was hypothesized that participants in this experiment would also present above chance performance in the identification task where discrimination between the minimal pairs with a non-native contrast.

2.6. Experiment 1

2.6.1 Method

Participants

We recruited a total of 27 participants (21 female; mean age: 29.4 years) from online recruitment platforms Prolific (www.prolific.co) and the University of Potsdam SONA systems. All were native German speakers and did not report any history of hearing, visual, or language difficulties. Participants received either payment or course credit in return for their participation.

Stimuli

There were 8 pseudowords recorded for the experiment. These were two-syllable pseudowords with a CVC-CV structure. Among these, there were three minimal pairs. Two of the pairs, /taːpsaː/ - /ʈaːpsaː/ and /goːstaː/ - /goːsʈaː/, contained the non-native dental [t] vs. retroflex [ʈ] contrast and were used in the target trials. The first pair, /taːpsaː/ - /ʈaːpsaː/, was used both in the pre- and post-training discrimination task, the training phase, and the final word-identification task. This pair is hence referred to as the trained pair. The second pair, /goːstaː/ - /goːsʈaː/, only appeared in the pre- and post- training discrimination tasks and is thus referred to as the untrained pair. It was included to measure generalization of the ability to discriminate the contrast. The third minimal pair, /tɪlmɛː/ - /dɪlmɛː/, contained the native voicing contrast [d] vs. [t]. This pair was used in control trials to see if participants understood the task and were paying attention throughout. Lastly, the two remaining pseudowords, /lɛːbnoː/ and /pɪnrɛː/ were a not minimal pair and did not contain the non-native contrast. This last pair appeared in the training phase and word recognition task as control items to see if participants were learning the object labels. The last pair does not appear in the AXB task.

A male native Hindi speaker recorded the two pairs containing the non-native contrast as well as the single non-minimal pair without the contrast. The minimal pair with the native voicing contrast was recorded by a male native German speaker. Recordings were made in a recording booth with Audacity (Audacity Team, 2021) and Praat (Boersma & Weenik, 2023) was used for cutting the wav files. All recordings were normalized to 70 dB.

For the training phase and word recognition task, we used images of 4 unfamiliar objects taken from the Novel Object Unusual Name (NOUN) Database (Horst & Hout, 2016). The objects were numbers 2005, 2013, 2033, and 2052. A unique label from the pseudowords we created was assigned to each of the objects. The object and label assignments are depicted in Figure 2.5.

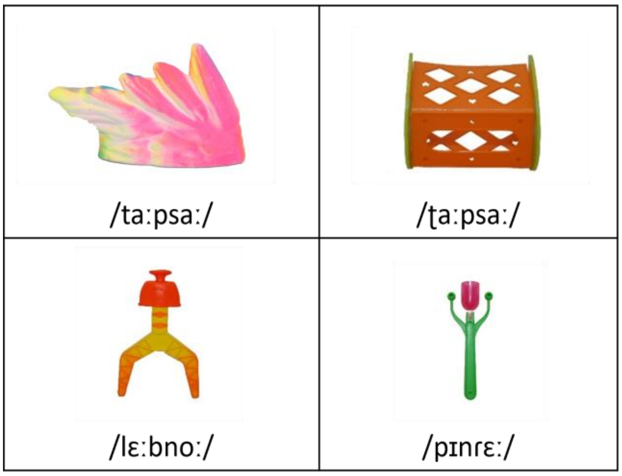


Figure 2.5.

Unfamiliar objects and their assigned pseudowords

Design and Procedure

The experiment was conducted online using the Gorilla Experiment Builder (www.gorilla.sc) (Anwyl-Irvine et al., 2018a). The experiment link could only be opened on a desktop computer. Once participants opened the link, they were guided through a browser audio and headphone check adapted from the dichotic listening task by Milne and colleagues (2021). Participants who did not successfully complete this check were not allowed to continue with the experiment. They were informed that their device was not suited to run the experiment properly.

The experiment followed a pretest posttest design. It consisted of three different tasks: a sound discrimination task, a training phase, and a word recognition task. The sound discrimination task was run at the beginning of the experiment as a pretest and repeated after the training phase as a posttest to measure the effects of training on the discrimination performance for the non-native sounds. The training phase was a cross-situational word learning task. Finally, after the posttest discrimination task a word recognition task was run to test whether the participants could associate the labels from the training phase to their corresponding objects.

Sound Discrimination Task. We used an AXB task to assess participants' ability to discriminate the non-native contrast. In this procedure, three words were played consecutively in each trial. Using a keypress, participants were to indicate if either the first or the third word was more similar to the second word. If the first sound is more similar, participants had to press the “y” button on their keyboard; if their answer was the third sound, they had to press the “m” button.

A trial began with a fixation cross accompanied by silence for 1000ms, followed by the three words playing one at a time with a 500ms interstimulus interval. A different recording was used for the “X” and ‘A” / “B” tokens. This is to prevent participants from comparing physical properties of the recordings, which only requires discrimination at the acoustic and not phonetic level. Participants had 2000ms after the third sound to give a response before the trial timed out, but they could respond as early as after the second sound.

There was a total of 36 trials in the entire AXB task, equally divided among the three minimal pairs used in the task. Each minimal pair was tested 12 times and balanced for the four possible sound sequences (AAB, BBA, ABB, BAA) in AXB tasks. The 36 trials were divided into four blocks with an equal number of trials per minimal pair. The trials within a block followed a pseudo-randomized order such that no three consecutive trials tested had the same “X” token to discourage discrimination at the acoustic level. Furthermore, the order of the blocks for each participant was randomized. Participants were allowed to take a short break after each block. Participants completed this task twice: once before and once more immediately after the training phase.

A practice version of the AXB task was also implemented. It consisted of eight trials and were shown prior to the test trials. The stimuli used here were all different from the test trials and only the instructions were the same. Participants were asked to discriminate non-minimal pair pseudowords that only contained native speech sounds. Participants were required to obtain at least five correct responses to move past the practice phase. If they failed to reach the pass criterion, they were allowed to repeat it once more after reviewing the instructions. Failing the practice round twice resulted in a participant being screened out. They saw a screen informing them that their scores indicated that they were likely not paying attention to the task and were not eligible to continue with the experiment nor receive the incentive. They were, however, also provided a link to a form where they can contest this decision and report any other issues (e.g., technical difficulties) that might have caused their poor performance.

Several checks for attention were also included throughout as this was an unsupervised online experiment. For the sound discrimination task, two values were used. The first value was participants’ total number of correct responses to the native minimal pairs at the end of this task. Participants were required to accurately respond to 8 of the 12 native minimal pair trials as one measure of sufficient attention. Participants were expected to perform at ceiling for these trials since they only contain native speech sounds. This measure was intended to screen out participants who were responding randomly without actively attending to the stimuli. The second value that was considered was the total number of timed out trials. These were trials where participants did not respond to within the time limit. Participants had to provide a response to at least 25 trials or 70% of the total trials. This was meant to screen out participants who started the experiment but did not intend to perform it. Participants needed to satisfy both criteria to move on to the next task. As with the practice phase, participants who were screened out were informed as to why they could not continue with the experiment but were also given a chance to contest the decision if it was due to an error on the experiment's side. This is also true for the succeeding tasks as well.

Training Phase. After the pre-test discrimination performance was measured, participants continued onto to the training phase. They watched a three-minute video where they were asked to learn the names of the four unfamiliar objects, two of which were labelled with either the dental (/taːpsaː/) or retroflex (/ʈaːpsaː/) member of the minimal pair which was present in the sound discrimination task. The remaining two objects were labelled with one of two words not forming a minimal pair (/lɛːbnoː/ - /pɪnrɛː/, which participants did not encounter in the sound discrimination task.

There were 24 learning trials in total and in addition an attention-getter screen appearing every three trials. Trials initially showed a fixation cross for 500ms, followed by two objects appearing simultaneously side by side. Their appearance was accompanied by a silence for 500ms. Afterwards, the labels were played one at a time with an interstimulus interval of 500ms. Each label was only played once for each trial. A silence lasting 500ms occurred before the next trial started. This sequence is illustrated in Figure 2.6.

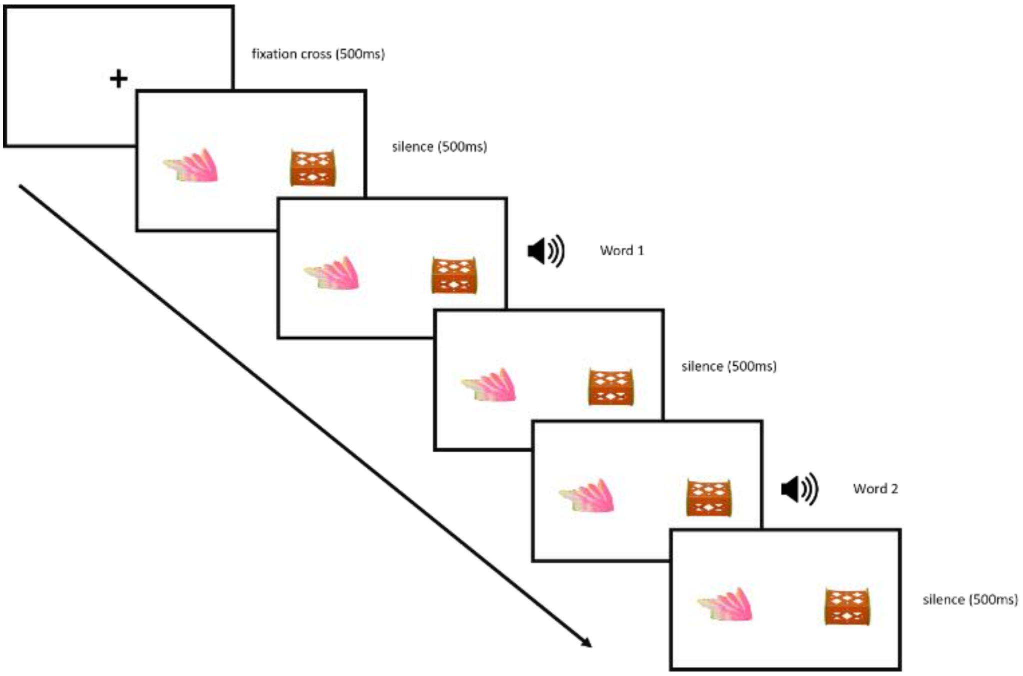


Figure 2.6

Sequence of each trial during the training phase

Each of the four objects occurred with their corresponding label 12 times. There were six learning trials where the objects were labelled with the dental (/taːpsaː/) and the retroflex (/ʈaːpsaː/) together. The remaining learning trials were a combination of any of the four objects that did not form a minimal pair. As is common in CSWL experiments, the position of the objects on the screen had no fixed correspondence to the order of mention of the labels. Participants were not informed of this and were expected to resolve this ambiguity independently. The position of the objects on the screen, the order of mention of the labels, and object pairings were balanced. Each object had a 100% co-occurrence frequency with its label, meaning that each time the object appeared on screen its label was also one of the words mentioned in the trial. The training phase alternated between two different recordings of the dental and retroflex words to label the objects. Half of the exposures for the object occurred with one recording and the other half with another. The variation in sound tokens was also to ensure that participants were not associating the labels to the objects based on acoustic properties. This resulted in two balanced training lists. We randomly assigned an equal number of participants to each.

Prior to the training trials, we included two familiarization trials where common objects (i.e., ball, dog, car) were named. This is to emphasize the referential relationship of objects and the labels that they heard. Each of the labels was preceded by a carrier phrase

"Look! A _____". This carrier phrase was no longer used for the unfamiliar objects in the training trials.

Participants were not required to respond during the training phase, but a short test of attention was implemented after. Participants were asked to press "y" if they had seen this object during the training phase and "m" if they had not. They were shown a total of six objects, only four of which were present in the training phase. Participants needed to answer all questions correctly for them to proceed to the post-test of their discrimination abilities.

Word Recognition Task. For the last task, participants were tested on their knowledge of the object-label relationships they learned in the training phase. We implemented a two-alternative forced choice task consisting of 10 trials.

The test trials began with a fixation cross shown for 500ms, then two objects from the training phase shown in silence for 500ms, and four repetitions of one label. Each repetition was separated by an interstimulus interval of 500ms. Participants had to then click on the object that corresponded to the label after the fourth repetition. This is illustrated in Figure 2.7. There was no time limit for choosing a response. To familiarize the participants with the task, we also added two practice trials at the start of the task where they were asked to identify familiar items (e.g., car and dog).

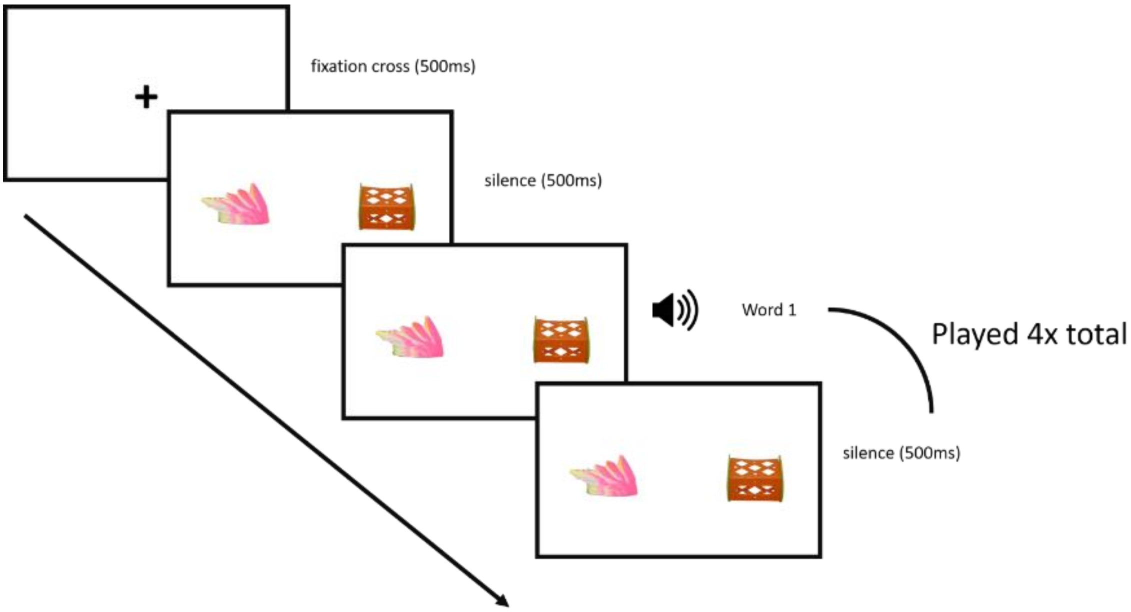


Figure 2.7

Trial sequence for word identification task

Of these 10 test trials, six of them were "minimal pair" trials in which the choices were between the two objects that had been labelled with a member of the minimal pair during training. Successfully answering these trials required an ability to discriminate the non-native contrast. Non-minimal pair trials were a choice between one member of the minimal pair and one of the two non-minimal pair words. The position of the target object on the screen was balanced such that it occurred once on each side in both minimal pair and non-minimal pair trials. As with the training phase, there were two different recordings used to label the object named after a member of the minimal pair. The trials were arranged in a pseudo-randomized order wherein no three consecutive trials tested the same word. Lastly, we interspersed attention getters every two to three trials in this task.

2.6.2 Results

Logistic mixed effects models were used in analysis with the “lme4” package (Bates et al., 2015) in RStudio (R Core Team, 2023). The fixed effects, which were all categorical, were treatment coded. The pre-test was the reference level (0=pre-test, 1=post-test) for the task variable and the trained pair was the reference level for the word pair type variable (0=trained, 1=untrained). Model selection was based on the results of the likelihood ratio test carried out using the anova function (Baayen et al., 2008). The performance on the word recognition task made use of a one-sample t-test to determine chance performance. We also conducted an exploratory analysis of the correlation between post-test sound discrimination and word recognition performance.

Sound Discrimination

All the participants successfully passed the attention checks and were included in the analysis. The proportion of correct responses was used as a measure of their performance in the AXB task. Participants exhibited ceiling performance in the control trials with the native voicing contrast /t/ and /d/ both in the pre-test (M=0.98, SD = 0.12) and in the post-test (M=0.98, SD = 0.13). In principle, participants sufficiently attended to the discrimination task all throughout.

Figure 2.8 displays participants’ performance on the target trials. For the trained word pairs, participants performed similarly before training (M=0.52, SD = 0.50) and after training (M=0.52, SD = 0.50). Neither of the score for the trained pair in the pre-test (t (27) = 0.64, p > 0.05) nor the post-test (t (27) = 0.74, p > 0.05) differed from chance. This was also the case for the untrained word pair when we look at the mean proportion correct in the pre-test (M=0.54, SD=0.50) and post-test (M=0.50, SD=0.50). One sample t-test also showed chance performance for the untrained pair before (t (27) = 1.66, p > 0.05) and after (t (27) = -0.01, p > 0.05) training.

It was hypothesized that the participants would exhibit above chance accuracy in their discrimination abilities after the training phase for both the trained and untrained word pairs. A model with accuracy (0=incorrect, 1=correct) as a function of condition (0=pre-test, 1=post-test) and word pair type (0=trained, 1=untrained) with by-participant and by-items slopes and intercepts for the conditions was initially fit. This model resulted in a singular fit. Random effects structure was simplified until the model converged. The model with the best fit had only by-participant and by-item intercepts in its random effects. Statistical comparison of nested models found no effect of condition (χ2 (2) = 0.71, p = 0.7). The estimates of the full model are shown in Table 2.1

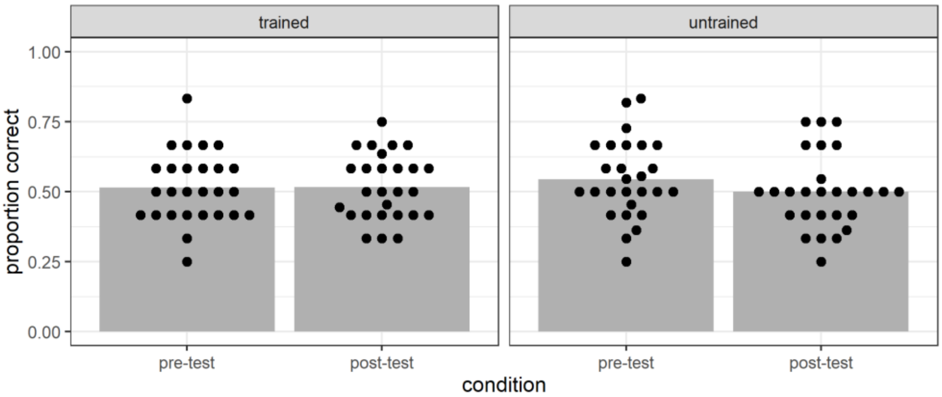


Figure 2.8

Performance on AXB task measured in proportion of correct responses before and after training based on word type.

Effect	Estimate (log odds)	S.E.	z-value	p-value
(Intercept)	0.19	0.35	0.36	
Condition-post-test	-0.11	0.13	-0.84	p > .05
Word Type-untrained	0.06	0.49	-0.12	p > .05

Table 2.1

Model on proportion of correct responses in the sound discrimination task before and after training and in trained versus untrained stimuli

Word Recognition

We also took the proportion of their correct responses in the word recognition task to examine learning of novel label-object mappings cross-situationally (Figure 2.9). Participants exhibited ceiling performance on the non-minimal pair trials (M = 1, SD = 0), which indicates that they were attending sufficiently to deduce the correct associations that existed. The minimal pair trials indicate if this learning extended to objects that contained a non-native contrast in the novel labels. Participants were less accurate on these trials (M=0.71, SD=0.22), but performance was significantly above chance level based on a one-sample t-test (t (26) = 5.12, p<0.001).

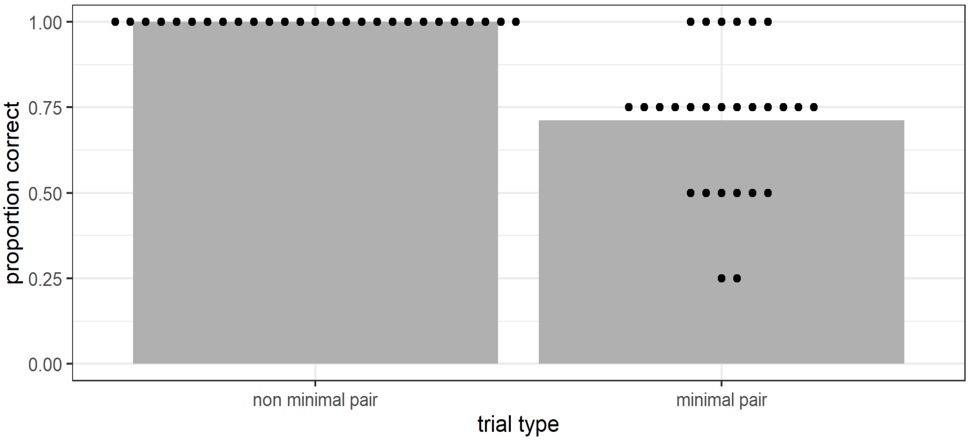


Figure 2.9

Performance on word recognition task measured in proportion of correct answers grouped by trial type

Correlation between post-training sound discrimination and word recognition performance

Given the variability present in the participants' sound discrimination performance, we decided to explore if there was any relationship between this and their performance on the word recognition task. We ran a Pearson's correlation on the proportion of correct responses participants obtained on the post-training AXB task for the trained items and the proportion of correct responses on the word recognition task. As illustrated in Figure 2.10, there was no correlation between these scores (r (25) = -0.27, p > 0.05).

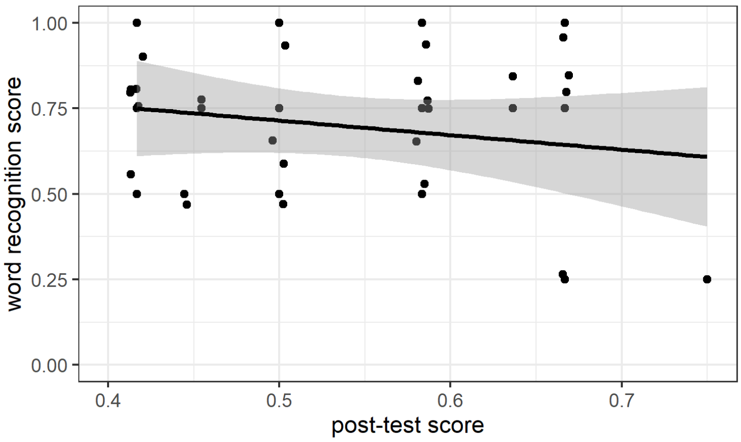


Figure 2.10

Correlation between proportion correct scores on the word recognition task and the trained items in the post-training AXB task

2.6.3 Intermediate discussion

In this experiment, participants performed at chance on the AXB task after undergoing training. While this finding alone might indicate that participants did not develop an ability to discriminate non-native sounds in our experiment, their performance on the word recognition task indicated otherwise. Participants were above chance at selecting the object that occurred most frequently with a novel label, even when the choices were between two objects labeled with a minimal pair that contained a non-native sound. This required some ability to determine whether the label they heard contained the dental or retroflex voiceless stop, and thus some ability to discriminate between the two sounds.

This pattern of performance does not align with the findings of other studies that made use of minimal pairs and images. In the other studies (Fuhrmeister & Myers, 2017; Fuhrmeister et. al, 2020), using minimal pairs in training of non-native contrasts resulted in improved performance in both the post-test sound discrimination task and above chance performance on the word recognition task. Hayes-Harb (2007), on the other hand, only found improvements in the sound discrimination task and suggested that the lack of discrimination in the word recognition task was due to the additional lexical demands required. However, this explanation is not supported by the opposite pattern of performance exhibited by our participants. It appears that the word recognition task, despite the additional lexical component, showed better scores than the test of pure sound discrimination.

The measure of sound discrimination used in the studies cited did differ from what was used in this experiment. While this study made use of an AXB task, the studies above only made use of an AX task. The key difference is the number of stimuli that participants need to recall as they decide. An AXB task requires participants to recall three tokens in total, but the AX task only requires recall of two and respond if they are same or different. The second experiment explores the task effects of the measure of discrimination on performance by implementing the same training but with an AX task as a measure of discrimination.

2.7 Experiment 2

Previous studies have found that discrimination performance is strongly affected by the type of task used (Massaro & Cohen, 1983; Pisoni & Lazarus, 1974; Gerrits & Schouten, 2004). The AXB task and its varieties were found to promote categorical perception, which makes it difficult to discriminate between non-native contrasts that are assimilated into one L1 category. This is said to arise from which information stored in short term memory is used during the task. According to Fujisaki and Kawashima's model of categorical perception (1969; 1970), both auditory and phonetic information about sounds is stored separately in short-term memory. Auditory information is extracted first from the acoustic features of the stimuli and phonetic information is extracted by comparing auditory information to existing categories in long-term memory. Phonetic information is useful when comparing sounds that belong to different phonetic categories in the L1, but auditory information is required if within-category discrimination is required such as differentiating two sounds assimilated into one L1 category. However, auditory information fades quicker in the short-term memory. Therefore, sound discrimination tasks that require the listener longer to respond might lead to auditory information to be less available in short-term memory and thus result in categorical perception.

Pisoni and Lazarus (1974) directly compared the ABX task and the 4IAX task using the same stimuli. The ABX task is another version of the AXB task, but the reference sound appears last. The 4IAX, on the other hand, presents participants with two pairs of sounds and requires them to determine if each individual pair has the same or different sounds. Their findings showed better performance on the 4IAX task in discriminating different tokens along the /i/ and /ɪ/ continuum, while the ABX task resulted in more categorical perception. The performance in the ABX task was reportedly due to participants having to hold both the A and B part of the stimulus before comparing it to the X. Participants needed to wait for all three stimuli to be heard as the comparison being made is relative to the reference sound. This could have led to the auditory information fading in the short-term memory store and participants making a response based on the remaining phonetic information. On the other hand, participants are suggested to require less time to decide in the 4IAX task which makes it more likely that the auditory information was still available. The comparison in the 4IAX task is performed pairwise and thus participants only need to remember two sounds at a time. In essence, the 4IAX task poses less demands than the ABX task which makes it more likely that the required information for within-category information is still available in short-term memory.

A paper on auditory working memory capacity in speech perception supports this explanation. Speech perception was found to be more categorical with increasing memory load and continuous when a single item needs to be remembered. Joseph et al. (2015) measured participants' capacity to remember fine-grained phonetic information through the quality of the representations in their memory as opposed to traditional quantitative measures such as working memory span. By turning a dial, participants could vary a token they were hearing along F1 and F2 dimensions and were asked to recreate the sound they previously heard. They either did this for a single token or multiple tokens at a time. Results showed that participants were more accurate in the single token which implies that they could recall more accurately the acoustic features of the sound. Performance declined as the number of sounds increased, which supports that auditory working memory resources are distributed across demands rather than having a hard limit where performance suddenly declines.

In the second experiment, a new set of participants were recruited and tested using an AX task as a measure of discrimination. This is to further test if the discrepancy in performance was due to the AXB task placing heavier demands on working memory leading to categorical perception. Another change implemented was the order of the post-training tasks. The order of the post-training AX task and word recognition task were balanced. This is to control for the possibility that participants were benefitting from the additional exposures in the post-training sound discrimination task that might have affected their word recognition performance.

2.7.1 Method

Participants

A new set of 26 new participants were recruited (14 female; mean age: 25.3 years) using the same platforms: Prolific (www.prolific.co) and the University of Potsdam SONA systems. Similar to the first group, they were all native German speakers without any self-reports of sensory deficits or communication disorders. Participation was compensated with either payment or study credits.

Stimuli

All the pseudowords, recordings, and images used in experiment 1 were reused in this experiment.

Design and Procedure

The experiment was also conducted online using the Gorilla Experiment Builder (gorilla.sc) (Anwyl-Irvine et al., 2018a). Participants could also only access the link on a desktop computer and proceed to the task once they completed browser audio and headphone checks (Milne et al. 2021). The key differences between this experiment and the one prior are the use of a difference sound discrimination task as well as the order of the post-training assessments.

Like experiment 1, this second experiment consisted of four phases. However in experiment 2, the order of the post-training assessments were counterbalanced. The order of the first two tasks was identical for both experiments (i.e., pre-training discrimination, training phase). For the post-training tasks half of the participants were assigned to the order used in experiment 1 (i.e., post-training discrimination followed by word recognition), while the other half completed the word recognition task before the post-training discrimination task. This was done to control for potential effects of task order within the experiment on participants' performance.

Sound Discrimination Task. This version of the experiment implemented an AX task instead, which required participants to indicate with a keypress if two sounds are the same or different. In each trial, only two words were played consecutively. Afterwards, participants either had to press the "y" button on their keyboard if the two words were the same and the "m" button if otherwise. Participants had to complete this task at the beginning of the experiment as a pretest and once more either directly after the training phase or after the word recognition task as a posttest.

In each trial, a fixation cross appeared for 1000 ms, followed by the first word, a 500ms interstimulus interval, and finally the second word. Participants needed to provide a response within 2000ms after the second word was played.

The AX task also had a total of 36 trials, equally divided among the trained, untrained, and native minimal pairs. These resulted in 12 trials testing each pair equally distributed across four blocks. Trials within a block followed a pseudo-randomized order such that not more than two consecutive trials tested the same word. The order of the four blocks was randomized across participants. Likewise, the possible sound sequences (AA, BB, AB, BA) and the use of two different recordings for each of the target minimal pairs were balanced. This resulted in two lists to which an equal number of participants were randomly assigned to. The same practice phase and criteria for sufficient attention used in experiment 1 were implemented.

2.7.2 Results

A similar analysis to the first experiment was performed. However, one of the participants was excluded as they had participated in another experiment with identical stimuli. Data from 25 participants were analyzed.

Sound Discrimination

All the participants were performing at ceiling level on the native contrast in the pre-test (M = 0.96, SD = 0.2) and post-test (M = 0.98, SD = 0.13). Proportion accuracy on the non-native contrast trials was lower for the trained items (pre-test: M = 0.46, SD = 0.50; post-test M = 0.45, SD = 0.50) compared to the untrained items (pre-test: M = 0.52, SD = 0.50; post-test M = 0.52, SD =0.50) (Figure 2.11). However, trained pair performance remained at chance before (t (25) = -1.89, p > 0.05) and after training (t (25) = -2.06, p > 0.05). This was also the case for the untrained pair (pre-test: (t (25) = 0.77, p > 0.05); post-test (t (25) = 1.13, p > 0.05))

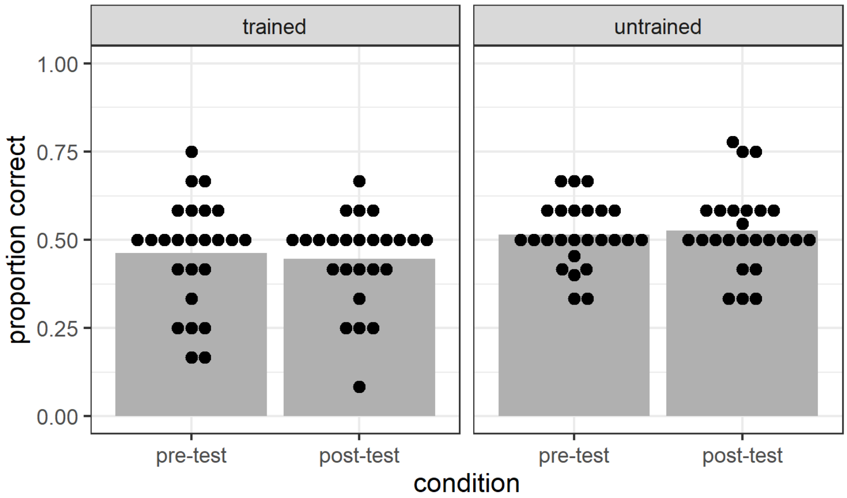


Figure 2.11

Proportion of correct responses in the AX task for the non-native contrast before and after training in the trained and untrained word pair

The maximal model with random slopes and intercepts for items and participants was initially fitted, but this time included the order of post-test assessments as a main effect as well. The maximal model also resulted in a singular fit. Models with simplified random effects structures were compared and the model with by-items intercepts and item by condition slopes is reported (Table 2.2). Statistical comparison of nested models found no effect of performance in pre-test vs post-test (χ2 (3) = 04.53, p = 0.2, word pair type (χ2 (1) = 2.67, p = 0.1), or order of post-test assessments (χ2 (1) = 1.86, p = 0.2).

Effect	Estimate (log odds)	S.E.	z-value	p-value
(Intercept)	-0.30	0.26	-1.13	
Condition-post-test	$-4x10^{-3}$	0.27	0.02	p > .05
Word Type-untrained	0.30	0.18	1.68	p > .05
Post-training order – Identification-Discrimination	0.17	0.12	1.37	p > .05

Table 2.2

Model on proportion of correct responses in the sound discrimination task before and after training, in trained versus untrained stimuli, and order of post-training assessments

Word Recognition

Participants were still at ceiling for the non-minimal pair trials (M=.96, SD=.09). On the minimal pair trials, their performance was less accurate (M=.75, SD=.26) (Figure 2.12) but above chance level (t (25) = 4.09, p<0.001).

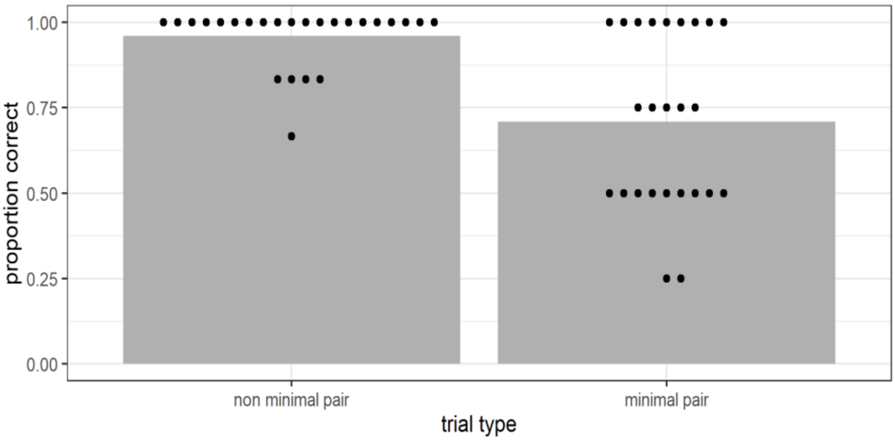


Figure 2.12

Accuracy on object-label relationships in the word identification task

Correlation of Post-Test Discrimination Abilities and Word Identification Scores

A correlation between the scores on the post-test AX task on trained trials and the word identification task was also tested. Results suggest no evidence of a correlation between the scores (r (23) = 0.39, p > 0.05) (Figure 2.13).

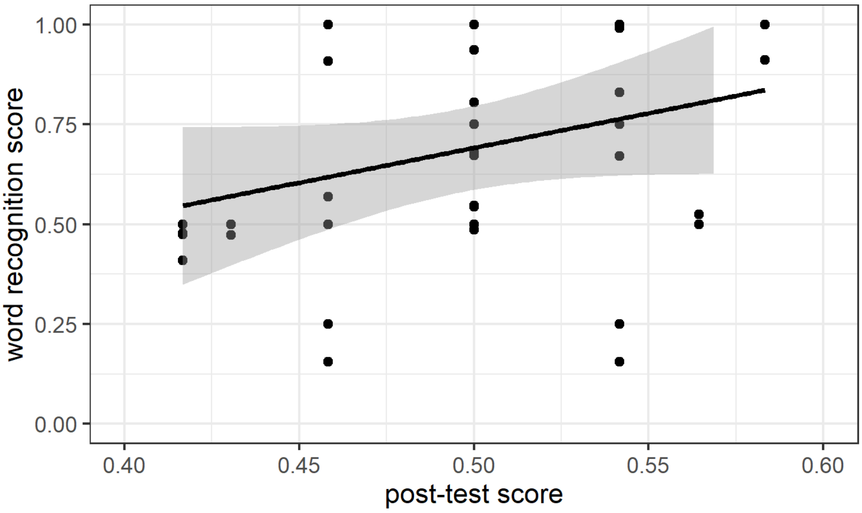


Figure 2.13

Correlation between proportion correct scores on the word recognition task and the trained items in the post-training AX task.

2.8 Overall Discussion

We investigated whether cross-situational word learning can train native German speakers to discriminate the Hindi dental-retroflex voiceless stop contrast, and in the process discover word-referent relationships containing this non-native contrast. We implemented a pre-test post-test design, initially assessing discrimination abilities using an AXB task and word learning through a word recognition task. While we found no evidence for an improvement in discrimination abilities in the AXB task, participants were above chance on discriminating between minimal pair words containing the contrast in the word recognition task. Previous studies have found that discrimination tests like the AXB task impose significant memory demands that might cause participants to discriminate based on a memory trace of the sound rather than the actual acoustic features of the stimuli (Massaro & Cohen, 1983; Pisoni & Lazarus, 1974; Joseph et al., 2015). To address this, the first experiment was replicated but instead used an AX task that requires reduced memory load compared to the AXB. This still resulted in chance performance on the AX task and an above ceiling performance in the word recognition task. It appears that discrimination performance varies between measures that are sound based versus those more lexical in nature. We discuss possible explanations for this below.

The L2 speech perception models we have discussed agree that the influence of the L1 phoneme categories hinders successful non-native perception in naïve L2 listeners. The influence of the L1 was found to increase when non-native speech perception tasks also presented lexical demands, which led to poorer discrimination performance (Freeman et al., 2021). Responding correctly to AXB and AX tasks only requires attention to the acoustic features of the stimuli, whereas a word recognition task requires lexical access (Curtin et al., 1998). Despite this, we have found that participants were performing better on the word recognition task compared to these measures that only required pure sound discrimination. This pattern of performance differs from participants in Hayes-Harb's (2007) non-native speech perception training study. Their finding was that participants performed above chance on either the AX or AXB tasks but not in the word recognition tasks. Although participants exhibit a distinction of the non-native contrast at the phonetic level, this knowledge is not immediately applied into a lexically meaningful context (Hayes-Harb, 2007).

This would be the case if one only considers a bottom-up pathway in developing non-native speech perception abilities. This explanation assumes that learners must first establish distinct phonetic categories for the new contrast that they are acquiring in order to successfully discriminate them at the lexical level. However, our results showed no correlation between the AXB and AX scores and word recognition scores. It was not the case that those who were able to discriminate purely at the sound level were the ones who had an advantage in applying perceiving the contrast in a lexical context.

This possible dissociation between perceptual and lexical abilities in L2 discrimination is accounted for in the L2LP (van Leussen & Escudero, 2015). The L2LP presents computational evidence that parallel processing achieves the same learning as sequential processing in discriminating an L2 contrast. Although the L2LP model states that this is possible, it does not provide an explanation as to why participants would have the strongest connection at the lexical level following CSWL training. A possible explanation could come from the Exemplar Theory as applied to speech perception and word recognition (Goldinger, 1996). This theory states that participants store exemplars of word forms that contain semantic, phonological, visual, talker-specific, and task-specific information. This can include the referential information provided during CSWL which was an additional link presented during word recognition but not sound discrimination. This might have assisted participants in discriminating the non-native contrast by presenting the linked objects to the lexical forms providing them an additional level of representation to distinguish the contrast.

Evidence from two spoken word recognition studies in the L2 also contradicts this idea of the primacy of phonetic categories in L2 discrimination of words containing non-native contrasts. In two eye tracking studies, participants were presented with target words that contained a non-native contrast in a word recognition task (Weber & Cutler, 2004; Cutler et al., 2006). The target word either contained the member of the contrast which had a direct equivalent in their L1 (i.e., native sound) or the sound which was incorrectly assimilated into the same L1 category (i.e., non-native sound). Participants showed increased looks to the competitor when the target contained the non-native member of the contrast but not when the target had contained the equivalent sound in their native language. The authors interpret this result as participants maintaining a distinction between the two sounds at the lexical level despite them being confusable at the phonetic level. Despite words containing a confusable contrast, they can still be stored as distinct forms in the lexicon. Perhaps our participants were establishing some form of distinction between the non-native sounds at the lexical level but not yet at the phonetic level. Thus, this would account for their better performance on the word recognition task despite chance level performance on tests of pure discrimination.

It appears that it is not always necessary for learners to begin discrimination at the acoustic or phonetic level. This implies that there is still a pathway to discriminate at the lexical level in the word recognition task despite our participants not showing the same abilities at the earlier acoustic or phonetic levels.

The findings of this study also differ from those which have found above chance performance in both the sound discrimination and word recognition tasks (Fuhrmeister & Myers, 2017; Fuhrmeister et al, 2020). While the contrast and measures of performance are similar, the largest difference is in the training procedure implemented. Fuhrmeister and colleagues made use of ostensive naming contexts and feedback which were absent in our procedure by virtue of what CSWL is. It appears that the lack of ambiguity and feedback during training provided sufficient support to improve performance across both measures, but nevertheless our participants were still able to learn phonetic information in a word learning context based on their word recognition performance.

Our study is limited in its ability to answer how individual variability played a role in training outcomes. One concern about internet-based studies is that participants were performing the experiments in different circumstances. This is specifically a concern regarding the quality of headphones used and whether it was enough to convey the important acoustic features of the contrast. However, the study being a within-groups comparison reduces the impact of variability in environment and equipment because we evaluated participants' performances against their own. They were also required to finish the task in one sitting to make sure that all tasks were completed in conditions that were as similar as possible.

More relevant sources of individual variability would be participants' language backgrounds, second-language learning experience, as well as their phonological working memory capacity. While none of the participants reported any experience with Hindi or participated in other experiments training the same contrast, people who have more extensive experience with learning languages may have benefited more from the training design of the study. Ideally, participants' previous language learning experience is something that is controlled for. Previous research also cites that phonological working memory is a predictor of an individual's non-native perceptual abilities (McLaughlin et al., 2018). This would be informative for a fine-grained analysis of task effects in this study. There was considerable variability in the AXB and AX performance that could be tested for correlations with phonological working memory measures. This can be in turn used as potential evidence that individual capacities interact with processing demands of discrimination tasks.

We suggest future studies to explore administering CSWL training with more intensity, such as increasing the number of trials or training sessions. The learning phase was only approximately three minutes in length and because of the ambiguity present in the learning situation, participants might benefit from longer training to create improvements in the discrimination tasks. As with many other non-native speech perception training studies, there is also a need to measure performance after a delay. This would inform how effective CSWL is to create lasting improvements in speech perception or to measure how performance changes over time. Collecting longitudinal data is more accessible with internet-based studies considering the convenience it affords participants.

In summary, we found that learners can learn a non-native contrast during cross-situational word learning. We found an asymmetry in their discrimination performance, with participants able to discriminate the contrast in a lexical based task but not in a task requiring only pure sound discrimination. This contradicts previous findings that discrimination tasks that had lexical demands were inherently harder and would result in reduced accuracy compared to AXB or AX tasks. This also presents an alternative to the account that distinct phonetic categories must be established prior to the contrast being used in the lexical context. Taken together, these results show that learners can simultaneously track phonetic information alongside co-occurrence frequencies between words and objects.

Chapter 3. Comparing Adults and 4-Year-Old Children's Abilities to Perceive the Non-Native Hindi Dental-Retroflex Contrast following CSWL-based Training

In the previous experiment, we found that training a non-native contrast during CSWL increased sensitivity in an identification but not in a discrimination task. We attributed this asymmetry to the contrast being represented at the lexical level but not in distinct sound categories. Previous studies have found that adults experience disproportionate difficulty in achieving proficiency in L2 phonology (Pallier et al., 1997; Oyama, 1976; Flege et al., 1999) compared to other domains such as syntax (Boxtel et al., 2003). It is possible for an L2 learner to speak with perfect grammar and an eloquent choice of words only for their accent to give them away as a non-native speaker. The age of L2 acquisition has a linear relationship with the degree of foreign accent and young L2 learners remain capable of producing speech without a detectable accent (Piske et al., 2001; Flege et al., 1999). This suggests that children could also exhibit better non-native speech perception abilities compared to adults and could show improvements from the CSWL-based training where adults did not. We conducted this experiment to directly compare the performance of adults and children.

3.1 Non-Native Speech Perception in Development

Lenneberg's (1967) Critical Period Hypothesis proposes that there is a sensitive period for language acquisition and consequently any learning that occurs after this period would be more difficult. While Lenneberg places the critical period between two years until puberty, the sensitive period for the attunement of speech perception to the sound properties of the ambient language is much narrower and occurs within the first year of life (Maurer & Werker, 2014; Werker & Hensch, 2015). Very young infants are able to discriminate phonetic contrasts even without prior experience with a language (Eimas et al., 1971; Aslin & Pisoni, 1981). For example, English learning 7-month-olds exhibited comparable performance to Hindi native speaking adults in their ability to discriminate the dental-retroflex stop contrasts (Werker et al., 1981). Adult native English-speakers, on the other hand, performed poorer than the other two groups even when trained.

However, infants slowly exhibit changes in their non-native speech perception abilities as development proceeds. A follow-up study on the Hindi dental-retroflex stop contrasts found that there is a steady decline in the perception of the same contrast across groups of

English-learning infants at 6 to 8, 8 to 10, and 10 to 12 months old (Werker & Tees, 1984). The oldest group of infants showed similar performance to the native English-speaking adults. In another study by Kuhl et al. (2006), Japanese and American infants were compared in their ability to perceive the /r/ and /l/ which is a contrast native Japanese speakers have difficulty with. Both groups performed equally well on discriminating the /r/ and /l/ contrast at 6 to 8 months, but at 10-12 months the Japanese infants showed a decline in performance whereas the American infants exhibited improvements. This decline in non-native speech perception abilities has been replicated for other contrasts in cross-linguistic speech perception studies (for a review see Maurer & Werker, 2014). At 10 months, infants show increased sensitivity to phonetic contrasts that are meaningful in their native language and an accompanying decline to contrasts outside of that. This has been referred to as perceptual narrowing. It may be unexpected that development is accompanied by a decrease in sensitivity especially during a time where infants gain many new linguistic competencies. However, an alternative description would be that reorganization is occurring where speech perception prioritizes sounds that create meaning differences in the native language. Models of speech perception development offer differing explanations on the mechanism of this reorganization.

Under the Native Language Magnet Theory Expanded (NLM-e), native phoneme categories instead arise from infants' sensitivity to distributional patterns in the language input and infant-directed speech further exaggerating these differences (Kuhl et al., 2008). Earlier research shows that 6-month-olds' perception of vowels varied as a result of language-specific experience (Kuhl et al., 1992). The acoustic patterns detected in the speech create changes at the neural level referred to as native language neural commitment by the NLM-e. The increase in sensitivity to native phoneme categories is a result of this commitment, which further reinforces the detection of these categories in the ambient language. Consequently, this commitment decreases the sensitivity to non-native sounds. The inverse relationship between native and non-native speech perception is supported by its links to future linguistic abilities. A group of English-learning 7-month-olds were tested on their native and non-native speech perception abilities and had their linguistic abilities assessed at several future timepoints at 14, 18, 24, and 30 months (Kuhl et al., 2005). Results showed better native speech perception abilities at 7 months predicted better language abilities at older ages, whereas better non-native speech perception abilities predicted reduced later language abilities.

The formation of sound categories centered around the L1 strengthens with age and experience. These native categories become the reference for any non-native sounds that need to be perceived. Therefore, the earlier that L2 learning occurs the less influence the native categories exert on speech perception. This can explain why better L2 speech outcomes are tied to earlier age of learning (Oyama, 1976; Piske et al., 2002) and why children are less likely to assimilate L2 sounds into L1 categories compared to adults (Baker et al., 2002). It appears that the more extensive experience of adults in the L1 makes them more susceptible to it influencing L2 perception when compared to children.

3.2 Non-native speech perception abilities in children and adults

The studies cited have so far discussed performance of infants within the first two years of life with some comparison to adults. Studies comparing L2 learners show that the relative disadvantage of adults in non-native speech perception abilities also extends to older children. A comparison of Japanese adults and 10-year-olds learning English as a second language found that children tend to outperform adults on measures of non-native sound discrimination (Aoyama et al, 2003). Snow & Hoefnagel-Höhle (1978) tested native English-speaking children, adolescents, and adults learning Dutch as an L2 on several language measures including those for speech, vocabulary, morphology, and syntax. Tasks assessing speech included a test of auditory perception by identifying minimal pairs as well as spontaneous and imitated productions of words. Adolescents performed significantly better compared to the other groups across all measures. Adults outperformed children on all measures except on auditory perception where the children who were 6 to 7 and 8 to 10 years old made significantly less errors. In these studies, participants had experience with their L2 for at least 6 months in settings such as school or work. It appears that where L2 language learning occurs in naturalistic settings, children achieve better outcomes.

However, studies which train non-native speech perception in learners without prior experience with the language show equivalent or better performance by adults compared to children. Wang & Kuhl (2003) trained groups of native English speaking 6-year-olds, 10-year-olds, 14-year-olds, and adults on Mandarin tones and found that performance increased with age. Heeren & Schoten (2010) trained native Dutch speaking adults and 12-year-olds on the Finnish length contrast and found that adults scored significantly higher than children in discrimination and identification tasks. Although, the comparison of their performance across the five training sessions showed that they were learning at the same pace. Fuhrmeister et al.

(2020) also found that adults were scoring higher compared to 10–16-year-old children on discrimination and identification tasks taken immediately after training. However, results from the same assessments taken the next day show that children's identification scores increased significantly compared to the day before while adults' scores remained the same. It is possible that the cognitive and linguistic skills of adults allow them to outperform children on initial training, but it seems that children are able to catch up with increased learning after a period of consolidation. This may account for the better performance achieved by children in studies comparing L2 learning in naturalistic situations.

3.3. The role of reference in non-native speech perception training

A crucial consideration when working with children is adapting experimental tasks to a format that they could easily understand to ensure that performance reflects the actual variable of interest. One way to do that in these training studies is to associate the sounds to pictures. The images can represent categories of sounds such as when Wang & Kuhl (2003) used animals as labels for different categories of Mandarin tones (e.g., bird tone, frog tone, cow tone, and dog tone). The sound contrasts can also be embedded in minimal pair words that have a unique object assigned to them (Yeung & Werker, 2009; Yeung et al., 2014; Fuhrmeister et al., 2020; Esteve-Gibert & Muñoz, 2021). Brekelmans (2020) tested 8-year-olds and found that embedding a non-native contrast in minimal pairs and assigning a unique referent was more effective than assigning contrasts to categories represented by shapes.

The referential relationship is suggested to be key in enhancing the difference between members of the non-native contrast. Two groups of English-learning 9-month-olds were trained to discriminate Cantonese tones (Yeung et al., 2014). Each tone was paired with a unique object and infants were familiarized with these pairings during the learning phase. One group was assigned to the referential condition which began by naming familiar objects (e.g., banana, car, keys) before showing the novel object-tone pairing. The non-referential condition did not have the initial naming trials with the familiar objects. The ability to discriminate between the tones was subsequently tested using the alternating non-alternating paradigm. In this paradigm, infants were expected to show increased looking times when exposed to stimuli that played alternating tokens of each member of the contrast. Increased attention to this compared to the non-alternating stimuli is taken as perception of the difference between the tokens. The results showed that only infants assigned to the referential condition were able to identify the contrast, but only if they have high vocabulary scores (i.e., at the median or higher). The authors attempt to explain this relationship by proposing that infants with higher vocabulary scores are better word learners and are more sensitive to meaningful variability in phonetic information that signify changes in meaning.

The findings of Esteve-Gibert and Muñoz (2021) further support the importance of the referential relationship in lexical-based training of non-native speech perception abilities. In their work with Spanish-Catalan speaking 4-year-olds, they trained their ability to learn English contrasts when provided social, visual, or referential cues. In the social cue condition, the speaker alternated her gaze between the object and child when producing the target with the contrast to be learned. In the visual cue condition, the speaker looked directly at the child to emphasize the movement articulators while producing the word. In the referential condition, the speaker only looked at the object while naming it. This last condition resulted in the largest improvements in the AX task after training. Picture-based training procedures when the referential relationship between words and objects are emphasized appear to be a viable way of teaching a non-native contrast to children.

Perceiving non-native contrasts in words is much more representative of how children would encounter them in the L2 than hearing them in isolation. It is also more functional as the goal of perceiving contrasts correctly is for listeners to retrieve the correct meaning. If this is the case, then it is important to consider that the contexts in which children can first encounter novel words can be ambiguous. Speakers can talk about objects absent in the immediate environment or may not provide an additional cue to refer to what they mean. Quine (1960) famously illustrates this situation of a linguist hearing a native speaker of an unknown language utter “gavagai” and being overwhelmed with the variety of objects in the surrounding environment it can refer to. This indeterminacy problem is at the core of L1 word learning research (Golinkoff & Hirsch-Pasek, 2000). However, this can also be extended to children who have to learn new vocabulary in a second language alongside novel sound categories.

3.5 The current study

In the previous study, cross-situational word learning (CSWL) has been introduced as a method where learners can combine information regarding word-object co-occurrences to determine meaning. Children have also been successful in performing CSWL in lab-based experiments. Bunce and Scott (2017) found that 2.5-year-olds can learn the referent of a novel word in CSWL with four objects simultaneously present. School-age children who were 5-7 years old also exhibited a similar ability to learn cross-situationally (Suanda et al., 2014). There has been work on using CSWL to train school-aged children to learn novel words in their L2. Hu (2017) recruited Mandarin-speaking 8-year-olds in the third grade who have been receiving English instruction since the first grade. Children were trained to learn 4 real English words (i.e., clamp, wedge, snoot, dart) after 24 learning trials. Results showed that they were able to learn words above chance in their L2 during CSWL.

In this study, four-year-olds were recruited to participate in the training method we designed in the previous experiment. The main research questions of the first study were investigated in this age group, namely 1) does a CSWL-based training using minimal pair words lead to above chance improvements in discrimination of the Hindi dental retroflex contrast; and 2) can learners identify the word-object mappings in a recognition task. A third question is also added regarding the difference in performance of adults compared to children. A new set of adults were recruited to determine whether there will be a difference between their performance and the 4-year-old children in this study.

Given that younger age is a predictor of better outcomes in non-native speech perception (Baker et al, 2002; Aoyama et al, 2003; Snow & Hoefnagel-Höhle, 1978), the children in this study might present improvements in sound discrimination where adults previously did not. While the findings of Snow & Hoefnagel-Höhle (1978) did state that children 3–5-year-olds performed worse than adults particularly in tests of auditory discrimination, there is the issue that they made use of real minimal pair words in the L2. This has been addressed by adjusting the computation of the discrimination scores to exclude mistakes that were not a result of confusing the referents of minimal pairs for each other, yet a better approach would be to ensure that participants were familiar with all the referents presented beforehand. This confounds with receptive vocabulary knowledge which was also be lower in the youngest set of participants compared to adults.

Ideally, this study would have recruited 12-month-old infants who were close in age to the perceptual narrowing window and perform word-referent mapping during CSWL (Smith & Yu, 2008). The studies by Yeung and Werker (2009) and Yeung et al. (2014) successfully reinstated sensitivity to a non-native contrast in 9-month-old infants who were trained using a referential context, which gave the motivation to recruit an age close to this but also capable of learning from a CSWL paradigm. However, there was the practical consideration that internet-based testing for that age group is not well-established. This is a larger concern than when testing adults as the data collected from infants would primarily be implicit measures and it was unclear how good the quality of this data would be when collected with existing remote testing setups.

Thus, it was considered to recruit from older children who could also provide explicit responses to compare with more implicit looking data. After piloting, four-year-olds were found to present sufficient understanding of the adapted AX task as well as complete the expected duration of the task. Children this age still show the ability to learn word-referent mappings during CSWL (Bunce & Scott, 2017; Suanda et al., 2014). Learning minimal pair words which only include native phonemes was also found to be possible for infants during CSWL (Escudero et al., 2016a). Thus, these studies indicate that 4-year-olds would be able to learn word-referent mappings with fine phonetic detail during CSWL. Younger age being tied to non-native speech perception possibly provides them an advantage in perceiving and increasing their sensitivity to the contrast as they are less affected by L1 assimilation than adults. Based on this second point, it is hypothesized that children would perform better than adults in both measures as the ability to perceive the contrast also affects retention of word-referent mappings in CSWL involving non-native contrasts (Tuninetti et al., 2020).

The study provides a novel insight into how children respond to word-learning based training for non-native speech perception when the link between words and referents is not ostensive. The implementation of this study through an internet-based testing session also demonstrated how developmental research could be conducted online while maintaining data quality. In contrast to the adults, children performed the experiments with an experimenter present on a video call. Apart from ensuring that caregivers were not responding on behalf of the child, having the experimenter on a video call allowed the collection of looking behavior during the experiment through recording the video call. The looking behavior in these videos were coded and analyzed similar to an Intermodal Preferential Looking Paradigm (IPLP). The IPLP is a method commonly used with infants and children since they present longer looks toward stimuli that match the linguistic stimulus they just heard (for a review see Golinkoff et al, 2013). Data from the IPLP can measure children's performance without relying on an overt response.

The decision to move the study online was due to the restrictions on in-person testing during the pandemic. However, with advancing technologies online experiments should be considered seriously due to their cost effectiveness and scalability. This study also explored the viability of online methods to collect both explicit and implicit measures of language in preschool age children.

3.6 Method

Participants

We recruited 24 monolingual 4-year-old German speaking children (12 female; mean age: 57.5 months). We confirmed through parental reports that children did not have any history of developmental or language delays nor have had exposure to Hindi. Parents of children were contacted using the participant pool at the BabyLab in the University of Potsdam and the study was also advertised on the the website Kinder Schaffen Wissen (kinderschaffenwissen.eva.mpg.de) of the Max Planck Institute of Evolutionary anthropology. We also recruited a new group of 24 German-speaking adults with no prior experience learning Hindi (21 female; mean age: 24.5 years). Adult participants were recruited via the SONA systems of the University of Potsdam.

Stimuli

The sound discrimination task was divided into 3 rounds where the children encountered different cartoon characters. In the first round, children discriminated between animal sounds. We used cartoon images of a cat, horse, and rooster and recordings of their corresponding animal sounds. In the second round, they discriminated between a sound contrast found in German. We used cartoon pictures of two different girls and one recording each of a girl producing /ta/ or /da/ (Figure 1). In the final round, children were asked to differentiate between the Hindi dental and retroflex voiceless stop. We made use of cartoon images of red and blue aliens and the two recordings each of the pseudowords /tapsa/ (dental) and /ʈapsa/ (retroflex) (Figure 3.1).

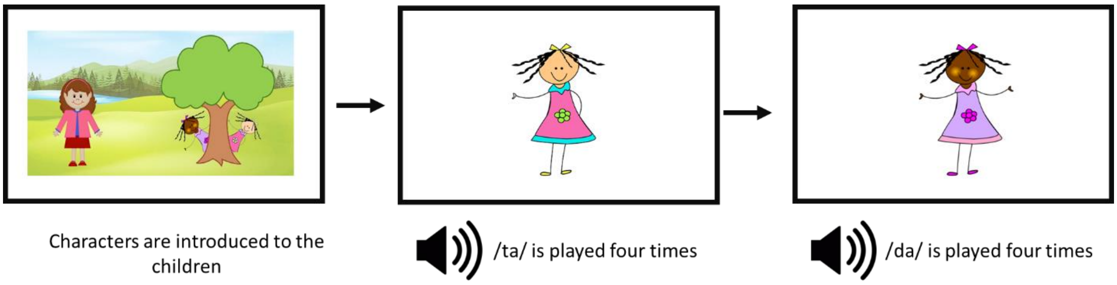

Characters are introduced to the children
/ta/ is played four times
/da/ is played four times

Figure 3.1

Association phase for Round 1 (Native /t/ and /d/ sounds)

The Hindi pseudowords had the target dental or retroflex phoneme followed by the vowel /a/. A male native Hindi speaker recorded at least five tokens of each of the words where two were selected. Audacity was used for recording (Audacity Team, 2015) and Praat (Boersma & Weenik, 2023) was used for cutting the wav files. All of the files were normalized to 70 dB. One native Hindi speaker listened to all of the recordings to check for pronunciation.

The cross-situational word learning (CSWL) training involved 4 novel words to be learned that each had an unfamiliar object as a referent. Two of the words to be learned are the same pair used in the sound discrimination task: (/tapsa/ and /ʈapsa/). The same sound recordings of the minimal pair in the AX task were used in the learning phase. The other two words to be learned didn't form a minimal pair: /lebno/ and /pinre/. These were also recorded under the same conditions described above, but only one token of each was used.

These four pseudowords were assigned to refer to four unfamiliar objects taken from the Novel Object Unusual Name (NOUN) Database (Horst & Hout, 2016). The same objects from the previous experiment were used in this version. During the CSWL training, an attention getter appeared every 3 trials. These were comprised of a static cartoon image accompanied by a related non-speech sound (Figure 3.2).

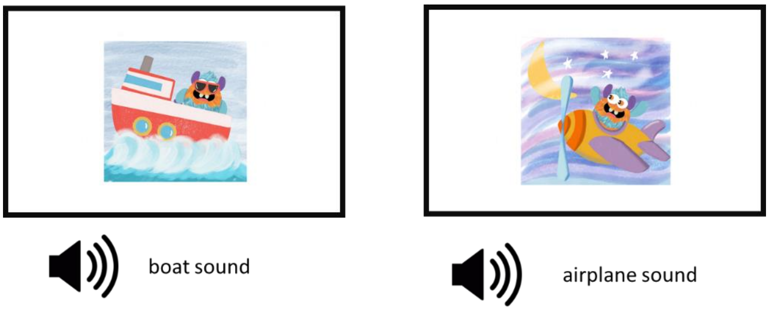


Figure 3.2

Example attention getter trials where a static scene was shown and an associated sound was played

At the beginning of the CSWL training, there were 2 familiarization trials to emphasize the referential relationship between the words and the referents. These made use of 3 familiar words and objects: car, dog, and ball. The pictures were of real objects and the labels recorded by a female native German speaker. The carrier phrase "Look! A ___" accompanied the labels.

The word recognition task made use of the same recordings and images from the CSWL training. At the beginning there were also familiarization trials that make made use of the same familiarization stimuli as in the training phase. The same attention getters are also used here which appear every 2 to 3 trials.

Both the tasks had the instructions pre-recorded and integrated into the experiments. The instructions were recorded by female native German speakers in child-directed speech.

Design and Procedure

The experiment was programmed using Gorilla (Anwyl-Irvine et al., 2018a), an online experiment builder. Testing sessions with children were supervised by an experimenter during a video call. Participants were also asked beforehand to complete a setup so that their device met the technical requirements of the experiment. Once the experiment began, the experimenter recorded her screen so that the gaze of children while doing the experiment could be captured and coded for later analysis. Adult participants, on the other hand, were provided a link to the experiment that they could complete independently.

Sound Discrimination Task. The sound discrimination task was a modified version of an AX task wherein participants needed to respond if the two sounds heard are same or different. This was administered before and immediately after training.

This task took the form of a listening game. Participants were asked to help three cartoon characters find who was hiding behind certain objects. They were asked to listen to two sounds played consecutively and determine if they only heard one character hiding behind the object (same sounds) or two different characters (different sounds). They indicated their choice through selecting from two choices: a picture showing two of the same characters or another picture showing two different characters. There were three different rounds: 1) help a farmer find his animals hiding behind a barn (animal sounds); 2) help a mother find which of the girls is hiding behind a tree (native German contrast); and 3) help a Martian find his friends hiding behind a rocket ship (Hindi dental/retroflex contrast). At the end of the Martian round, 4 attention check trials are administered where children are asked to differentiate between animal sounds again.

The first round served to familiarize participants with the rules of the game. We included two familiarization trials for the succeeding rounds. This introduced participants to the idea that each character "says" a unique sound. In the second round, one of the girls is shown along with four repetitions of /ta/, and the other girl shown with four repetitions of /da/. Similar trials were used for the third round wherein a different Martian is shown producing either a word with the dental or retroflex voiceless stop. The familiarization trials are illustrated in Figure 3.3.

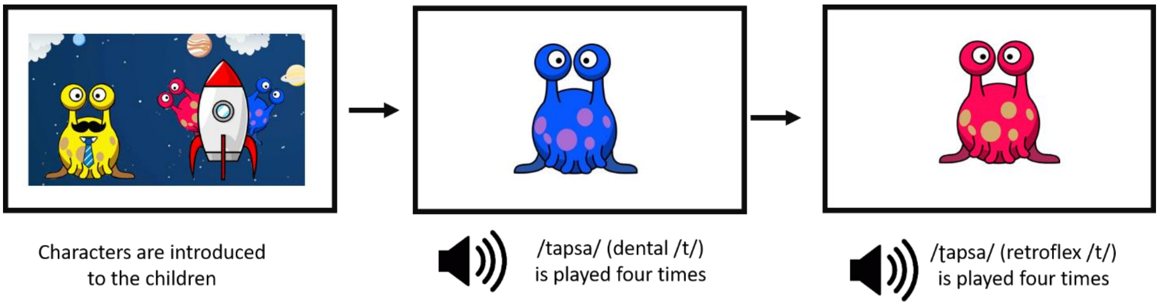

Characters are introduced to the children
/tapsa/ (dental /t/) is played four times
/ʈapsa/ (retroflex /t/) is played four times

Figure 3.3

Association phase for Round 3 (Hindi dental/retroflex sounds)

Each trial began with showing the object the characters were hiding behind and a silence of 1000ms. Afterwards, the first sound is played followed by a 1000ms interstimulus interval. The second sound followed and the two picture choices are immediately shown. Children could click on the picture of their choice or if they point to it the parent could ask the child to use the mouse or click on the child's answer for them. There was no time limit for responding. The trial sequence is shown in Figure 3.4.

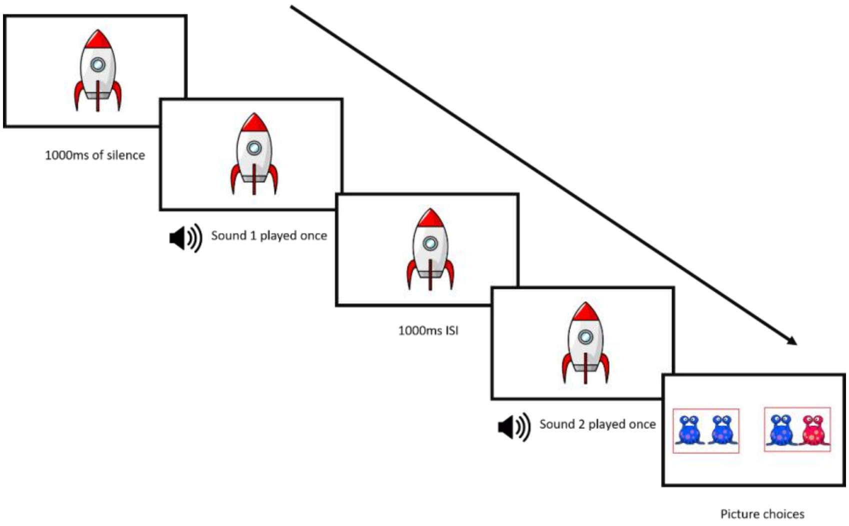


Figure 3.4

Trial sequence for Round 3 (Hindi dental/retroflex sounds)

In the first round with animal sounds, a consistent reinforcement schedule was used to help children understand the rules. For each correct answer, they were shown a screen with the farmer and his animals together and a recorded voice saying they did a good job. If they answered incorrectly, they were shown a picture of the farmer alone and heard a recording to listen again carefully. In the second round, not every trial was provided feedback. In the third round, no feedback was given at all on the accuracy of answers. However, every four trials showed a picture of the Martians riding the spaceship getting closer to their planet each time. These were also accompanied by a voice recording that tells the children that the Martians were almost home. This way, progress in the task was reinforced rather than accuracy.

The entire AX task consisted of 47 to 52 trials. Participants completed a different number of trials in the first round depending on how quickly they reached criterion. Participants needed a total of five correct responses to proceed to the next phase and they can answer at most 10 trials if they don't meet this criterion immediately. The number of trials in the succeeding rounds were fixed with the native sound contrast having 10 trials and the non-native contrast having 16 trials. Participants had to complete all of these. There were an equal number of same and different responses, an equal number of each possible combination in an AX task (AA, BB, AB, BA), an equal number of times that the correct image appeared on either side of the screen.

Training Phase. The training was designed as a task where participants needed to learn the names of a character's favorite toys. Participants only needed to watch the trials without having to respond.

Participants needed to learn 4 novel words that each have a unique unfamiliar object as a referent. Each word occurred with its corresponding object 12 times. To create the ambiguity found in CSWL, we presented 2 objects and 2 words at a time in one trial. The positions of the objects on screen had no relationship to the order of mention of the words. The pairings and position of the objects that appeared on each trial were also controlled so that none of the trials were exactly the same. There were 4 word-object pairs shown 12 times each that resulted in 48 exposures in total. There was a total of 24 learning trials in this task. Every 3 trials, an attention getter appeared. Each object appeared an equal number of times on either side of the screen. Each word was mentioned an equal number of times as either the first or the second word. There were also an equal number of trials where the words were either a minimal pair or not. This resulted in two lists wherein an equal number of participants were randomly assigned.

A trial began with a fixation cross shown for 500ms, then the two images were shown and were accompanied by a silence of 500ms. This was followed by the first word, an interstimulus interval of 500ms, then the second word, followed by a silence of 500ms at the end of the trial (Figure 3.5). The first two trials showed familiar words and objects, followed by the novel words to be learned. The training lasted for at most 3 minutes.

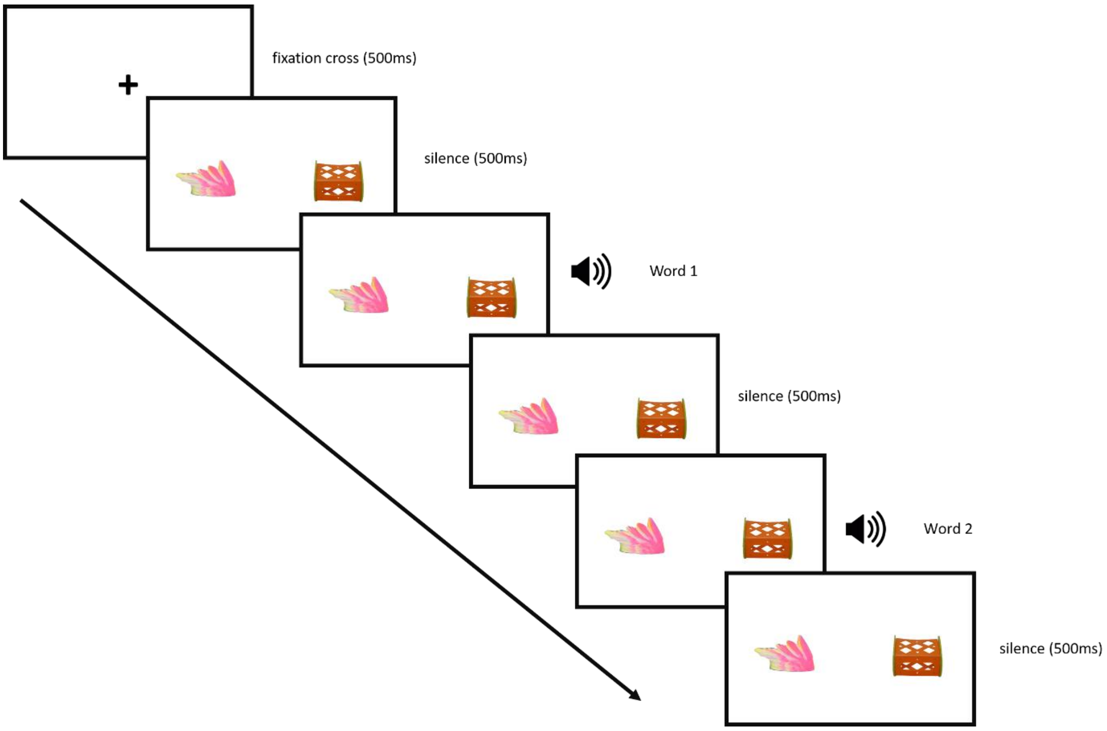


Figure 3.5

Trial sequence during CSWL word learning

Word Recognition Task. The last task was a two-alternative forced choice task with 10 trials. The recorded instructions stated that the participants needed to click on the picture that Molly the Monster named. Each trial began with a fixation cross shown for 500ms, followed by the two picture choices accompanied by silence for 500ms, and the target word. The target word was repeated four times with an interstimulus interval of 500ms each repetition (Figure 3.6). The first two items presented were familiar objects and labels (e.g., car and dog) to test understanding of the task. The non-minimal pair words were each tested twice and the minimal pair words with the target contrast were each tested thrice. Two of these test trials were with their minimal pair to test ability to discriminate between the two words and once with one of the non-minimal pair words.

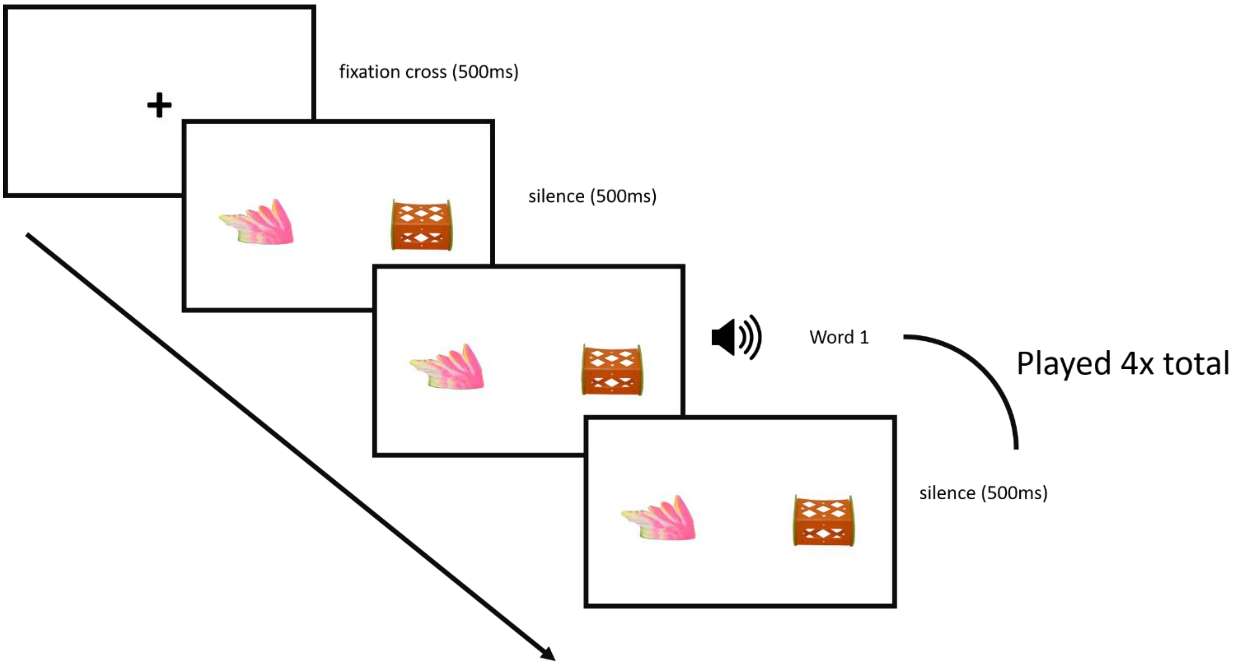


Figure 3.6

Trial sequence for word recognition task

3.7 Results

3.7.1 Behavioral Data

The behavioral responses of 22 children and the full set of 24 adults were analyzed. Two children were excluded as one parent responded that their child had prior exposure to Hindi, and another responded that their child had a history of language delay. The proportion of correct responses was used in visualizations of performance in both the sound discrimination and word recognition tasks. The effect of CSWL training was tested using was through logistic mixed effects models using the "lme4" package (Bates et al., 2015) in RStudio (R Core Team, 2023). Model selection was based on the results of the likelihood ratio test carried out using the anova function (Baayen et al., 2008). A one sample t-test was used to determine if performance was above chance. An exploratory analysis on the correlation between post-test sound discrimination and word recognition performance was conducted.

Sound Discrimination

The proportion of correct responses of participants on the animal and native contrast blocks was used to measure their understanding of the task. Adults were performing at ceiling for the animal (M=1, SD = 0) and native (M = .96, SD = .06) blocks. Children had more variable performance on both blocks (Figure 3.7). Mean performance on the animal trials

(M = .91, SD = .29) was above chance (t (21) = 12.64, p < 0.0001) indicating that children understood the task. Their mean performance for the native trials was lower compared to the previous blocks and when compared to the adults (M = .68, SD = .46) but was still above chance (t (21) = 4.18, p < 0.0001). Children do not perform at ceiling for a contrast that they should be able to discriminate.

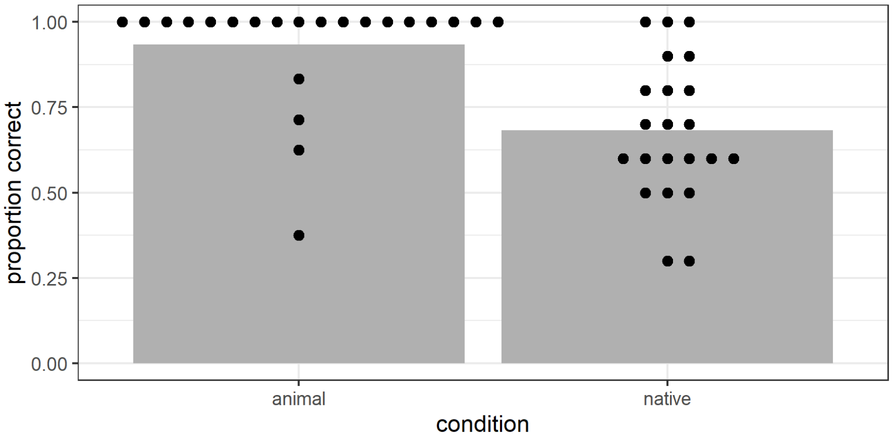


Figure 3.7

Proportion of correct responses by children in the animal and native blocks during sound discrimination

Mapping the individual performance of children in the animal and native blocks (Figure 3.8) shows that those children performing below ceiling in the animal block had a similar performance in the following block. However, even children that were at ceiling in the animal block generally scored lower in the native block as well.

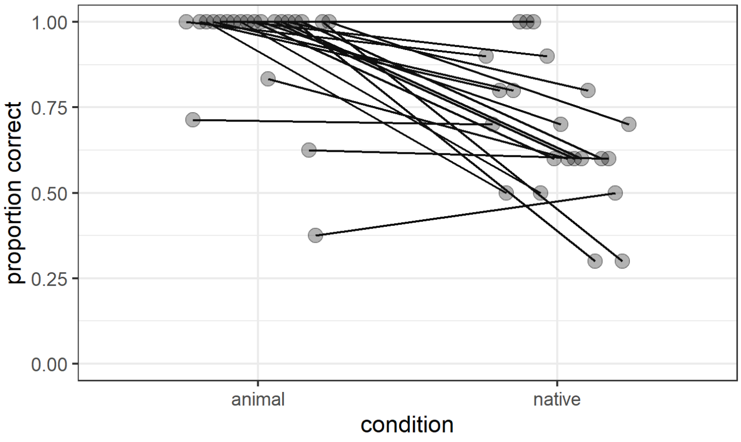


Figure 3.8

Relationship of individual performance of children in animal and native blocks

For the target trials, adults also have higher mean scores for the Hindi trials compared to the children (Figure 3.9), although pre and post training scores do not vary for both groups. The mean score for adults before (M = .71, SD = .46) and after training (M = 0.71, SD = .45) remains virtually identical, which is also the case for the children who scored lower than the adults both in the pre-training (M = 0.46, SD = .50) and post-training (M = 0.47 SD = .50). A one sample t-test of adults' accuracy showed performance to be above chance in both the pre-test (t (23) = 13.47, p < .001) and post-test (t (23) = 18.7, p < .001). Children, on the other hand, were performing at chance before (t (21) = -2.04, p = 0.05) and after training (t (21) = -1.18, p > 0.05).

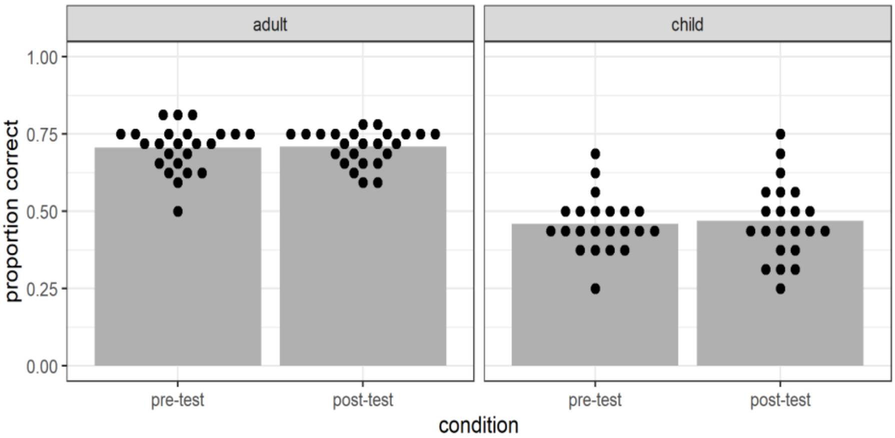


Figure 3.9

Proportion of correct responses by adults in the Hindi block during sound discrimination

In Figure 3.10, the performance of each participant in the pre-test and post-test were linked. The figure shows that some children exhibit larger increases in scores in the post-test within their group and compared to the adults but a decrease in scores is also common. Adults generally show less variability in individual performances compared to children with the majority exhibiting maintaining their scores before and after training.

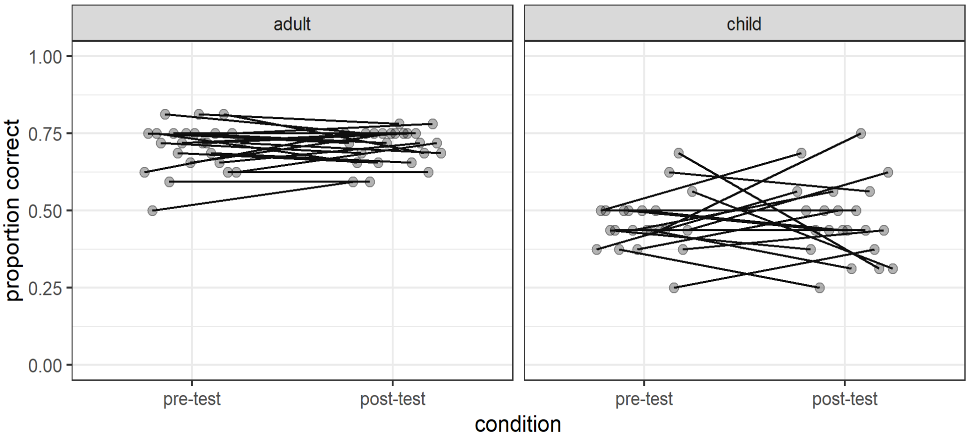


Figure 3.10

Proportion of correct responses by children in the Hindi block during sound discrimination

We fit a logistic mixed effects model for comparing the performance before and after training and between adults and children. Accuracy was coded as binary (0 = incorrect, 1 = correct). Treatment coding was used for the task (0=pre-test, 1=post-test) and group (0= adult, 1 = child) fixed effects. A maximal random effects structure with by participants and by items slopes was fitted, but this model failed to converge. This was then simplified to a model with by participants and by items intercepts. A model with a task by group interaction and one with no interaction were compared to the random effects only model. Neither the interaction (χ2 (1) = .002, p > .05) nor the fixed effects (χ2 (2) = .340, p > .05) were significant. Children showed a lower log odds of responding correctly than adults. The effect of training was not significant for both groups' performances on the sound discrimination task.

Word Recognition

Adults were also outperforming children in the word recognition task (Figure 3.11). Adults were at ceiling on non-minimal pair trials (M = 1, SD = 0) and performing above chance for minimal pair trials (M = 0.79, SD = 0.41) (t (23) = 6.58, p < 0.0001). Children's scores were also better for the non-minimal pair trials (M = 0.74, SD = .44) than their scores for the minimal pair trials (M = 0.49, SD = 0.50). One sample t-test scores for minimal pair trials show that children were performing at chance (t (21) = .23, p > 0.05) unlike adults who scored above chance (t (23) = 6.58, p < .001).

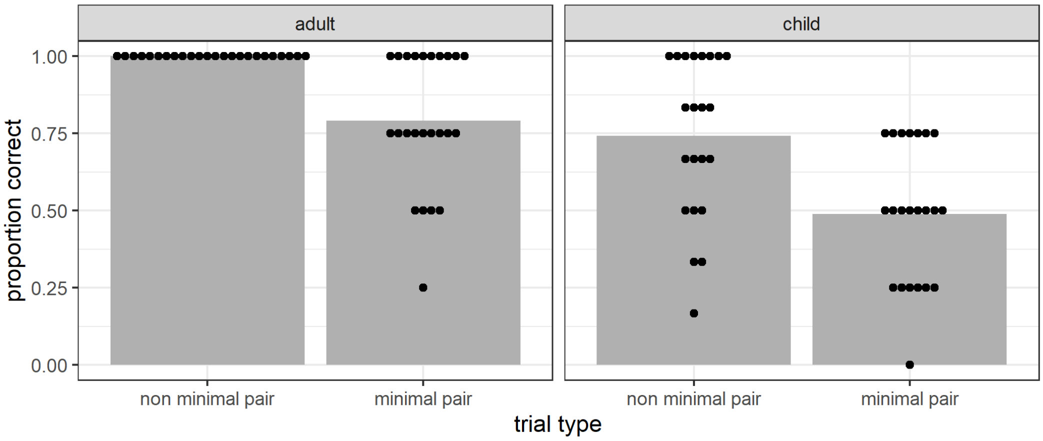


Figure 3.11

Proportion of correct responses on the word recognition task

We fit a logistic mixed effect model for the word recognition task with accuracy as the dependent variable and trial type (0 = non minimal pair, 1 = minimal pair) and group (0 = adult, 1 = child) as treatment coded fixed effects. Modeling the adult data with by participant and items slopes resulted in a singular fit and we simplified the random effects structure step-by-step and a by-participants only intercept was the best fit. We compared the random effects only model, a model containing an interaction between trial type and group, and a model with both fixed effects but without the interaction. The likelihood ratio test showed that interaction was significant (χ2 (5) = 15, p < .001). The child group had a significantly lower log odds of responding correctly compared to adults on minimal pair trials.

Correlation between post-training sound discrimination and word recognition performance

A similar exploratory analysis to experiment 1 was performed given the variability in sound performances in both groups. A Pearson's correlation on the proportion of correct responses participants obtained on the post-training sound discrimination task and word recognition task was run for the adults (Figure 3.12) and children (Figure 3.13). Neither performance from adults (r (22) = -0.28, p > 0.05) nor children (r (20) = -0.11, p > 0.05) showed any correlation.

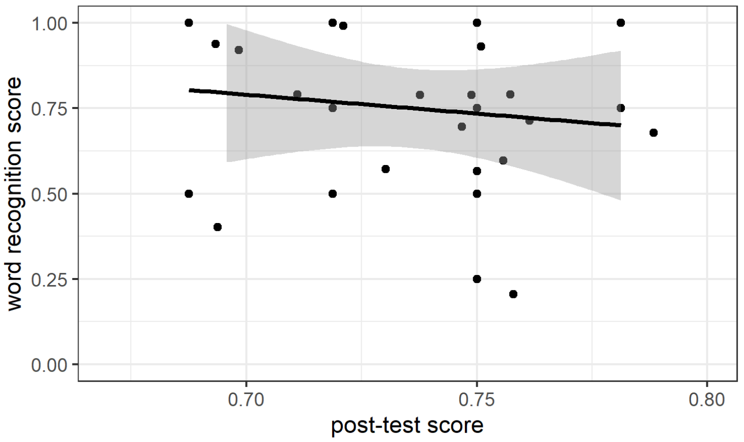


Figure 3.12

Correlation between proportion of correct responses of adults on the post-training measures (sound discrimination and word recognition)

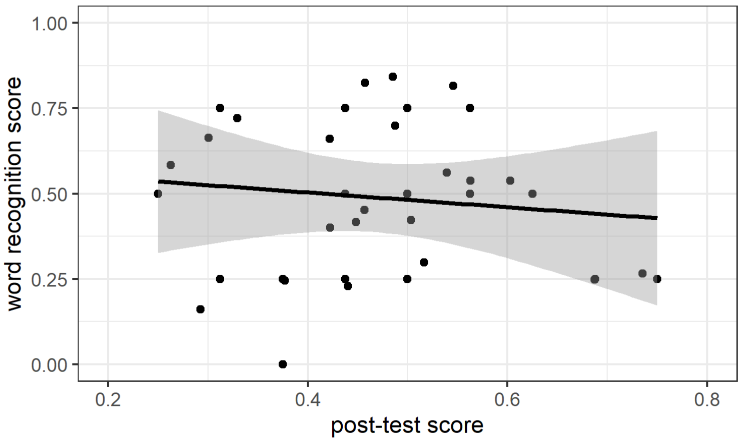


Figure 3.13

Correlation between proportion of correct responses of children on the post-training measures (sound discrimination and word recognition)

3.7.2 Looking Data

From 22 children in the sample, a total of 13 videos were coded for their looking behavior. There were four videos excluded due to a technical error. These were videos where the lighting was poor and so the gaze of the child was not clearly visible, setups where families were not using a front-facing web camera, and where the video recording was interrupted for more than five trials. Another five videos were excluded where children were not visible on the screen for five or more trials. The threshold of five trials was decided so that children would have data for approximately 70% of the trials, which was the value used for the attention checks in the first experiment.

For the sound discrimination task, the period between the offset of the second word to 2000ms after was analyzed. This was because the picture choices only became available after the second word, so looks would only become meaningful from this point. In the word recognition task, annotations began at the onset of the word to 2000ms post word onset as the choices were immediately available. The upper limit was chosen as looks after 2000ms were found to be no longer relevant to the processing of the auditory stimulus (Delle Luche et al., 2015). Looks were annotated using ELAN version 6.4 by two coders. Looks were coded as left or right, and off-screen looks were not coded. A second coder was recruited to test reliability of the coding criteria and annotated 10% of the total trials. We took the percentage of frames for which both raters were in agreement regarding the direction of the participant's looking direction and this totaled to 73.33%.

The mean proportion of looking time to target within a trial was used as a measure. This was calculated by first dividing the duration of an individual look to the target by the total duration of looks that were on the screen to both the target and distractor. Afterwards, the these looking durations to the target expressed as proportions were averaged.

Sound discrimination

The mean proportion of looking duration to the target before and after training is shown in Figure 3.14. There was an increase in the mean from the pre-test (M = .44, SD = .27) to the post-test (M = .57 SD = .31). The performance in the pre-test did not differ from chance (t (12) = -0.67, p > 0.05) nor did the performance in the post-test (t (12) = 1.09, p > 0.05). linear mixed effect model was fitted with the proportion of looking duration as the dependent variable predicted by task (treatment coded: 0 = pre-test, 1 = post-test). The random effects structure included by participant and by item intercepts as the more complex structure including slopes for each of the terms resulted in a singular fit. The increase seen in the post test did not reach significance (t = 1.46, p = 0.1). Although not significant, there is a trend towards an increase of looks to the target in the post-test that might reach significance with a larger sample size.

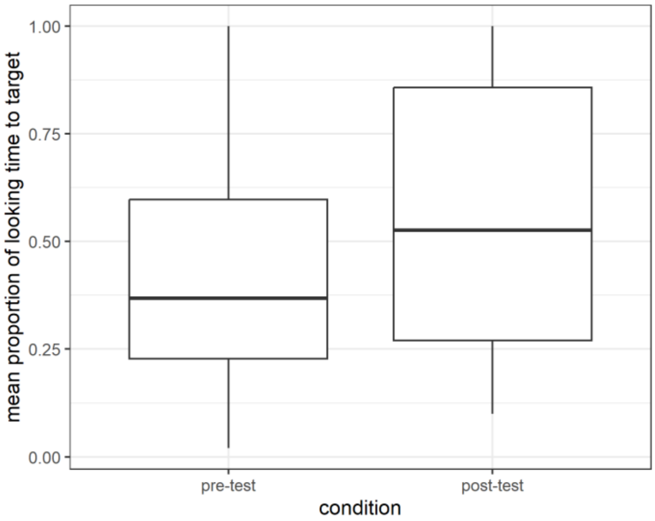


Figure 3.14

Comparison of mean proportion of looking time to target within a trial between pre and post -training sound discrimination tasks

Word recognition

The looking behavior of children in the word recognition task was compared based on trial types (Figure 3.15). The mean proportion of looking duration to the target was on average longer for the non-minimal pair trials (M = 0.42, SD = .29) compared to the minimal pair trials (M = 0.38, SD = .24). However, the performance for both trial types did not differ from chance (non-minimal pair trials: (t (12) = -0.72, p > 0.05); minimal pair trials: (t (12) = -2.40, p > 0.05)). A linear mixed model was fitted to model the proportion of looking duration as a function of trial type. A random effects structure with participants intercepts was used as a more complex model with slopes did not converge. The effect of trial type was not significant (t = -1.16, p > .05) meaning that the participants were performing equally for both trial types.

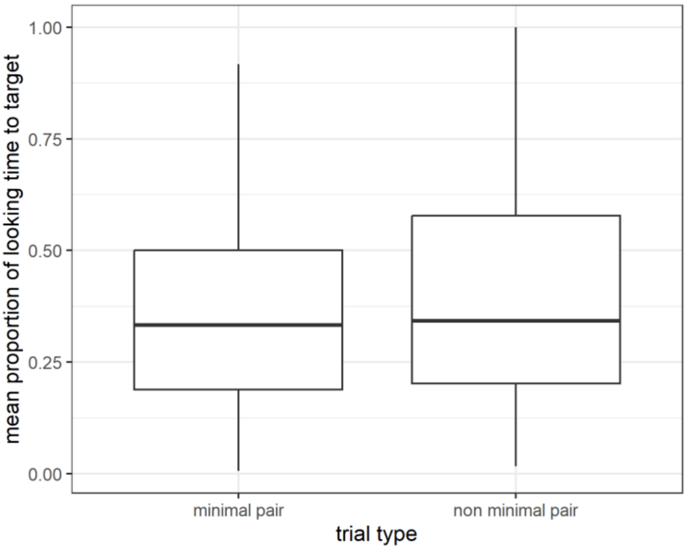


Figure 3.15

Comparison of proportion of looking time to target within a trial between word recognition trial types

The results from the looking data appear to contradict children's performance when providing an explicit response as they were responding above chance to the non-minimal pair trials (t = 4.36, p < .001). However, this analysis does not show where children are looking for the remainder of the time during the trial. Figure 3.16 shows that the mean proportion of time looking at the distractor varies between trial types. While children show no difference in looking behavior to the target, they looked longer to distractors in the minimal pair trials (M = 0.36, SD = .21) than the non-minimal pair trials (M = 0.30, SD = .19). Children were looking outside of the screen for the remainder of the trial in non-minimal pair trials, which does not necessarily contradict their above chance performance when asked to respond. On the other hand, they were looking towards the target and distractor for roughly the same duration in the minimal pair trials which is also consistent with behavioral data. It is possible that since they've already found the correct object faster in the non-minimal pair trials that children looked away from the screen for majority of the trial, whereas in the minimal pair trials they were looking at either the target or distractor as they could not immediately make a choice.

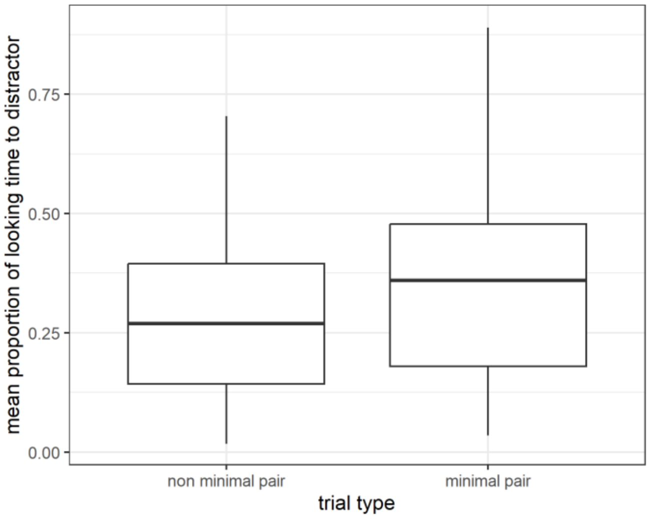


Figure 3.16

Comparison of proportion of looking time to distractor within a trial between word recognition trial types

3.8 Discussions

Non-native speech perception, compared to other linguistic areas, has shown stronger ties to age with children generally outperforming adults in long-term outcomes. This is attributed to the increase in sensitivity to L1 phoneme categories which is accompanied by a decrease in sensitivity to sounds beyond these. In general, children achieve better outcomes than adults in speech perception and production in naturalistic L2 learning contexts (Aoyama et al, 2003; Piske et al., 2001; Flege et al., 1999; Pallier et al., 1997; Snow & Hoefnagel-Höhle, 1978). Due to their younger age, they are seen as less susceptible to the L1 filtering out relevant acoustic properties relevant in discriminating a non-native contrast. To further clarify the developmental time course of non-native contrasts, we trained a group of adults and 4-year-olds to discriminate a non-native contrast through cross-situational word learning. Embedding the contrast in words paired with unique referents would enhance the difference in meaning each of the sounds create, while the ambiguity of CSWL would reflect naturalistic word learning situations. This would be that novel sounds can be encountered in words where the referent is not available or not yet learned. It was hypothesized that children, who are not as heavily influenced by their L1, would outperform adults on both measures of pure discrimination and on a word recognition task.

Our findings show that this was not completely the case. Although both adults and children show no difference in performance before and after training, adults generally score higher than children in both the sound discrimination task and word recognition task. The hypothesis that children would perform better due to being less susceptible to the influences of their L1 is based on findings on L2 learners acquiring the language in naturalistic settings (Oyama, 1976; Baker et al., 2002). On the other hand, there have been lab-based studies with both adults and children who have no prior experience with the contrast being trained that found better initial performance in adults over children (Fuhrmeister et al., 2020; Heeren & Schoten, 2010). However, this early advantage of adults disappears after multiple sessions of training where the younger participants either showed the same pace of learning or even outperformed them on some measures. This is possibly due to children requiring more time to consolidate learning across multiple sessions for them to eventually show superior learning to adults as was found by other studies. Overall, studies like those of Fuhrmeister et al., 2020 which was the basis of this chapter still concluded that better long-term outcomes for children were to be expected.

However, this study is unable to test if better performance from children does emerge across time as it did not measure performance after a prolonged period of delay. Similarly, it cannot make any conclusions on how lasting the effects of such training are. There is potential in internet-based studies facilitating more longitudinal investigations as parents might find logging onto a link more convenient than travelling to the lab over several sessions.

What remains consistent with the findings of the previous chapter is that adults' discrimination abilities did not improve after undergoing CSWL-based training. A notable difference is, however, that adults were scoring above chance on the discrimination task regardless of whether it was before or after training. This indicates that they were able to perceive differences in this version of the AX task.

It could be argued that this version of the sound discrimination task did provide additional support as it made use of pictures to represent whether the two sounds were different or the same and an accompanying familiarization phase of two trials at the start of the task. However, the association between the images and sounds in the discrimination task is different from the association tested in the word recognition task. The participants did not need to remember which alien corresponded with either the dental or retroflex. The images only served as a visual representation of whether what participants heard were the same or different. In the familiarization phase for the AX task, the words were also not used as names for either of the aliens but were introduced as something that they said. In the learning phase and word recognition task, the referential relationship between the words and the objects were explicitly stated. From what is known based on work with infants, this referential relationship was found to be key for the 9-month-olds in Yeung et al. (2014) to learn to discriminate between a non-native contrast. Simply pairing words containing a non-native contrast with unique objects was insufficient to emphasize the tone contrast being trained, it was only when familiarization trials where familiar objects were named that infants showed learning. While no similar test of the necessity of a referential relationship exists for adults, non-native contrast training studies with adults using minimal pairs were more effective than using non-minimal pairs (Davidson et al., 2007). This emphasizes the role of small phonetic differences and their effects on meaning to adults which aided them in learning a novel phonetic information beyond what is offered by the co-occurrence of linguistic and visual stimuli.

It is more likely that the difference in stimuli used in the AX task in the previous chapter and the one adapted for children is the types of stimuli used and the block presentations. The child-friendly AX task only made use of a trained pair and presented consecutively in a single block. In contrast, the earlier AX task interspersed presentations of the trained pair with the untrained pair and a native contrast pair. The inclusion of the native pair among the non-native stimuli could have made it more difficult for participants to overcome the influences of their L1 as they could still answer correctly on stimuli by relying on their L1 knowledge. Borrowing from the L2P2 model, for learners to be successful in discriminating a novel L2 contrast they must adjust the boundaries of the matching categories in their L1. This adjustment becomes difficult when participants are still encountering stimuli that can requires their L1 knowledge while simultaneously searching for the minute differences in the new L2 categories. Removing this dependence on the L1 listening mode in the child-friendly AX task possibly aided adults attune only to the relevant differences.

The simplified version of the AX task also appears to be effective considering ceiling performance of most children on discriminating animal sounds. However, the demands of discriminating speech sounds become clear even as children were asked to discriminate minimal pairs containing native contrasts. Children score lower on this block compared to the animal block. It appears that an additional demand is placed onto children to retain word forms containing fine acoustic differences which may have placed a demand on working memory. This is similar to findings of the previous study where working memory demands vary depending on the structure of the discrimination task.

Considering both the behavioral and looking data, it appears that children did not benefit from CSWL training for both measures of sound discrimination and word recognition. Children were able to comprehend the rules of the simplified AX procedure based on their ceiling performance in discriminating animal sounds. Their performance was lower in the succeeding block where they were asked to discriminate a native sound contrast. Despite effectively using this contrast in their L1, it appears that retaining small acoustic differences in an AX task increases the demand on children enough for their performance to drop below ceiling. Similar to the previous study, working memory demands might have contributed to children's poorer scores since they have to hold more detailed acoustic information to decide. While lab-based discrimination tasks with children do provide some insight on children's speech perception, it does not appear to perfectly map to how they functionally use this skill in naturalistic settings.

Children are also different from adults in that they do not benefit from the additional referential context provided by CSWL and the word recognition task. It is possible that this is caused by a shift in how children's word recognition develops from infancy to early childhood. There is evidence that children's lexical representations and processing in infancy are more holistic and underspecified but gradually changes to become more segmental in detailed in early childhood (Walley, 1993). This is driven by the increase in vocabulary which results in increased phonological competition during word recognition, hence justifying the need for more detailed lexical representations. If children were focusing on encoding these fine segmental differences in our study, these would have been masked interference from their L1. Unable to perceive these differences at the phonetic level, this then led to them also being unable to use their knowledge of the contrast to map word-referent relationships resulting in poor word recognition scores.

Our group of 4-year-olds is younger than the groups of children in other non-native speech perception training studies that tested children ranging from 6- to 16-year-olds (Fuhrmeister et al., 2020; Wang & Kuhl, 2003; Heeren & Schoten, 2010). According to the PAM (Best, 1991; Best et al., 2001) and NLM-e (Kuhl et al., 2008), our group would be at an advantage due to less interference from and less commitment of phonetic categories to the L1 expected in younger listeners. The difference might have resulted from the ambiguity inherent to the CSWL paradigm. Wang & Kuhl (2003) associated Mandarin tones to different animals and Fuhrmeister et al. (2020) associated novel words with unique images as we did, but these relationships were explicit and in the latter case reinforced during training. On the other hand, these associations were ambiguous in CSWL and participants were not given any feedback on whether they had already found the correct word-referent association. However, learning cross-situationally was not the problem in itself, as children showed learning for the non-minimal pair words. It is the additional demand of encoding detailed phonetic information containing a contrast that they could not perceive that had hindered learning minimal pair words.

There are findings in early development that children at 14 months old fail to use an L1 contrast to learn minimal pair words even if they can perceptually differentiate these sounds (Stager & Werker, 1997). However, infants at 20 months old have already overcome this limitation (Swingley & Aslin, 2000). This change is said to be driven by the increase in vocabulary which provides infants more experience with phonemes in their language appearing in distinct lexical contexts (Thiessen, 2007). While a distinct referential context is helpful for infants in detecting phonemic differences (Yeung et al., 2014), our use of minimal pair words did not provide a distinct lexical context. The children in this study have not been exposed to input where each member of the contrast appears in distinct lexical distributions. It might be that distinct lexical contexts in addition to referential contexts are necessary for children to begin discriminating a non-native contrast in words. While minimal pairs have shown to be more effective in emphasizing novel phonetic information to adults (Davidson et al., 2007), it might be that children would benefit more from non-minimal pair word stimuli.

Comparing the effect of distinct lexical contexts to minimal pairs in a CSWL-based training would be an interesting comparison to see if one is more facilitative than the other in emphasizing the differences created by phonetic contrasts. The findings of Thiessen (2007) show that distinct lexical contexts provide distributional information about the contrasts occurring in different phonetic environments. On the other hand, words that phonologically overlap in all but one phone force children to encode them more precisely (Storkel, 2002). Most lab-based studies that teach a non-native contrast use minimal pairs as it controls the lexical context in which the sound occurs. However, there has not been a direct comparison between the effectiveness of distinct lexical contexts compared to minimal pairs in training non-native speech perception abilities.

If this trend were considered, there is also a discrepancy in that this emerging increase was only observable in the looking data and not the children's pointing responses. The looking data presents children's response to the contrast immediately after hearing both stimuli, without the delay contained in their pointing response. It is possible that the auditory information had already decayed to some degree by the time children pointed to an answer and this would reflect why the improvement in performance is only reflected in the looking data.

In conclusion, this study found that adults outperform children on measures of sound discrimination and word recognition after CSWL training. While adults still did not display improvements in their discrimination abilities, it was found that they were performing above chance both before and after training in the sound discrimination task. The adapted version of the AX task which reduced the types of stimuli presented might have aided adults in focusing on a key difference in a limited context. Children, on the other hand, might be focusing on mapping the differences at the acoustic and phonetic level. This might be a result of how children approach lexical representations in early childhood with a focus on specifying lexical representations in detail to differentiate forms in their growing vocabulary. However, the influence of their L1 masks the acoustic differences of the contrast and therefore prevents them from properly differentiating the word forms and mapping the word-object relationships. These results along with those of the first study show that both adults are capable of simultaneously learning other types of linguistic information while learning word-referent mappings. Children are capable of learning words in CSWL but this ability is directly affected by the presence of a hard to perceive contrast in the L2.

Chapter 4. The influence of eye gaze and trial spacing on the proposed mechanisms underlying CSWL

The first two experiments have shown that novel phonetic information can be learned simultaneously while adults are tracking statistical information on word-object mappings. While not all novel words contain novel phonetic information, paying close attention to other co-occurring cues present during word learning could still be beneficial for the learner. An example would be social cues which are frequent in conversational speech and have a strong relationship to the communicative intent of a speaker. While studies have investigated the ability of cross-situational statistics and social cues to individually support the word learning process, less is known about how these two streams of information interact with each other.

Since the suggestion that learners keep track of multiple word learning cues has been made, the question of how and how much information they absorb from the rich learning environment becomes more relevant. The exact mechanism underlying CSWL is still under discussion with the degree of ambiguity, competition among referents, and delays between learning instances influencing findings (Bunce et al., 2016; Vouloumanos, 2008; Suanda & Namy, 2012; Smith et al, 2011; Aussems & Vogt, 2020). Social cues do signal the intent of the speaker and could reduce ambiguity, which would in turn affect the strategy undertaken by learners in considering co-occurrence frequencies. This experiment investigates the role of social cues in varying trial spacing conditions and examines the mechanisms behind the aggregation of co-occurrence frequencies.

4.1 Mechanisms underlying cross-situational word learning

There are two main accounts that attempt to explain how cross-situational statistics are accrued across multiple learning trials. The associative learning (AL) account suggests that learners keep track of co-occurrence frequencies for multiple events simultaneously and refer to this stored information in future encounters (Yu & Smith, 2007). Figure 4.1 shows how a learner performing associative learning would approach a CSWL task. Upon hearing a word in trial 1, the learner would then take note of all referents available in the learning situation. On the next trial, they update the information from the previous trial with the referents that co-occur with the word in this encounter. With the word already occurring twice with one of the objects, this association now becomes stronger compared to the others. The learner continuously updates their matrix with additional learning trials until one of the potential referents sufficiently reaches a higher co-occurrence frequency with the target word compared to the others. Trial 3 shows the process begins again when a novel word is heard where the learner tracks a new set of co-occurrence frequencies for the set of objects present in this encounter.

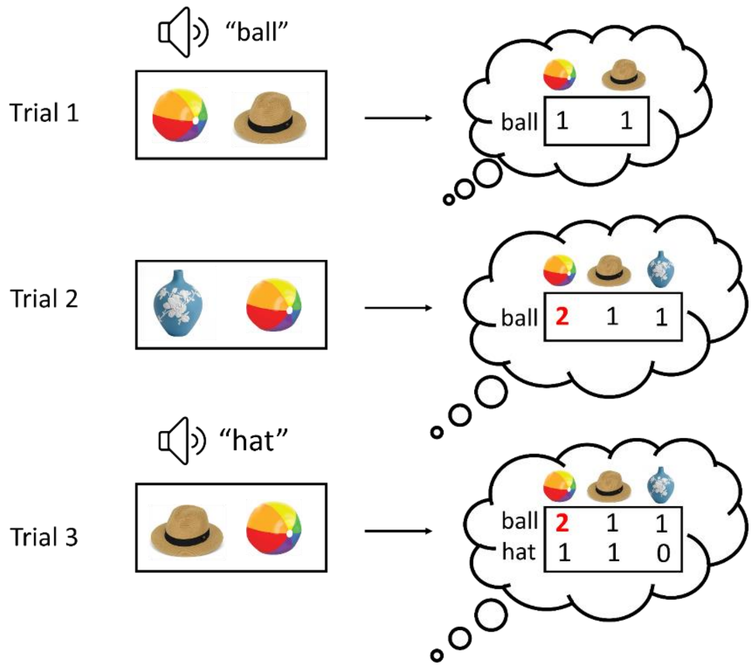


Figure 4.1

CSWL performed according to the Associative Learning account

In contrast to this is the hypothesis testing (HT) account (Figure 4.2). A learner performing HT would be more selective in the information they choose to store (Gleitman et al., 2005). In their initial encounter with a word and possible referents, the learner would select one of the referents at random to test in the next learning instance. In this example, the learner receives information in trial 2 that the referent they had selected prior is not among the current available referents. This leads the hypothesized referent to be rejected. The learner then proceeds to select another referent to confirm in the next encounter. Important to note here is that the learner lacks any information about which referents they had encountered before and thus cannot use this to choose a new hypothesis. In trial 2, the learner selects the correct hypothesis but this is without knowledge that they have encountered this object twice already with the same label. In trial 3, the learner receives feedback that their hypothesized referent is correct and will continue to use this referent as a hypothesis in trial 4. The learner will continue carrying on with a hypothesized referent as long as it is confirmed as correct.

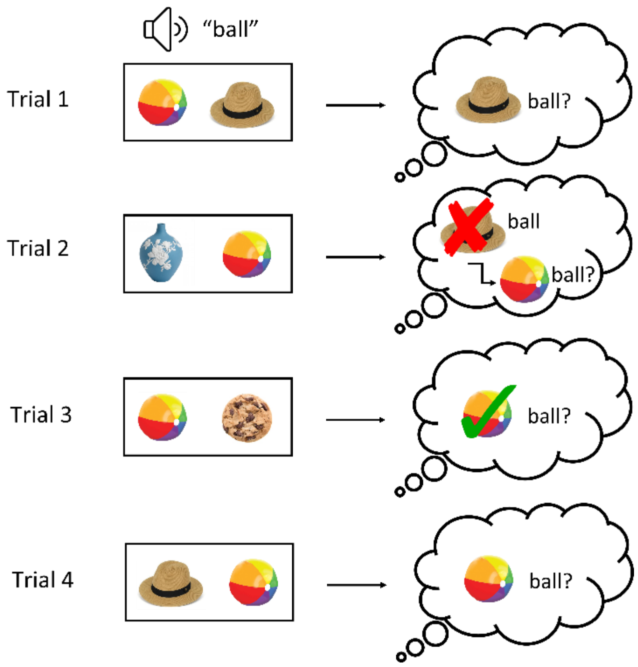


Figure 4.2

CSWL performed according to the Hypothesis Testing account

For both the AL and HT accounts, their descriptions have yet to offer an endpoint of when the correct word-referent mapping is established. The associative learner would need to consider what degree of co-occurrence frequency is sufficiently high to distinguish a referent from other spurious co-occurrences. This then leads to the question of how many possible mappings does the learner take into consideration when deciding the pairing with highest co-occurrence frequency. The number of candidates can grow quite large especially if the word being learned is one that can occur in multiple different contexts, with each unique context containing a new set of distractors. For example, hearing the word "water" during mealtime would expose the child to a different set of co-occurring objects as compared to hearing the word during bath time. The AL account could describe which candidates are brought into final consideration to determine the correct word-referent mapping.

The learner employing hypothesis testing would need to determine how many confirmations a hypothesis needs in order to establish it as correct. It might also be important to determine if consecutive confirmations are required, as the pure HT account states that an incorrect hypothesis is discarded immediately. It can happen that the correct word-object pairing is carried over as a hypothesis in the next learning situation only for it to be incorrectly discarded. A speaker can mention a word, but the correct object is absent in the environment as one can refer to concepts removed from the present situation. The HT account then predicts that the learner would discard this word-object mapping completely and would have to choose a new hypothesis again. The HT account would need to explain how the learner overcomes false negatives such as this and at what point is a hypothesis confirmed.

4.1 Evidence for and against the hypothesis testing account

Zhang et al. (2019) have written a review that outlines the supporting evidence for each approach. The main criticism against the AL account is that the attention and memory demands of keeping track of all referents present in the natural learning space would be too high. In response to this, the HT has been proposed. Medina et al. (2011) have attempted to simulate the experience of the naive word learner using CSWL in naturalistic settings through the Human Simulation Paradigm (Gillette et al., 1999). Adults were tasked to learn new word-referent mappings cross-situationally by watching 40-s videos of parents interacting with 15-20-month-old infants during naturalistic contexts (e.g., mealtime, play). In these vignettes, the audio was completely muted except for the part where the parent named an object where the label was replaced by a non-sense word. Participants were tasked to determine what the parent was naming across five different vignettes that referred to one object, providing opportunities to aggregate information across multiple encounters. To control for the ambiguity in each vignette, they manipulated a variable called "informativeness". The informativeness of a vignette was based on how well a separate group of participants were able to identify the referents. Vignettes with high informativeness were those where participants correctly identified the referent above chance. There was one vignette with high informativeness included out of the five shown for each object and findings showed that the order of this had a strong influence on the trial-by-trial accuracies of participants. The earlier participants encounter this trial, the earlier the sharp rise in their accuracy would appear across the five learning instances. This pattern is predicted by the HT account since once the learner gets confirmation of their chosen referent, they are to continue proposing this referent until the end of the learning phase. In contrast, a learner performing AL would display a gradual rise in trial-by-trial accuracy as it would take them longer to build up sufficient evidence for the correct referent even if a more informative exposure appeared early on.

This sharp rise in participants' accuracy following confirmation of their hypothesis found by Medina et al. (2011) was used by Trueswell et al. (2013) as a basis for analyzing the relationship of accuracy in the previous trial to that of the present trial in CSWL studies. They collected eye tracking as well as behavioral responses in a series of CSWL experiments. In every trial, participants saw one target and four distractors on the screen and guessed which object the word referred to after every learning trial. They found that participants were more likely to select the correct referent if they had also done so in the preceding trial, but that they remained at chance if they had guessed incorrectly before. The looks to the target compared to the distractor also followed the same pattern based on the previous trial's accuracy. This is in line with the HT account since participants will continue proposing the same referent if they receive confirmation of their hypothesis. If this referent happens to be the correct one, then their accuracy in future encounters would be dependent on their selection before.

Other CSWL studies do not often make use of previous trial accuracy to determine the strategy used by participants, although no studies have directly criticized this approach either. This can make it difficult to directly compare findings as other studies use different measures to analyze the underlying mechanism of CSWL. On the other hand, when Zhang et al. (2021) used the Human Simulation Paradigm used by Medina et al (2011), they found that learners were continuously integrating information during their learning rather than show a sharp rise in increase. However, they did not mention whether their findings were in support of either account but rather refer to other studies that propose an integration of the two. There have so far been fewer studies that support a pure HT account of CSWL as compared to those that support AL and varieties of it.

A common issue pointed out regarding the studies above is that they instruct participants to select or guess a referent after each trial, which arguably encourages them to make an explicit hypothesis during the learning phase. This could then influence the way learners aggregate co-occurrence frequencies to conform to the HT account. Direct comparisons of a learning phase requiring trial-by-trial responses versus passive watching found no difference in participants' accuracy or reaction times on test trials (Kachergis & Yu, 2014; Zhang et al., 2021). It can be argued that measures such as accuracy and reaction time only reflect the end state of learning, while the concern is that providing an explicit response every trial affects the process of learning itself.

4.2 Evidence for and against the associative learning account

Other studies investigate the underlying mechanism of CSWL through measuring participants' memory for any alternative referents during learning. The AL account claims that since learners are keeping track of multiple possible referents, then they should remember other objects besides the target on the screen. Another variable controlled by CSWL studies is the presence of an object that has a high co-occurrence frequency with the novel world. In real life situations, there will be instances where a learner hears "shoe" but also sees a sock among the objects in the scene despite it not being the correct referent. Learners would then have to overcome the additional competition from some distractors like these.

CSWL studies found that participants are more likely to select the referent with a higher co-occurrence frequency when the object with perfect co-occurrence with the word is absent (Bunce et al., 2016; Vouloumanos, 2008; Suanda & Namy, 2012). Bunce et al. (2016) performed a mouse tracking CSWL study where they kept track of participants' mouse trajectories when selecting a referent on the screen. Participants underwent a training phase and were asked to select a referent on each trial. Of the several objects on the screen, there was one object that occurred in 50% of the training trials for a specific word and was considered a high probability competitor. This competitor occurred less frequently than the target, but more frequently than the other distractors. Test trials were divided into those that showed this high probability competitor alongside the target and those that did not. While participants were equally above chance in accuracy for both kinds of test trials, their mouse movements showed less deviation in the trials where the competitor was absent. Differing mouse trajectories reveals some influence of the more frequently appearing distractor on the participants' decision-making process.

This assumption that learners keep track of an indiscriminate number of possible word-object mappings is the main criticism against the pure AL approach. An approach this exhaustive would surely not be possible given a study on 9–12-month-old infants show that their working memory for extends up to three objects and declines as they need to remember more specific features (Kibbe & Leslie, 2013). In CSWL, the higher the number of potential referents, the higher the referential ambiguity. Consequently, the success at word learning deteriorates with higher referential ambiguity (Smith et al., 2011; Medina et al., 2011; Yu & Smith, 2012). While the purest version of AL proposes that learners keep track of all possible referents, Koehne and colleagues (2013) found that learners only keep track of referents that they have considered before. Their CSWL experiment also required trial-by-trial responses as well as a distractor that appeared at a higher frequency with the word (50%) compared to the other distractors (17%). Participants would only select a high frequency distractor above chance level if they had selected it before during learning. They replicated these results also using a passive paradigm where they used eye tracking to record participants' eye movements during learning. If participants looked at the high frequency distractor more than 50% of the time during any of the learning trials, they were also above chance at selecting it during the test phase. Thus, learners do not keep a perfect record of co-occurrence frequencies and which referents they store is restricted by their attention. However, they do use the information they have stored regarding multiple referents when making decisions in future encounters.

4.3 Delays between learning trials in CSWL

The CSWL studies discussed so far manipulated the difficulty of the learning situation through the number and co-occurrence frequencies present. The difficulty of the learning situation could also be affected by delays between learning situations, which affect how participants aggregate cross-situational statistics. Ideally a learner would be able to receive multiple encounters with the word to confirm its meaning, but these learning opportunities could be separated by a period of time. Learners would then have to retrieve previous information after a delay whether its co-occurrence frequencies of multiple referents or a single proposed referent waiting to be verified.

Delay is introduced into CSWL experiments through manipulating trial spacing which compares consecutive versus interleaved orders of presentation. Consecutive presentations would have all of the trials related to a single word presented in immediate succession (Figure 4.3). Interleaved spacing, on the other hand, would insert other learning trials in between before the next encounter with a word (Figure 4.4). Interleaved spacing places demands on learners in terms of retention and retrieval which specifically could influence the mechanisms of CSWL. The delay would make it more challenging for learners to maintain multiple associations during learning and to retrieve them if the interval between exposures is too long. This could lead to learners keeping track of less referents or even a single referent.

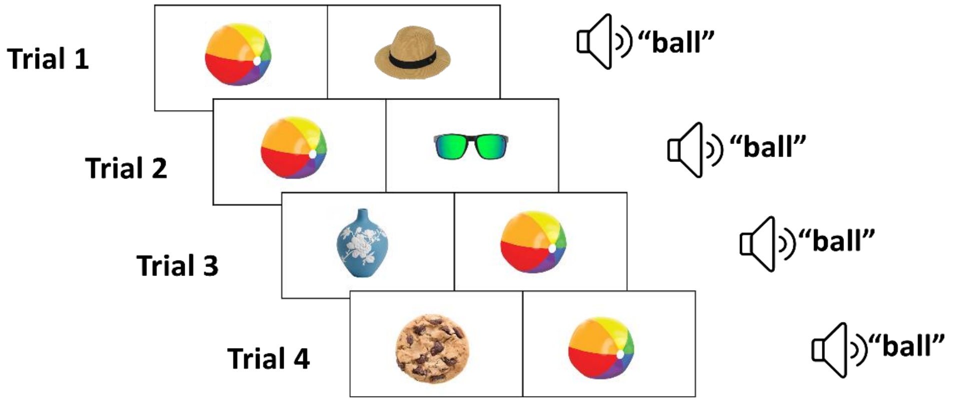


Figure 4.3

CSWL learning trials presented in a consecutive order

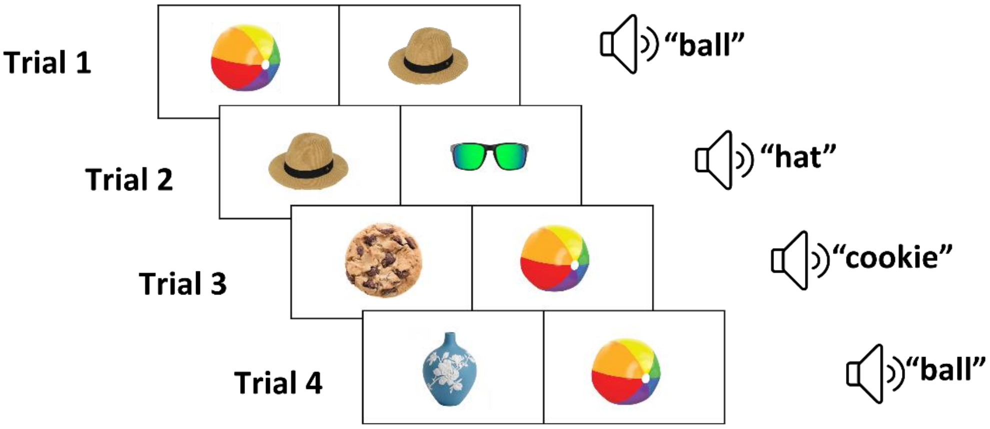


Figure 4.4

CSWL learning trials presented in an interleaved order

Smith et al. (2011) found support for this claim when they manipulated referential uncertainty as well as trial spacing during the learning phase in a CSWL experiment. Their experiment followed a 3 x 2 experimental design where there were 3 levels of referential uncertainty (2, 5, or 8 referents onscreen) and 2 levels of trial spacing (consecutive or interleaved). In the test phase, participants were asked to choose the target among all 15 referents that were shown during training with the word. The authors applied an algorithm to determine if participants were adhering to a specific strategy or were randomly responding in the test phase. The results showed that participants were more likely to keep track of the entire set of possible referents in the consecutive condition but were more likely to choose among possible referents with higher co-occurrence frequencies in the interleaved condition.

Aussems and Vogt (2020) also tested the effect of trial spacing but implemented a passive, vision-based paradigm with eye-tracking instead of a forced-choice task. They used the same algorithm as Smith et al. (2011), but this time to explicitly compare the AL and HT accounts. They found that no participants used a pure HT approach in either spacing condition. Their looking behavior fell into a "conservative CSWL" model wherein their learners did not propose a referent until the later learning trials and were aggregating information for the initial trials. They also found evidence of a mixture of strategies within spacing conditions and also within participants. In either case, interleaved presentations did lead to less referents being tracked but did not result to participants only testing a single referent at a time.

The findings of both studies suggest that learners tend to favor an AL strategy and shifted to recalling less objects if either: 1) they have encountered several learning trials that provides them some information to make a hypothesis or 2) they encounter an additional task demand, such as interleaved trial spacing, that makes remembering more objects more demanding. While neither of the studies found evidence for a pure HT approach, it does appear that interleaved presentations encourage a shift of strategy from AL to HT.

4.4 Social cues and CSWL

CSWL studies have extensively taken into account how the complexity of the naturalistic learning context burdens the learner. Current findings show that increasing degrees of referential ambiguity or delays between learning trials have not been sufficient for participants to adopt a pure HT strategy. However, the results also do not favor the image of a perfect associative learner but one who has distributed attentional resources to multiple selected referents. Additionally, learners who begin to keep track of multiple referents can eventually shift to proposing a single referent when they've gathered enough statistics. This is in line with the end goal of the word learner which is to find the referent with sufficient evidence among a plethora of options.

On the other hand, the richness of the naturalistic learning context can also provide the learners multiple cues to support their word learning. There is evidence that learners would eventually propose a single referent if they have amassed sufficient co-occurrence frequencies but providing them additional supporting information could also encourage them to settle on a single referent earlier on. Co-occurrence frequencies, sufficient as they are in the lab, are not the only way to determine referents in the real world. Learners encounter different types of information in the language input that could bolster word learning. Going back to findings of Medina et al. (2011), their participants showed behavior consistent with the HT approach when asked to guess the name of the object being named in several vignettes of parent-child interactions. The videos themselves contained additional cues in the communicative interactions that would remain despite removing the accompanying linguistic information. This could be social cues which could reduce the ambiguity within a learning situation to help learners propose a hypothesis sooner.

Studies have shown that co-occurrence frequencies can simultaneously be tracked alongside other cues that support word learning (Monaghan et al., 2017; Dautriche & Chemla, 2014). Monaghan et al. (2017) tested the effect of distributional, prosodic, and gestural cues appearing at varying frequencies during CSWL. Distributional cues were an additional non-word that appeared before the word that referred to the object, similar to how an article would accompany a noun. A prosodic cue was emphasis placed on the referring word which would otherwise be monotonous. The gestural cue was an image of a hand pointing to the object that matched the referring word. These cues occurred at frequencies 25%, 50%, 75% and 100% of the time in learning trials. The test trials were shown without any cues. Results showed that accuracy increased along with the number of types of cues present, but the gestural cue produced significantly better performance compared to the other two cues. Although performance also increased with the frequency of cues, it was the 75% condition that produced the best outcomes even compared to the condition where additional cues were present 100% of the time. The authors provided two possible explanations for this. One comes from the degeneracy model which states that a 100% reliable cue is less helpful to a learner as it will become a crutch (Monaghan, 2017). It will make it difficult for the learner to attend to other highly correlated cues in situations where the perfectly reliable cue is absent. It appears that learners can tolerate and do benefit from imperfect cues as long as there are other overlapping cues present that would help them fill the gap.

Social cues, in particular, have been shown to be widely used by novice learners to learn words in early development (for review of evidence see Baldwin & Moses, 2001) and are central to the social-pragmatic theory of word learning (Tomasello, 2000). Formal simulations show that cross-situational information paired with social cues improve the learning outcomes (Yu and Ballard, 2007; Frank et al., 2009; Lazaridou et al, 2016).

MacDonald et al. (2017) tested the mechanisms of CSWL while including an eye gaze cue directed towards the target during learning. Participants were asked to select the object which they think the word referred to during the learning phase and the majority of them selected the one which was accompanied by eye gaze. In order to test if participants' memory for alternative referents was affected by the presence of a social cue, the authors made use of switch trials in the test phase. Switch trials are named so because the referent that participants chose in the learning phase is "switched out" and another object is put in its place. This alternative object has co-occurred with the word in the learning phase but was not previously selected by the participant. This alternative object would be considered as the correct answer in switch trials, as the remaining options have not co-occurred with the word at all. If participants were performing AL, then they would choose the alternative referent as it still has a higher co-occurrence than the remaining choices. Participants who were assigned to the gaze condition failed to correctly select this object, which indicates that they were not tracking referents apart from the one which was the object of a social cue. Notably, the presence of eye gaze did not result in a significant boost in performance in the "same" test trials where the objects that participants originally selected were present. Due to the lack of better performance in the same trials with gaze, what seems to be primarily affected by the social cue is the memory for number of alternative referents rather than the strength of a chosen hypothesis.

MacDonald et al.'s (2017) findings suggest that eye gaze can encourage learners to track a single hypothesis during learning supporting the HT account. However, their study exposed learners to only one learning trial for a word. The next exposure they receive would be a test trial. This means that learners had limited opportunities to amass co-occurrence statistics before they are tested and had the additional expectation of only selecting one referent. The limited number of learning trials would not be conducive to an AL strategy, which could explain why behavior consistent with HT was observed. Increasing the number of learning trials before testing would provide more opportunities to see when and to what extent participants begin relying on the social cue. It would test if providing a reliable cue for reference encourages learners to rely on the social cue as a heuristic instead of gradually building up evidence.

Another interesting comparison is between the influence of task demands and supportive cues on the learner's choice of strategies. Generally, learners tend to be more cautious and favor recalling multiple referents, but this is affected by trial spacing. Previous findings have already shown that participants would opt to track multiple referents in a consecutive spacing condition as compared to interleaved presentations. Would this still be the case if a reliable social cue were to reduce referential uncertainty to the point that being cautious presents no additional benefit? Alternatively, learners could also give equal importance to both and rely on a strategy based on a tradeoff between cognitive demands and the presence of supportive cues.

4.5 The present study

In the current study, the influence of trial spacing and eye gaze on the underlying mechanisms of CSWL was investigated using webcam-based eye tracking in an internet-based experiment. This study combines the use of high frequency distractors (HFD) which were used in other CSWL studies (Bunce et al., 2016; Vouloumanos, 2008; Suanda & Namy, 2012; Koehne et al., 2013) and switch trials as used in MacDonald et al. (2017) to determine memory for alternative referents during learning. Whether or not participants are sensitive to the co-occurrence frequencies of other objects beyond the target was used to determine whether participants were engaging in AL or HT during learning. This was measured through their looking time to the HFD.

The target in CSWL is usually the word that co-occurs 100% with the novel word and distractors usually have a similar co-occurrence frequency that is below this. A high frequency distractor is an object that has a higher co-occurrence frequency than the other distractors but still less than the target and its main function is to provide additional competition to the target. In this study, the co-occurrence frequency of the HFD is at 50% based on previous findings that participants were able to consider it as an alternative referent at this rate (Bunce et al., 2016; Vouloumanos, 2008; Suanda & Namy, 2012). If participants are engaging in HT, then they would not present a difference in looks to the HFD and distractors because they are only keeping track of the target. This is because the target is the only object that would not be rejected as a hypothesis since it has a 100% co-occurrence frequency. If participants are engaging in AL, then they would show a difference in pattern of looks to the HFD versus regular distractors since it has a higher probability of being the correct referent. An associative learner would be able to recall as the goal of AL is to aggregate information for all possible referents, not just the correct one.

The role of the HFD becomes important during the switch trials, which are only present during the test phase. The switch trials replaced the target object with the HFD with the goal of determining if participants were recalling the co-occurrence frequencies of any object aside from the the target. If participants were engaging in HT, then there would be no difference in looks between the HFD and distractors. However, an associative learner would be able to present more looks to the HFD since it is the next best option in the absence of the target. Previous studies that found evidence favoring the AL approach found that participants would choose the HFD if the target was absent (Bunce et al., 2016; Vouloumanos, 2008; Suanda & Namy, 2012).

Separate research questions and hypotheses were put forth for trial spacing and social cue, as well as regarding the interaction between the two. These are summarized in Table 4.1. The first question was specific to the effects of trial spacing on the mechanism underlying CSWL. Based on the findings of Smith et al. (2011) and Aussems &Vogt (2020), consecutive spacing would allow participants to perform AL while interleaved spacing is expected to lead to behavior more consistent with HT. Phrasing these hypotheses in terms of the switch trial paradigm used in this study, participants are expected to present more looks to the HFD compared to the other distractors in the consecutive condition compared to the interleaved condition.

The findings of Smith et al. (2011) and Aussems & Vogt (2020) stated that their participants did not behave in a manner consistent with the pure HT approach, only that they were less effective in remembering alternative referents due to the memory demands of interleaved spacing. This was interpreted as the learning situation influencing participants to be more selective in their tracking of referents similar to what is suggested to HT. If participants were performing pure HT, this would be seen as no difference in looking times between the HFD and distractors in the switch trials. This is because learners performing pure HT are only expected to keep the correct hypothesis, the target, and discard referents that are inconsistent with its occurrence with the label such as the HFD.

The individual effect of social cues on CSWL was examined in two ways. The first also relied on analyzing performance on the switch trials. Since social cues reduce ambiguity, it could encourage learners to keep track of a single referent supported by the eye gaze which would be more in line with a HT approach. This would lead to no difference in looks to the HFD and distractors in the switch trials for the with social cue condition. This is in line with what MacDonald et al. (2017) found.

The second way to test the effect of social cues is to look at trial-by-trial gaze data during the learning phase. Medina et al. (2011) and Trueswell et al. (2013) implemented paradigms where they asked participants to select a referent for each learning trial. They found that responding correctly on the previous trial increased the probability of being correct on subsequent trials as well manifesting as a sudden rise in accuracy. This is said to be due to learners implementing hypothesis testing where a correct hypothesis is carried forward onto the next encounter as long as it is proven to be correct. In this experiment, learners who find the correct target would be expected to show a sharp increase of looks to the target compared to the other distractors during the learning phase that would remain consistent across the subsequent trials. If participants were paying attention to the referential intent behind the eye gaze provided by the onscreen speaker, then this increase in looks to the target would appear only in the with social cue condition. Participants in the without social cue condition, however, are expected to show a steady increase in their looks to the target for each trial since they have to compile co-occurrence frequencies without the aid of a cue that reduces referential ambiguity.

Finally, this experiment also tested how trial spacing and social cues interacted with each other. There were three different possibilities. The first was that trial spacing exerts a stronger influence than the social cue. This means that regardless of social cue, participants are expected to perform AL in the consecutive spacing condition and a less efficient version of AL in the interleaved condition. If this were the case, then participants are expected to have greater proportion of looks to the HFD in the switch trials in the consecutive spacing condition regardless of whether a social cue was present or not. Participants would proceed to cautiously aggregate statistics despite a reliable social cue being present, resulting in behavior more aligned with AL. This is in line with what Smith et al. (2011) and Aussems &Vogt (2020) found that participants are by default tracking co-occurrence frequencies of objects.

The second scenario was that the social cue would be more influential than trial spacing. If this were the case, participants would perform HT regardless of spacing condition because a reliable cue is present. Even if the task demands of the consecutive spacing are low, they would use the social cue to identify the target than rely on a slower AL approach. This would present as no difference in looks HFD compared to other distractors on the switch trials regardless of spacing condition which is more in line with the HT approach.

The third possibility would be that trial spacing and social cue are equally important in the learner's choice of strategy. Smith et al. (2011) and Aussems &Vogt (2020) found that interleaved presentations alone were insufficient for learners to engage in pure HT. This study hypothesizes that the combination of task demands from interleaved spacing added to the reduced ambiguity provided by a social cue could be sufficient for participants to adopt this approach. This would mean that participants would not show any difference in looks to the HFD compared to distractors in the switch trials in a condition. Likewise, a sharp rise in looks to the target would also be expected for the interleaved with social cue condition in the learning phase.

Hypothesis	**Measure**	**Outcome**
Individual effect of trial spacing		
Consecutive spacing encourages AL; Interleaved spacing less so but not pure HT	Switch Trials	Difference in looks between the HFD and other distractors is larger in the consecutive vs interleaved spacing condition
Individual effect of social cue		
With social cue condition encourages HT	Switch Trials	No difference in looks between the HFD and other distractors in with social cue condition
	Learning Phase	Sudden increase in looks to the target vs distractors in the with social cue condition
Interaction between trial spacing and social cue		
Trial spacing is more influential than social cue and results in AL	Switch Trials	Regardless of social cue condition, difference in looks to the HFD and distractors
Trial spacing is less influential than social cue and results in HT	Switch Trials	Regardless of spacing condition, no difference in looks to the HFD and distractors
Both are equally important and HT is expected when situation is both too demanding to do AL and referential ambiguity is low	Switch Trials	No difference in looks to HFD and distractors only the interleaved spacing with social cue condition
	Learning Phase	Sudden increase in looks to the target vs distractors only the interleaved spacing with social cue condition

Table 4.1

Summary of hypotheses for each research question

4.6 Method

Participants

We recruited a total of 60 native British English speakers from Prolific (www.prolific.co) (35 female; mean age: 32.0 years). Participants had no self-reported history of neurologic injury, self-reported as neurotypical, and did not wear glasses or contact lenses. They also passed the hardware check implemented by the experiment platform that determines if their device could run the experiment. An equal number of participants were assigned to each of the trial spacing conditions.

Stimuli

There were 8 novel words used split evenly across the two social cue conditions. These words followed the phonotactic rules of English and were created using the software Wuggy (Brysbaert, 2010). The pseudowords were: ufface (/ʌfɪs/), ponem (/pɔnɘm/), cungle (/kʌngɘl/), bewy(/bjuwi/), marsy(/mɑrsi/), recubs(/rikʌbs/), sether(/seθər/), pimcent (/pɪmsent/). Two female native British English speakers recorded the words. There was a different speaker for each of the social cue conditions. The same speakers were also featured in the video recordings used to show an eye gaze on screen.

Images were taken from the set of novel unfamiliar objects created by Smith et al. (2011). There was a total of 96 distinct objects used. Each word had one object assigned as the target, another as the HFD, and 10 objects which served as distractors. A separate pool of distractors was used per word to prevent participants from using mutual exclusivity from learning trials for different words. The target had a 100% co-occurrence frequency (i.e.,, appears in all learning trials) with the label and will be the target of the eye gaze in the social cue condition. The HFD had 50% co-occurrence frequency with the label and appeared in half of the learning trials. This was less frequent than the target object, but more frequent than the remaining distractors. The remaining 10 other objects all had a co-occurrence frequency of 25%, less than the target and HFD.

A video recording of a person served as an onscreen speaker providing a social cue. The speaker who produced the words in the no social cues condition began the video looking straight ahead, said the word, and then continued to look ahead. There were four videos like this, one for each of the words presented without a social cue. The speaker assigned to the social cue condition also began the video looking ahead, produced the word, turned their head to one of the quadrants on the screen, and finally turned their head back to the center and continued looking ahead. As the target could appear in one of the four quadrants, each word had four different videos showing a matching social cue (Figure 4.5). There were 16 videos in total for the social cue condition. The videos were only used during the learning phase. Only audio recordings were used during the test phase. Since there were already eight objects on the screen in the test phase, including the speaker onscreen would further add to the visual stimuli. It was decided to remove this to concentrate participants' looking times to objects.

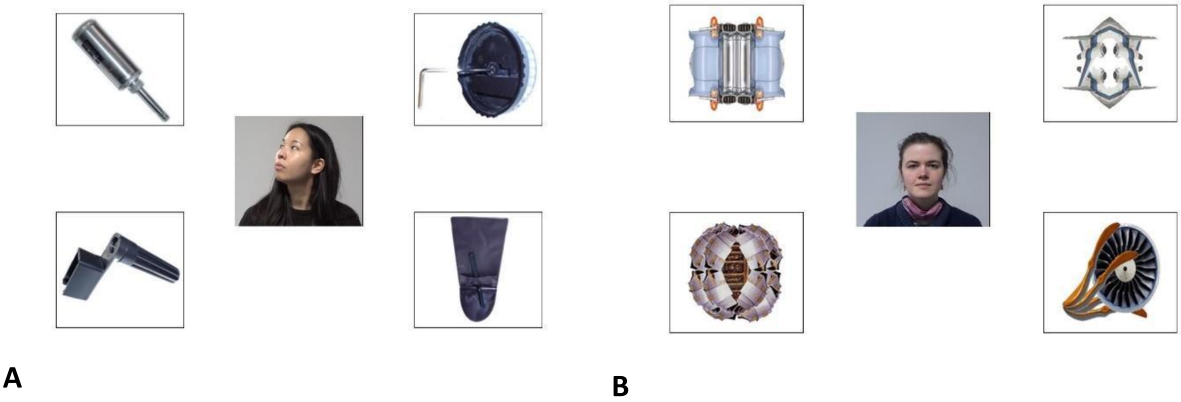


Figure 4.5

Learning phase trials showing objects presented with (A) and without a social cue (B)

Procedure

The experiment collected eye tracking data using users' webcams through Labvanced (www.labvanced.com). There were no other behavioral measures collected as not to interfere with the collection of eye tracking data.

Participants were randomly assigned to one of the spacing conditions: consecutive or interleaved. In each trial spacing condition, participants completed two blocks of learning and testing phases: once with the presence of a social cue (Figure 4.5a) and once without (Figure 4.5b). The order of these blocks was counterbalanced across participants. At the beginning of the experiments, participants were informed that they are participating in a word learning task. They were asked to figure out which of the objects on the screen corresponds to the word the speaker said. They were not asked to look at one specific object in the learning phase but were only told to look at the screen at all times. In the test phase, they were specifically instructed to look at the object they think the word referred to.

Participants were exposed to a total of 32 training trials in one learning phase block. Learning trials presented an object in each quadrant and the speaker in the middle. Each word had eight learning trials. The target object appeared in all eight trials, twice in each of the quadrants. The HFD appeared in four trials, once in each quadrant. The remaining 10 distractors each appeared twice with the word, randomly assigned to a quadrant in each of their occurrences. Figure 6 shows the difference between regular learning trials and HFD trials.

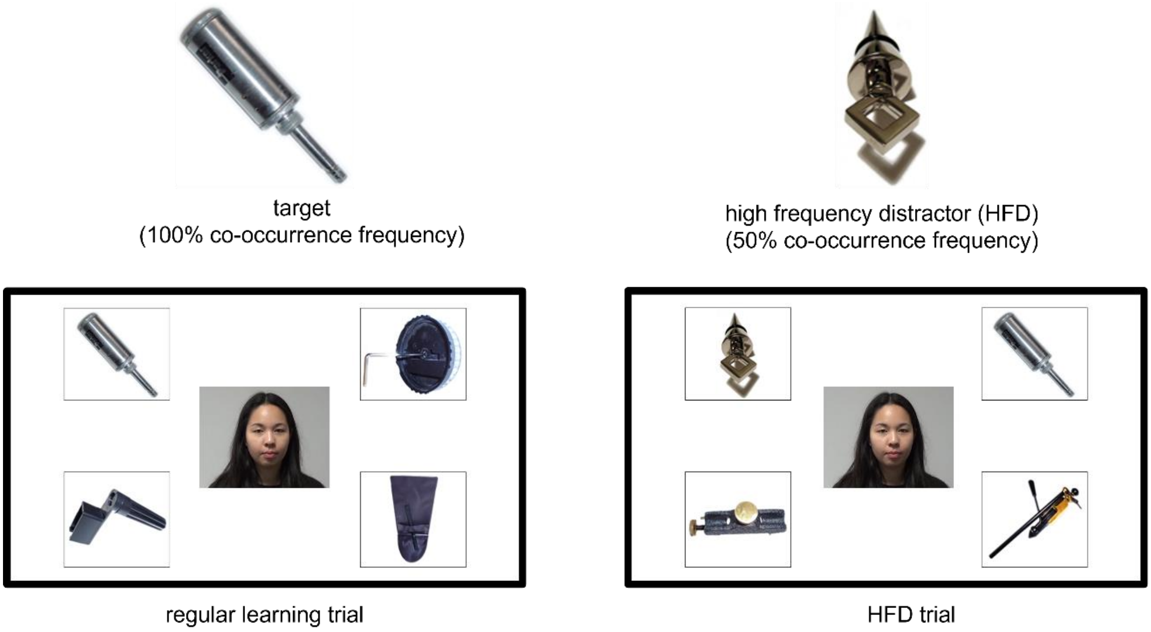

target
(100% co-occurrence frequency)
high frequency distractor (HFD)
(50% co-occurrence frequency)
regular learning trial
HFD trial

Figure 4.6

Learning phase trials showing regular learning trials where the HFD is absent and HFD trials where the HFD is present

The beginning of a trial showed a fixation cross for 500ms. This was followed by the four objects presented on the screen for 4000ms to allow participants enough time to examine each of the objects. The speaker then appeared on screen and remained facing forward for 500ms. She then produced the target word. This was how the trial ended for the no social cue condition. In the with social cue condition, the speaker turned their head towards the target for 1000ms after producing the word. She then looked back to the center and remained onscreen for 1000ms more.

In the consecutive trial spacing condition (Figure 4.7), learning trials were grouped based on the word. The 8 learning trials for each word were successively presented. There was a randomized presentation among the eight trials for each word to prevent any order effects. In the interleaved condition (Figure 4.8), the eight trials for each word were spaced. After a learning trial for a word, there were three trials in between before the same word was encountered again. The interleaved trials were those of the other words to be learned.

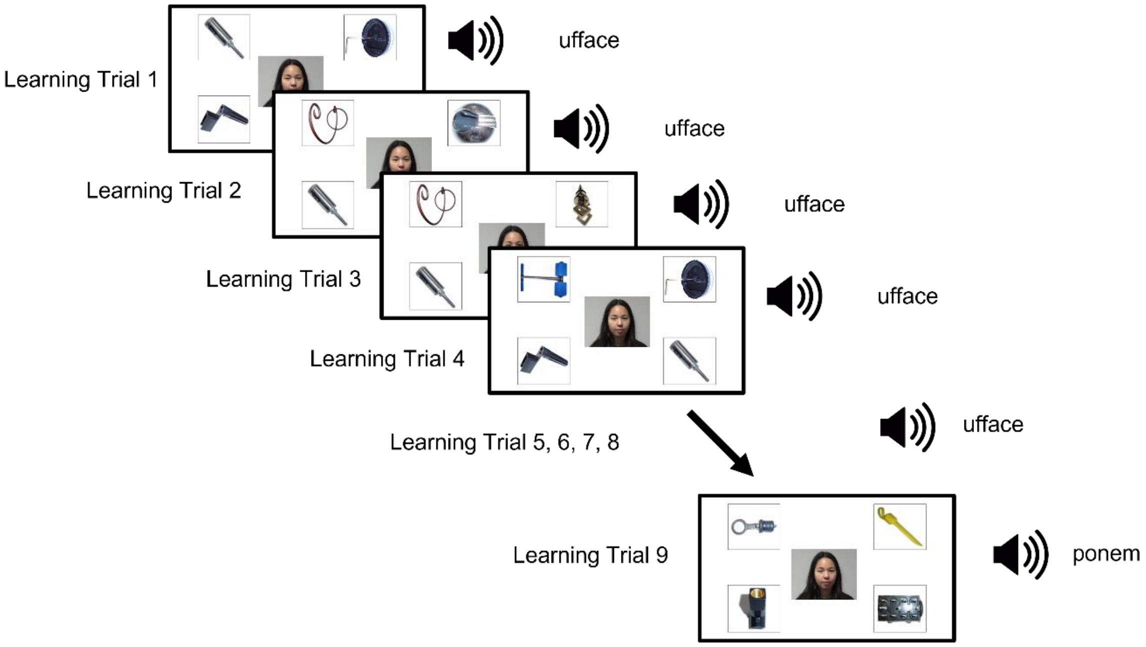


Figure 4.7

Consecutive block where all learning trials for the word are grouped together before learning trials for the next word are shown

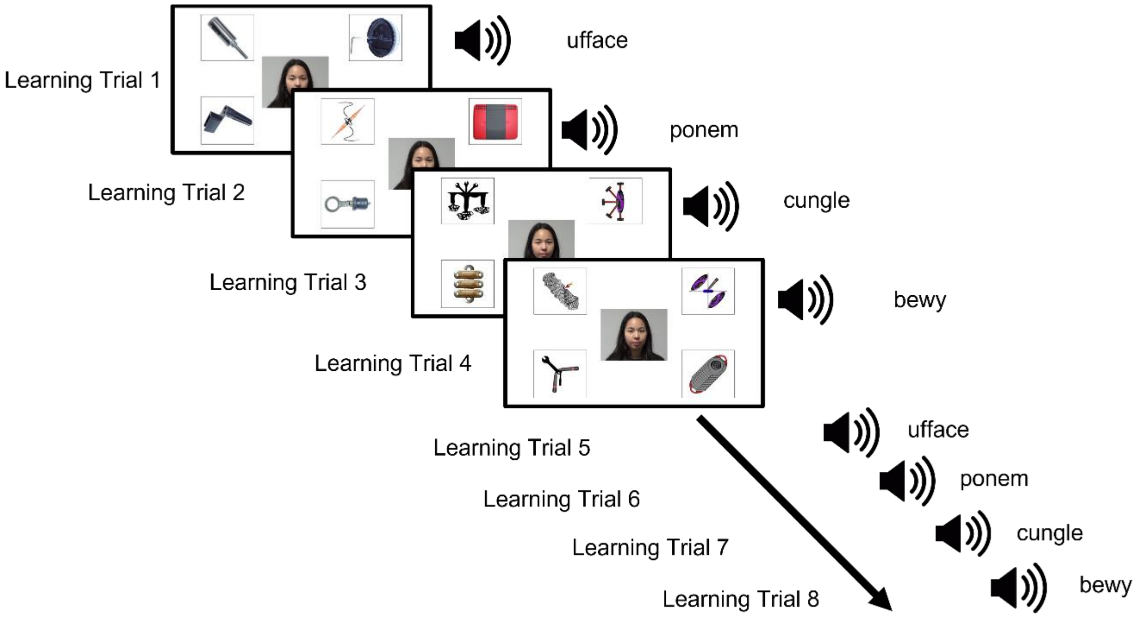


Figure 4.8

Interleaved block where learning trials for the four words in the block alternate with each other until all eight learning trials for all words are presented

After each training block, there were eight test trials in total (Figure 4.9). Four of the trials were same trials where the target was present. These appeared first. They were followed by four switch trials where the target was absent and replaced by a regular distractor, but the HFD would be considered the correct response. Therefore, each word had a corresponding same and switch trial. The objects were presented in a 2x4 grid, and the speaker did not appear on screen as the word played.

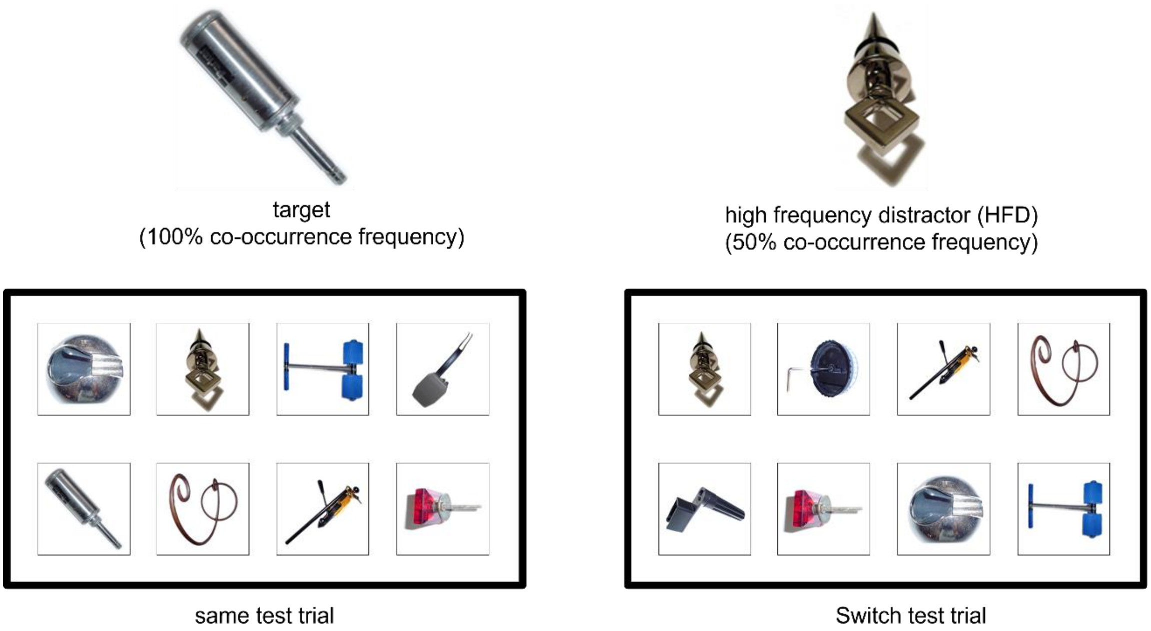

target
(100% co-occurrence frequency)
high frequency distractor (HFD)
(50% co-occurrence frequency)
same test trial
Switch test trial

Figure 4.9

Two kinds of test trials where same trials feature the target as the correct response and the switch trials switch the target out for the HFD as the correct answer

4.7 Results

The analysis only included 59 participants. The file from one participant had technical errors and their test phase responses could not be analyzed. Pre-processing of the eye tracking data used the saccades package (von der Malsburg, 2015) to identify fixations. The window of analysis was from 0 to 3500ms after word onset and data were grouped into 500ms bins. Growth curve analysis (Mirman, 2016) was used to analyze gaze to the target, HFD, and remaining distractors using the lme4 (Bates et al., 2015) package on R (Baayen et al., 2008). The growth curve analysis allows an analysis of the time course of cognitive processing by adding a high-order polynomial function to capture the changing trajectory of the dependent variable across time and conditions (Mirman et al., 2008). A quadratic terms was added based on visual inspection that the data approximated a curve with a single peak. The inclusion of the quadratic term in each analysis was assessed using the likelihood ration test to compare model fits with and without the quadratic term.

Models were initially specified with by-participant and by-item varying slopes for the linear and quadratic time term in the random effects structure, but this resulted in singular convergence. The random effects structure was gradually simplified by first removing the quadratic term, which still resulted in a singular fit. The by-participants and by-items slopes for the spacing condition and social cue conditions were then removed, but these models were overparameterized and the existing data could not support them. Models containing only by-participants and by-items intercepts were returned without warnings and this was the random effects structure used for the test phase and learning phase models.

The main effects spacing condition (0-consecutive, 1-interleaved) and social cue (0-without social cues, 1-with social cues) were treatment coded. For the same trials and learning phase, Helmert coding was applied to area of interest where the estimate for target was compared to the estimates for both HFD and distractors combined and the estimates of the HFD was compared to only that of the distractor level. The contrast between the HFD and the other distractors is to measure if participants were sensitivity to the higher co-occurrence of the HFD with the target word compared to the other distractors. In the switch trials, area of interest was treatment coded (0- HFD, 1- distractor) since the target was absent. The learning phase had an additional main effect, trial number, which used forward difference coding and compared the one level to the one immediately after it. This variable had four levels. Despite the learning phase having eight trials in total, there were only four which showed the HFD. These trials were included in the analysis and were coded as a number between 1 to 4 depending on their order of appearance in the entire learning phase.

Model comparison was performed using the likelihood ratio test and was used to evaluate the effect of each added fixed effect on model fit. The base model for both the test phase and learning phase included only the linear and quadratic time terms as fixed effects and the by-participants and by-items intercepts. This was compared to a model with area of interest as a main effect to confirm if participants were looking to specific objects on the screen versus not looking at any object in particular. If this comparison was significant, trial number (only for the learning phase), spacing condition, and social cue were added as main effects one-by-one without any interactions and only included the linear term. Another comparison was made where there were interactions among the main effects and the quadratic time term was included. The model with significantly lower AIC component was used to interpret parameter-specific estimates.

4.7.2 Performance on test phase same trials

Participants proportion of fixations to the target compared to the different distractors in the same trials were first analyzed (Figure 4.10). This is to determine they have learned the word-object mappings and confirm if they had truly experienced a "switch" in the switch trials as their preferred referent was no longer there. The model with only the linear and quadratic time terms was compared to one where area of interest was added as a main effect which significantly improved model fit (χ2(2) = 2181.08, p < .001). Spacing condition (χ2(1) = 8.22, p < .05) and the quadratic term (χ2(4) = 2260.26, p < .001) all significantly improved model fit when included along with their interactions with area of interest, but social cue did not improve model fit (χ2(1) = 1.01, p > 0.05). Participants were looking significantly less to the target in the interleaved condition (t = -3.31, p < .001), but there was no difference in their gaze behavior to the HFD compared to the other distractors (t = 0.60, p > 0.05). The model with the best fit is shown in Table 4.2.

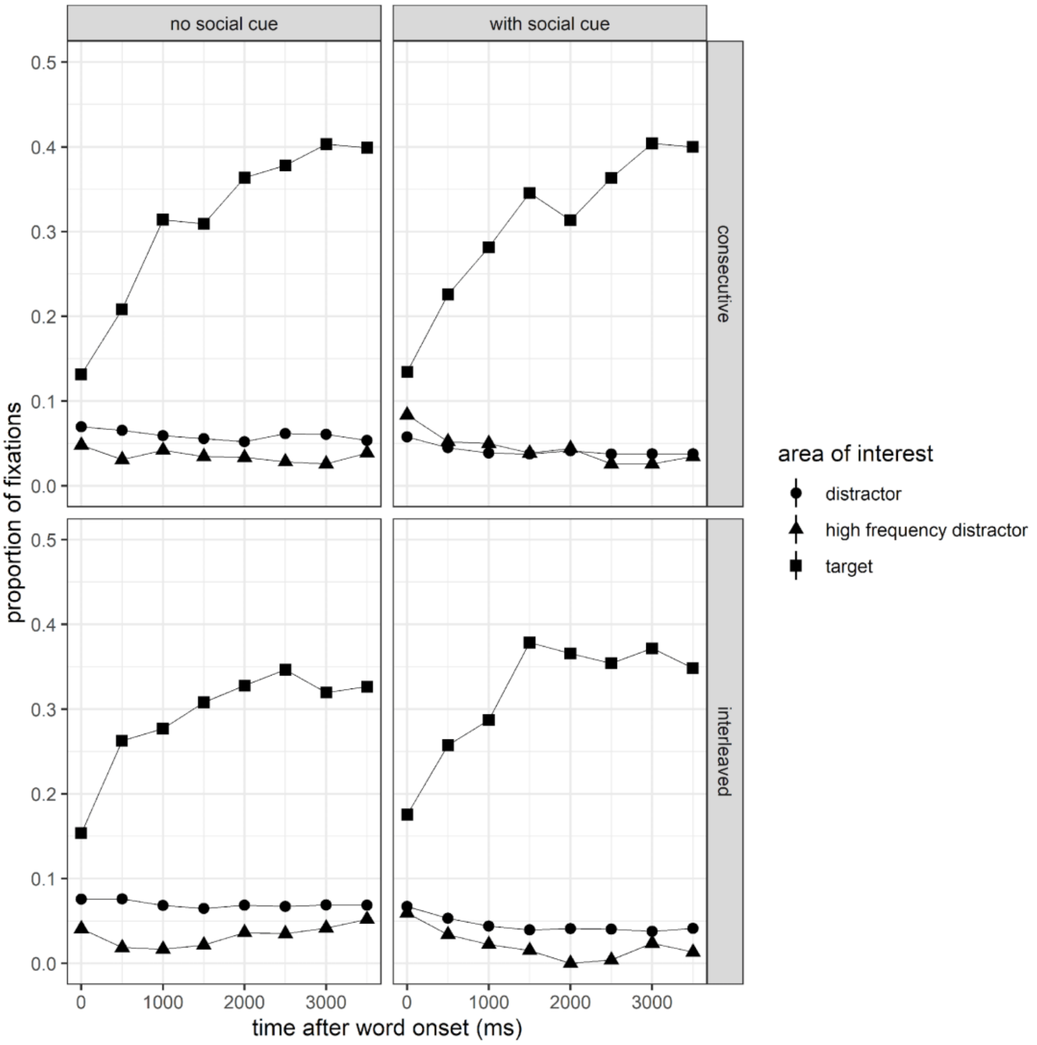


Figure 4.10

Proportion of fixations to the different areas of interests in test phase same trials across spacing and social cue conditions

Effect	Estimate	S.E.	t-value	p-value
(Intercept)	0.13	0.01	9.77	p < .001
Main effect of time				
Linear	0.07	0.01	6.44	p < .001**
quadratic	-0.02	0.01	-1.75	p > .05
Main effect of area of interest				
target	0.27	0.01	34.97	p < .001**
high frequency distractor	-0.02	0.01	-1.24	p > .05
Main effect of spacing condition				
interleaved	0.00	0.02	-0.24	p > .05
Interaction between area of interest and time				
linear:target	0.25	0.02	11.75	p < .001**
linear:high frequency distractor	-0.02	0.04	-0.57	p > .05
quadratic:target	-0.09	0.02	-4.08	p < .001**
quadratic:high frequency distractor	0.01	0.04	0.25	p > .05
Interaction between spacing condition and time				
linear:interleaved	-0.02	0.01	-1.73	p > .05
quadratic:interleaved	0.00	0.01	-0.11	p > .05
Interaction between area of interest and spacing condition				
target:interleaved	0.00	0.01	-0.41	p > .05
high frequency distractor:interleaved	-0.03	0.02	-1.59	p > .05
Interaction among area of interest, spacing condition, and time				
linear:target:interleaved	-0.10	0.03	-3.31	p < .001**
linear:high frequency distractor:interleaved	0.03	0.05	0.60	p > .05
quadratic:target:interleaved	-0.03	0.03	-0.85	p > .05
quadratic:high frequency distractor:interleaved	0.01	0.05	0.26	p > .05

Table 4.2

Model on proportion of fixations in same trials with main effects and interactions among areas of interest, spacing conditions, and linear and quadratic time terms

4.7.3 Performance on test phase switch trials

Performance on the switch trials were then examined (Figure 4.11). The model with area of interest as a main effect did not differ significantly from the base model (χ2 (2) = 2.83, p > 0.05). Participants appear to be looking equally at the HFD and distractors. Spacing condition did not improve the model fit (χ2 (1) = 0.78, p > 0.05) nor did social cue (χ2 (1) = 0.00, p > 0.05).

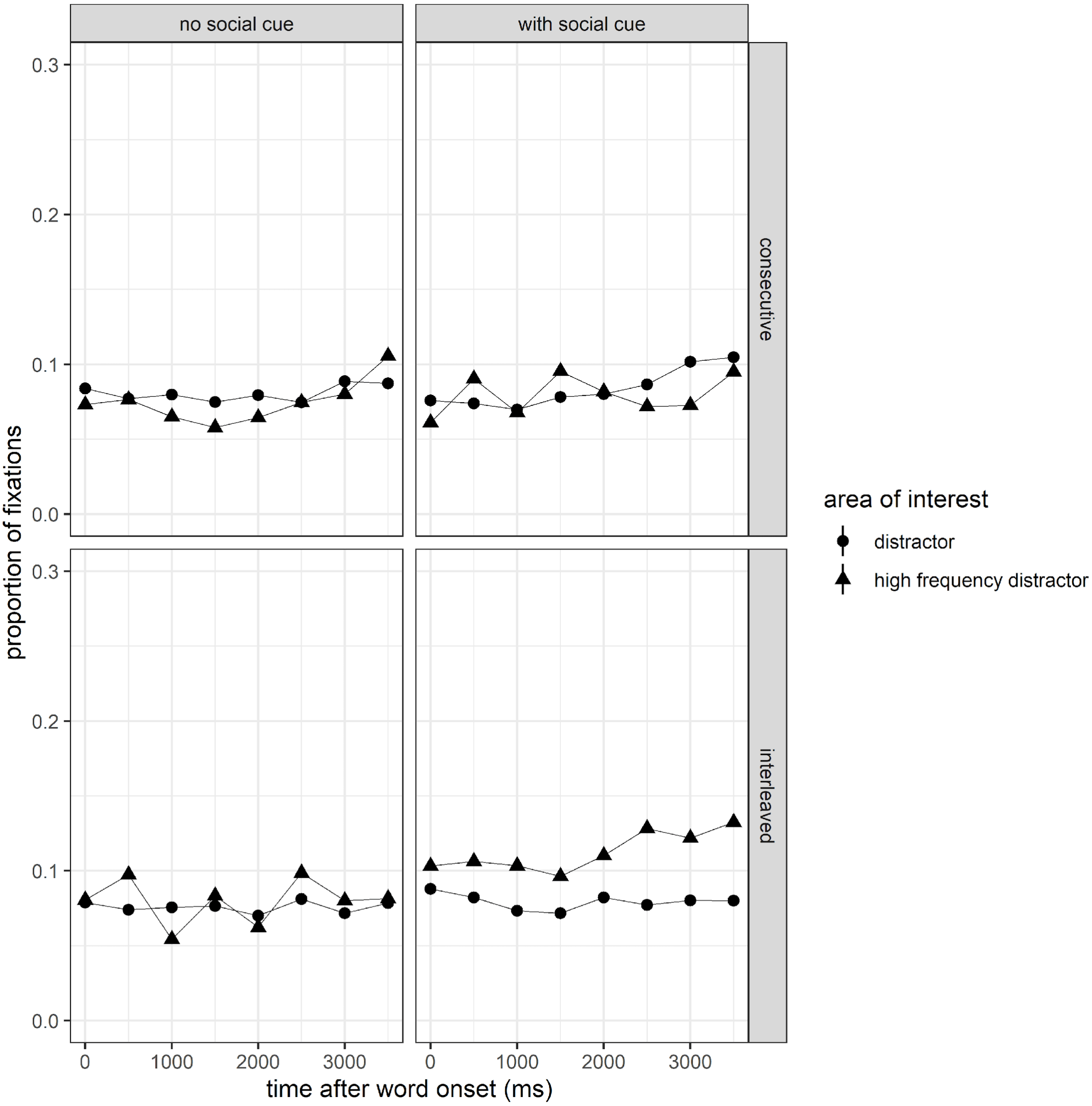


Figure 4.11

Proportion of fixations to the HFD vs. distractors in test phase switch trials across spacing and social cue conditions

4.7.4 Performance during learning phase

The data for the learning phase involves a trial-by-trial analysis to examine how participants were developing knowledge of the word-object mappings (Figure 4.12). The main effects of area of interest (χ2 (2) = 5198.09, p < .001), trial number (χ2 (62) = 5520.62, p < 0.001), spacing condition (χ2 (22) = 5451.54, p < .001), and social cues (χ2 (22) = 5451.54, p < .001) were significant as were their interactions and the inclusion of the quadratic term.

As the hypotheses for the learning phase involved comparing the looks to the HFD to the regular distractors, only trials where the HFD was present were included. Otherwise, including all learning trials created artifacts in the data that made it seem that participants showed no fixations to the HFD when it was actually not present among the options. This resulted in four out of the eight learning trials used in this analysis for the analysis to be possible. The HFD was randomly distributed across the eight learning trials, therefore the trial number refers to the order of when the trial appeared compared to the other HFD trials. In other words, trial 1 would be the first learning trial that contained an HFD and not necessarily the first learning trial overall. Consequently, this contributes to the higher looks to the target as early as trial 1 in Figure 12.

The learning phase allows the question of whether the presence of a social cue allows word-object mappings to be learned faster (Appendix A). Generally across the learning phase, there is a larger proportion of fixations to the target versus all other distractors on each trial (trial 2 (t = 2.90, p < .05); trial 3 (t = 4.20, p < .05); trial 4 (t = 4.07, p < .05)). This is not the case when comparing the difference in looks to the HFD to other distractors (trial 2 (t = 1.01, p > .05); trial 3 (t = 0.93, p > .05); trial 4 (t = 0.46, p > .05)) which implies that participants did not consider the HFD as a potential referent at any point. To determine the effect of social cues on learning, the interaction among the time, area of interest, and social cue terms was examined. There was a significant increase of looks to the target compared to distractors in the with social cue condition (t = 3.96, p < .001) and the rate of this increase is also significantly faster (t = 2.50, p < .05). When including the interaction with trial number, only the quadratic time term on trials 3 (t =-2.76, p < .05) and 4 (t =-2.51, p < .05) were significant. Participants were faster at identifying the target with a social cue later in the learning phase and did not show a sudden rise in looks to the target compared to distractors early on. There was no difference in the proportion of looks to the target nor the rate of increase of these looks between the spacing conditions (t = -1.57, p > .05).

Since there were no differences in the results of the test phase, we examined the learning phase data further to determine if a combination between trial spacing and social conditions influences the mechanism of CSWL. From the same model, we took the interaction among all fixed effects. There are significantly more looks to the target at trial 2 (t = -2.03, p < .05) when comparing the interleaved with social cue condition to the consecutive no social cue condition. This difference does not exist in later trials (trial 3 (t = -1.49, p > .05); trial 4 (t = 1.87, p >.05)). It appears that the interleaved with social cue condition influenced participants to select a target early on in the learning phase, while learners in the consecutive no social cue condition have not made a choice yet. As this difference is no longer significant in succeeding trials, participants in both conditions have appeared to look at the target the same amount by the third trial.

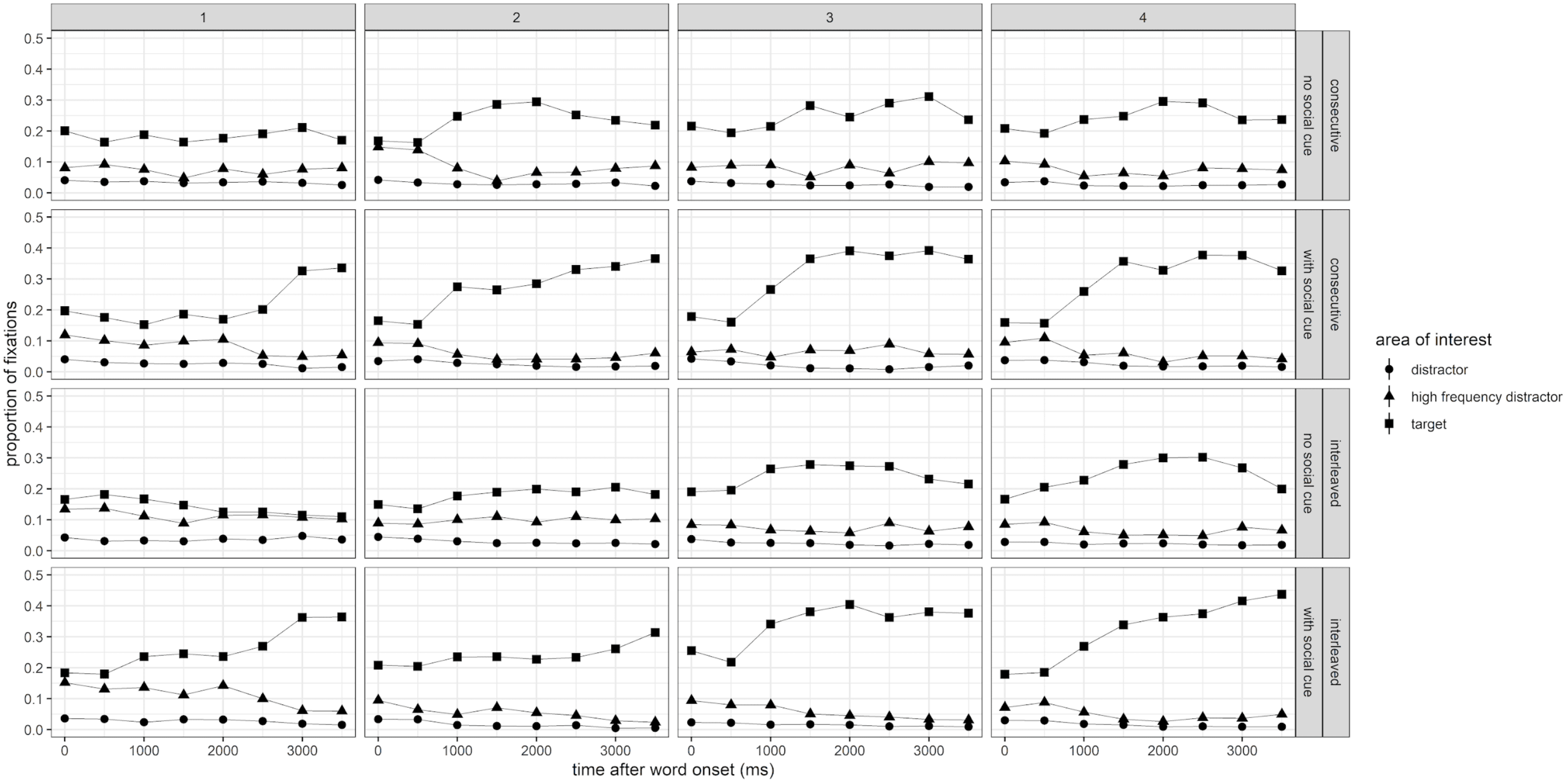


Figure 4.12

Proportion of fixations during the learning phase across trials, spacing conditions, and social cue conditions

4.8 Discussion

This online eye tracking experiment further investigated the underlying mechanism of CSWL. Trial spacing and the presence of eye gaze were manipulated to increase demand and provide support in the learning situation, respectively. The hypotheses were tested using a high frequency distractor (HFD) and switch trials in the test phase. The switch trials did not show the target, the object with perfect co-occurrence with the label, and only had the HFD alongside other distractors. The HFD stood as the most probable referent since it has higher co-occurrence with the label compared to the remaining distractors. Longer looks to the HFD during the trial would indicate memory of alternative referents during learning which would imply an AL approach. The looks to the HFD in the test phase switch trials were used to determine the individual influences of trial spacing and social cue on which approach better described participants' looking behavior, as well as whether or not one variable exerted a stronger influence than the other or were they both equally considered.

The first question was regarding the influence of trial spacing on how participants aggregated co-occurrence frequencies. The hypothesis was that participants would present more looks to the HFD than to the distractors in the switch trials, more so in the consecutive spacing condition compared to the interleaved condition. Conversely, it was expected that there would be less looks to the HFD in the interleaved condition.

Results show that regardless of spacing condition, participants did not look longer to the HFD compared to the other distractors on the switch trials. This would imply that they did not consider the HFD as a possible referent more than they did the other distractors. These findings do not completely agree with those of Smith et al. (2011) and Aussems and Vogt (2020) in that their participants showed behavior more in line with AL in both consecutive and interleaved conditions. It was only that the interleaved condition reduced the tendency to perform AL due to the additional memory demand, yet they did not find evidence completely favoring HT. This is where the findings of this study differ as the lack of difference in looks in the switch trials seem to support a full HT approach.

Although not part of the hypothesis, the looking behavior in the same trials presents a picture more in line with Smith et al. (2011) and Aussems and Vogt (2020). In the same trials, participants in both spacing conditions looked significantly longer to the target which means that they did identify the object with the highest co-occurrence frequency. However, participants did have significantly less looks to the target in the interleaved condition compared to the consecutive condition in the same trials. It appears that participants assigned to the interleaved condition experienced more competition from the distractors and supports that including a delay between learning trials places additional memory demands which impacts learning. These findings are in line with other studies that interleaved condition presents additional difficulty for participants based on their performance on the same trials.

In other domains outside CSWL, interleaved presentation has been found to create better learning outcomes such as in second language learning (Kim & Webb, 2022; Nakata & Suzuki, 2019; Pan et al., 2019), mathematics (Rohrer et al., 2015), and science (Eglington & Kang, 2017). The opposite appears to be true in CSWL that interleaved presentations in ambiguous learning situations create difficulty for the learner.

One account claims that the benefit of interleaved presentations is derived from the more effortful retrieval required when forgetting what was learned after a delay (Delaney et al. 2012). There is a larger chance that participants will forget the information being learned when interfering trials are presented as compared to presenting learning trials consecutively. If information is forgotten after this delay, then this would result in more effort to retrieve the information. The ambiguity present during CSWL might prevent participants from fully benefitting from this retrieval regardless of the way participants aggregate co-occurrence frequencies.

If participants adhere to an HT account, participants might make several incorrect hypotheses and would be retrieving the incorrect word-object link. If they adhere to the AL approach, they will need to retrieve several possible word-object links which would distribute the benefits from effortful retrieval across objects that aren’t even the correct referent. Learning during CSWL requires multiple exposures before the correct word-object link could be established and interleaved presentations would further contribute to this delay. Therefore, similar benefits from interleaved presentations in other domains of learning might not be available in CSWL.

The second research question investigated the effects of a social cue on CSWL mechanism. It was hypothesized that during the testing phase, participants would show no difference in looks to the HFD versus the distractors for the switch trials in the with social cue condition. Without a social cue, participants were expected to show more looks to the HFD for these trials, which would be more in line with the AL approach.

Similar to trial spacing, participants did not exhibit higher looks to the HFD compared to the distractors for the switch trials in both social cue conditions. This suggests that participants were tracking only one referent at a time as they showed no memory for the HFD which had a slightly higher co-occurrence with the word. MacDonald et al. (2017) also found that participants were less likely to look at alternative referents in switch trials when they were presented an eye gaze cue which implies that the presence of a reliable social cue influences participants to adopt a HT strategy. However, the same conclusion cannot be derived from this study since performance was the same regardless of social cue condition.

Participants' performance during the learning phase was also hypothesized to reflect a HT pattern for the with social cue condition. In this case, a sharp rise in looks to the object supported by the social cue was expected. If participants were performing AL, a gradual increase would be seen. Results from the learning phase show that the social cue aided them in finding the target faster, but only at the later trials. This does not seem to be a sudden rise in accuracy as Medina et al. (2011) and Trueswell et al. (2013) have found. If participants were performing pure HT, then participants should have been faster in locating the target as soon as the second trial given that they had found the target in the first. Looking times to the HFD were not significantly different from the distractors which implies that participants have not considered it as a possible referent despite its higher co-occurrence frequency. This might explain why they also showed no preference for this in the switch trials. As Koehne et al. (2013) have found, it was only when participants looked at the HFD more than 50% of the time in any of the learning trials did they were above chance at selecting it in the test phase.

An additional interpretation of the results from the same trials is also presented as it complements some of the findings in existing literature. It appears that the presence of a social cue does not lead to increased looks to the target in the test phase same trials. Participants appear to learn look the same degree to the target regardless of the social cue condition, which was also in line with the findings of MacDonald et al. (2017). On the contrary, Monaghan et al. (2017) found that a gestural cue significantly improved accuracy in CSWL learning when they made use of a pointing cue. There is evidence in adult learning that while eye gaze was effective in enhancing looking towards an area of the screen, pointing increased ease of visual search, attention towards stimuli, retention, and transfer (Pi et al., 2019).

The last research question asked if there was an interaction between trial spacing and social cues during CSWL. There were three possible outcomes: 1) that trial spacing had a stronger influence than social cues; 2) that social cues had a stronger influence than trial spacing; 3) that these two factors were equally considered. If the first option were true, participants would show behavior more consistent with AL regardless of social cue condition since they would not use gaze as a shortcut to determine the referent. If the second option were the case, participants would exhibit HT behavior regardless of spacing condition since gaze reduced ambiguity enough that they would prefer not to engage in AL even though the task demands were low enough. The last option was possible if a combination of high task demands presented by interleaved spacing and reduced ambiguity due to the presence of eye gaze sufficiently encouraged learners to adapt an HT strategy.

The results from the switch trials are not really informative as to whether there is any interaction between trial spacing and social cues, as these results currently suggest that participants perform HT regardless of condition. The additional analysis of the data in the learning phase shows that participants looked earlier to the target in the interleaved social cue condition compared to the consecutive without social cue condition. We hypothesized that this combination of trial spacing and social cue would result in participants opting for an HT approach, since the memory demands placed by spacing and supportive cue provided by eye gaze would encourage the tracking of a single referent. In this case, it could be that the effects of these two factors were more evident during learning as it shows the trial-by-trial change in how participants were tracking objects.

These findings should be taken into consideration alongside the limitations of this study. Hypotheses were set in terms of participants' looking patterns towards the HFD, but the distribution of trials containing the HFD might have impacted how participants considered it as high frequency. Each word had eight learning trials and four of these showed the HFD, but the distribution of the HFD trials were random for every participant. This means that for some participants, the HFD trials could have appeared in the latter half of the learning phase. If this were the case, participants have already received greater exposures to the target that might have made the HFD less competitive. In order to fully tease apart participants' preference for the object of a speaker's eye gaze, the HFD should be presented alongside the target in the early half of the learning phase. This would allow the social cue to be the only disambiguating factor early on until the point where the target's co-occurrence frequency begins to exceed that of the HFD. This would be a fairer comparison of the two possible referents.

Future lines of research could investigate how participants respond to cues that compete with co-occurrence statistics rather than support it. Although majority of social cues coincide with the object or even a speaker is referring to, there is still a percentage of utterances where a speaker could talk about something that is absent in the environment (Kyger, 2013; Harris et al., 1984). It is also interesting to investigate how different types of cues, syntactic, social, and prosodic, integrate with CSWL in early development. There seems to be a pattern where 12-month-old infants prefer salience cues over social cues in deciding a word-referent relationship (Hollich et al., 2000). This could add further insight to which cues present in early development do young learners rely heavily on to get word learning off the ground.

Another line of research would be to investigate if trial spacing affects retention of mappings after a period of delay. Results from the test phase already show that participants in the interleaved condition perform poorer compared to those in the consecutive condition, although both groups did learn the word-referent mappings. There is the question of how performance would change after a period of consolidation and if the disadvantage of interleaved spacing would still hold.

In conclusion, this experiment found that participants were not showing memory of alternative referents when tested on word-object links but appear to be gradual in their selection of a referent during the learning phase. The timing of when a competitor is encountered may affect whether or not learners consider it as possible referent. Interleaved presentation of learning trials, contrary to benefits seen in other domains, presents an additional challenge to learners engaging in CSWL. Social cues, on the other hand, aid learners in locating the referent faster as they are learning but does not result in a significant improvement when learning is measured.

Chapter 5. General Discussion

5.1 Overview of Findings

This dissertation investigated how cross situational word learning (CSWL) integrated with phonetic learning and social cues which are domains closely tied to early word learning development. It also studied the underlying mechanism of compiling co-occurrence frequencies. The first study tested if learners could differentiate a non-native contrast by being exposed to unique word-referent co-occurrences in CSWL. Their ability to discriminate the contrast in pure sound discrimination and word recognition tasks was tested. The second study extended these questions to 4-year-old children and compared performance between adults and these young learners. The third study investigated the mechanism underlying CSWL and tested if participants' memory of alternative referents was affected by the presence of social cue and trial spacing, both of which influenced the difficulty of the learning situation.

The first study was comprised of two experiments conducted with adult native German speakers who were trained to discriminate between the Hindi dental and retroflex voiceless stop. The participants completed four tasks: a pre-training sound discrimination task, a CSWL-based training phase, a post-training sound discrimination task, and a word recognition task. The contrast was embedded in a minimal pair and each member was assigned a unique referent. The word-referent pairings were introduced through CSWL along with two other word-referent pairings that were not minimal pairs and did not contain the contrast that served as a control condition. The sound discrimination task was used to test if participants were able to learn the contrast by encountering them in different referential contexts. The word recognition task tested if participants learned the word-referent mappings that contained a non-native contrast. The two experiments differed in the sound discrimination task they used.

The first experiment made use of an AXB task where participants responded if it was the first or third sound that was similar to the middle sound. The participants were at chance level in the AXB task before and after training, however they were able to respond above chance in the word recognition task. This still required some discrimination abilities as there were trials where the two choices were between the objects that were linked to the minimal pairs. It was considered that the AXB task places a heavy working memory load that may encourage categorical perception instead of within-category perception (Massaro & Cohen, 1983; Pisoni & Lazarus, 1974), which is why the second experiment used a different version of the sound discrimination task.

A second group of participants were recruited to perform the experiment again but using an AX task as a measure of sound discrimination abilities. This version of the task only required participants to respond if the two sounds were the same or different. The order of the post-training sound discrimination task and word recognition task was also balanced among the participants to ensure that the better performance in the word recognition task was not due to additional exposure during the previous task. Nevertheless, participants still showed a discrepancy in their performance. They remained at chance level in the sound discrimination task, but above chance in the word recognition task.

There is evidence that L2 phonology is more closely tied to age as compared to other areas (Pallier et al., 1997; Oyama, 1976; Flege et al., 1999), which is why study 2 replicated the second experiment in the previous chapter with 4-year-old German-speaking children and a new adult control group. This study used a similar pre-test post-test design but adapted the AX task to make it more child friendly. Instead of asking participants to press specific keys to respond same or different, the task was structured into a game where participants would indicate by pointing to pairs of images of speakers to indicate if the sounds heard were same or different. At the start of the block, children were given two familiarization trials for the Hindi dental and retroflex voiceless stop. In each familiarization trial, the child was shown one alien of a specific color paired with one member of the minimal pair played four times. In the test phase, participants were asked to choose the image showing the same aliens if they heard the same sound and the image showing different aliens if different sounds were played. An experimenter was present in a video call with the children as they completed the experiment. This allowed the experimenter to video record the call and capture children's looking behavior as they completed the task, providing an implicit measure of performance as they completed the task.

While adults again showed no improvement in their discrimination abilities, their performance on the AX task was above chance both before and after training. Children, on the other hand, were at chance for both sound discrimination and word recognition measures based on their behavioral data. Their looking data was also at chance, but there appears to be a trend approaching significance for the looking data in the sound discrimination task. Overall, it appears that adults were outperforming children on both tasks.

The last study investigated the underlying mechanisms of CSWL using an online eye tracking study with native English-speaking adults. The experiment tested whether there was more evidence for the associative learning (AL) or hypothesis testing (HT) approach and whether this varied with the difficulty of the learning situation using eye gaze and trial spacing. Participants were assigned to learn words with learning trials presented consecutively or with interfering trials in between (i.e., interleaved condition). Findings from a previous CSWL study showed behavior more similar to associative learning when learning trials are presented consecutively (Smith et al., 2011).

This study assigned participants into different trial spacing conditions where they completed both blocks of CSWL training with and without an accompanying eye gaze from an onscreen speaker. Eye gaze is a cue for referential intent and participants in one CSWL study that used eye gaze showed performance more similar to hypothesis testing (MacDonald et al., 2017). To determine which approach best described participants' looking behaviors, we made use of a high frequency distractor (HFD) and switch trials in the test phase. The HFD is a distractor that co-occurred only 50% of the time with the novel word and would be less preferred than the target as the correct referent. In the test phase, we had switch trials where the target was absent and only distractors and the HFD were present. In switch trials, the HFD would be the correct answer as it had a higher co-occurrence frequency compared to other distractors. Engaging in an AL approach would require keeping track of multiple possible referents which would allow the HFD to be selected as the correct answer in the absence of the target.

The participants did not consider the HFD as the correct answer in the switch trials regardless of trial spacing or social cue condition. It appears that they did not store multiple referents during the learning phase, and this is more in line with an HT approach. However, their behavior during the learning phase does not fully support this. Eye gaze in the learning phase helped participants locate the target object faster in the later trials but not in the earlier ones, which means that participants were not completely relying on eye gaze to locate the target at first. Their performance presented a more gradual increase in looks to the target which is not what is predicted by the HT account. An exception to these findings is the interleaved with social cue condition. Participants were more likely to select the target in the first trial in the interleaved with social cue condition as compared to other trial and spacing conditions. There are also more looks to the target in the first trial of interleaved social cue condition which implies that this learning situation encourages more HT-like behavior.

5.2 CSWL and Phonetic Information

On the first two experiments with adults, participants were at chance in measures of pure sound discrimination but above chance when discrimination is tested in a word recognition task. A previous study by Tuninetti et al. (2020) found that word learning during CSWL was affected negatively if words contained difficult to discriminate non-native contrasts but also found that learning word-referent mappings was still possible. However, this study differed from theirs in that this study sought to use CSWL to train a single contrast and included measures of pure sound discrimination. Their study exposed learners to multiple non-native contrasts during CSWL with varying degrees of difficulty and the majority of the difficult contrast they used followed a subset pattern according to the L2LP (van Leussen & Escudero, 2015) which would be easier than the new pattern that the Hindi dental-retroflex contrast used in this study falls under. This is because there is still some possibility to discriminate a contrast under the subset pattern based on one sound being a good exemplar for the L1 category it assimilates to and the other being a poor exemplar, while in the new pattern both members of the contrast assimilate equally well into the same L1 category. This dissertation found behavior in adults in line with the Tuninetti et al. (2020) study that learning word-referent mappings is still possible despite the presence of a non-native contrast, yet this contrast remains difficult to perceive in tasks of pure sound discrimination.

The ability to perceive the contrast in a referential context but not in a phonetic context is accounted for by the L2LP by allowing for parallel processing and acknowledging that links at some levels of processing might be stronger than others. In context of these findings, it appears that new links at the pre-lexical level, namely the acoustic and phonetic levels, have not been sufficiently strengthened to allow discrimination in the AX and AXB tasks. The L2LP assumes that L2 sound categories are direct copies of L1 categories and for learners to successfully discriminate a contrast like the Hindi dental-retroflex a new completely new sound category must be formed. Based on their sound discrimination performance, adults have not achieved this. They do, however, appear to benefit from the association with the unique referent that possibly strengthened connections at the lexical level enough for them to perform above chance in the word recognition task.

The next question is why CSWL-based training strengthens connections at the lexical level and not at the pre-lexical levels. One explanation could emerge from the Exemplar Theory as it applies to speech perception and word recognition (Goldinger, 1996). In this theory, listeners store detailed episodic traces of lexical forms rather than abstract categorical representations. These individual exemplars share common semantic, phonological, visual, talker-specific, and event-specific information that allows them to be grouped into categories. As exemplars are memory traces, they do fade over time. However, new exemplars quickly take their place and the more similar these are to those already existing in the cognitive space the more the categories are reinforced. In our CSWL training, participants might have benefitted from the additional referential information provided in the word recognition task since this would be information linked to the exemplar for these lexical forms. Although the sound discrimination tasks made use of the exact same tokens which would contain the exact same acoustic and phonological information, it is distinctly missing the referential context. Therefore, the word recognition task might have allowed adults to benefit from additional referential information linked to the lexical forms that was not available in the tests of pure sound discrimination.

The comparison of performance on the AXB and AX task did not yield any differences which implies that the working memory demands reported to differ between the two was not observed in Study 1. However, the design of these sound discrimination tasks and the demands they place on participants does emerge in Study 2 with the child-friendly adaptation of the AX task. In this version, the above chance performance of the adults regardless of training may be due to the difference in stimuli presented alongside the non-native contrast. In Study 1, both the tasks included native minimal pairs as attention checks interspersed among the stimuli containing the non-native contrast. The inclusion of trials wherein participants can rely on their existing L1 categories to respond correctly may have prevented them from tuning into the differences relevant for the L2 stimuli. As the L2LP model proposes, successful discrimination of difficult to perceive L2 contrasts requires adjustment of L1 phonetic boundaries. Participants would be successful without these adjustments for a third of the trials which may have posed an additional challenge in reducing perceptual assimilation effects. In contrast, success in the child-friendly AX task was dependent on the ability to adjust these L1 boundaries. Given that participants were performing above chance, it appears that presenting only L2 stimuli allowed them to attune to the relevant phonetic detail.

Children, on the other hand, appear to not have benefitted from this referential information at all. They were performing at chance in discriminating the contrast both in the AX task and the word recognition task. In other lab-based studies that trained a non-native contrast, adults were found to outperform children initially (Wang & Kuhl, 2003; Heeren & Schoten 2010, Fuhrmeister et al., 2020). It was only after multiple sessions that this advantage of adults disappeared and children either performed equally (Heeren & Schoten 2010) or outperformed adults on measures of discrimination (Fuhrmeister et al., 2020). Since this study only performed immediate testing after training, there are no longitudinal measures to ascertain that children in our study will show learning over time.

Currently, it appears that a CSWL-based training does not appear beneficial for promoting discrimination of a non-native contrast in 4-year-old children. It also appears to effectively block their ability to learn word-referent mappings containing difficult to perceive L2 sounds. As was found in adults (Tuninetti et al., 2020), the difficulty of the contrast directly impacted ability to map word-referent relationships in CSWL. This appears to be the case for children in this study too. In addition to this, it appears that children also did not benefit from the additional referential context provided by the word recognition task as adults did.

There are findings that speech perception moves from underspecified holistic representations in infancy to a more segmental approach in early childhood due to growth in vocabulary (Walley, 1993). As children increase the number of words in their vocabulary, a subsequent increase in phonological neighbors leading to a need for more detailed lexical representations to distinguish among these entries during word recognition. As such, children in this study might have been unable to benefit from strengthening the links at the lexical level as they are more inclined to focus on encoding fine phonetic detail at the segmental level. However due to the difficulty of the contrast and the influence of their L1, this has led to unsuccessful discrimination of the contrast in the AX task. Since they are also focusing on the pre-lexical levels as they do so in their L1, this might have prevented them from employing an exemplar-based approach as adults have done which manifested as poor performance in the word recognition task as well.

There is a finding in development that infants at 14 months old fail to pair words with objects when words contain an L1 minimal pair contrast even if it is one that they successfully discriminate in a speech perception task (Stager & Werker, 1997). Infants could overcome this as their vocabulary increases and they experience the contrast in more distinct lexical distributions (Thiessen, 2007). The children in our study, although much older, also have encountered the L2 contrast in minimal pairs. It is possible that they need to encounter this novel contrast in distinct lexical contexts that would provide them more information to differentiate the contrast not just in terms of acoustic-phonetic differences but also using distributions. Relying purely on acoustic-phonetic differences would be difficult due to the L1 masking the differences.

Based on these results, adults are able to benefit from the additional referential context learned during CSWL training to discriminate a non-native contrast. It also appears that CSWL training promotes representations of the contrast at the lexical level but not necessarily at the phonetic level. Children did not show immediate learning of an L2 contrast after similar training. There is, however, potential that learning may emerge from multiple training sessions due to learning across time and using implicit measures of sound discrimination such as looking times.

5.3 CSWL and Social Cues

This dissertation also found that eye gaze is utilized by adults in aiding them to locate the referent faster in CSWL. However, gaze did not result in better performance in the test phase. This coincides with the findings of MacDonald et al. (2017) where their participants in a CSWL experiment chose the referent which was the object of an onscreen speaker's gaze during the learning phase, but participants who were provided a gaze cue did not exhibit longer looks to the target in the test phase.

It is also worth noting that locating the target using the social cue only occurred in later trials and not at the onset. It is possible that once participants learned that the co-occurrence frequencies are supported by the direction of the gaze of the onscreen speaker, they were more likely to use it during CSWL. The results from study 3 also show that learners were more likely to select the target in the earlier trial when the learning context was more difficult due to interleaved presentations and when accompanied by a social cue. This shows that participants were more likely to make use of a reliable cue when the learning situation was more difficult.

The findings from all three experiments show some evidence that even when learning words under ambiguity, learners can also integrate phonetic and social information during word learning.

5.3 Mechanisms underlying CSWL

This dissertation had mixed findings on the mechanism underlying CSWL. High frequency distractors (HFD) were included to measure participants' memory of alternative referents. Their looking behavior in the learning and test phase shows that they did not prefer the HFD as a referent compared to other distractors in switch trials in the test phase. This is contrary to the findings of other studies that made use of a high frequency competitor and switch trials, where their participants were more likely to select the HFD when the target was absent (Vouloumanos, 2008; Suanda & Namy, 2012). Based on this finding, participants appear to be employing an HT approach.

Studies that support the HT approach have found a sharp rise in accuracy or looks to the target during the learning phase once participants receive confirmation that their hypothesis was correct (Medina et al., 2011; Trueswell et al, 2013). It was only the participants in the interleaved with social cue condition that followed this pattern early in the learning phase. For all other conditions participants only looked to the target faster in the social cue condition in the later trials. This suggests that they did not immediately rely on the social cue immediately in the learning phase and were considering other objects on the screen. This pattern is more characteristic of the AL approach. Aussems and Vogt (2020) had similar findings in a similar paradigm using eye tracking where their participants were aggregating information in the initial trials and only proposed a referent in the later learning trials. However, none of their participants showed a pure HT approach while the participants in our interleaved with social cues condition did.

5.4 Considerations regarding memory demands

This dissertation has mentioned memory in two specific contexts. The first is related to how sound discrimination is directly tied to the demands on working memory placed by the task used. This is directly relevant to study 1 which was compared the AXB and AX task to determine if heavier memory demands in the sound discrimination task led to the discrepancy in the results compared to the word recognition task. The second context is that of CSWL was a process and experimental paradigm itself. A focal point of discussion in CSWL is the number of potential referents the learner has to store presumably across different stages of learning word-referent mappings. Specific to Study 3, the interleaved trial spacing condition has been mentioned to place additional memory demands compared to the consecutive condition. Although memory has often been mentioned in the context of CSWL, there is a dearth of literature to how exactly these two processes interact.

It must be emphasized that study 1 did not include any measures of working memory, so its effects on sound discrimination cannot be commented on. Working memory as a possible contributing factor arose when participants in study 1 showed discrimination performance on the word recognition task but not on the AXB task. Direct comparisons of the AXB task to other tasks involving only same-different judgments found that the AXB presented heavier working memory load (Pisoni & Lazarus, 1974; Gerrits & Schouten, 2004). This was particularly important because acoustic information can decay over time and after that participants would be making a judgment on their memory of the stimuli rather than the actual features themselves. The word recognition task only required participants to recall one token at auditory token at a time which might have presented a lower demand than the AXB, thus motivating the second experiment using the AX. However, this still resulted in chance performance on the AX task while maintaining above chance performance on the word recognition task. Based on these findings, the AXB and AX task presented equal demands for participants in this study although no conclusions regarding working memory can be made.

In study 3, participants in the interleaved condition had less looks to the target in the test phase same trials compared to participants in the consecutive condition. Both groups, however, did manage to correctly map word-referent relationships. Aussems and Vogt (2020) also claimed that the interleaved condition was more challenging compared to the consecutive condition due to memory demands.

However, it is unclear which type of memory they are referring to. One possibility would be to discuss this demand in terms of the Baddeley and Hitch (1974) working memory model. In CSWL, learners would need to store co-occurrence frequencies or a potential hypothesis in short-term memory. On the next encounter, they would need to use working memory to update these statistics or evaluate their current hypothesis. The challenge might arise from coordinating multiple of these hypotheses or co-occurrence frequencies during interleaving since each target word has a unique set of possible referents. A participant in the interleaved condition would require switching across the matching sets of referents or hypothesis for a target word since they are simultaneously learning multiple words. In contrast, a participant in the consecutive condition can simply work with learning one word at a time.

There have not been any studies that have tested working memory in relation to CSWL. There is a finding that the recognition memory abilities of 4-year-olds is the strongest predictor of CSWL performance (Vlach & Debroch, 2017). This finding lends itself to support the view that language is only one component of CSWL and other cognitive mechanisms are in involved. This dissertation is unable to comment on which mechanisms those are. It does agree with the findings of Aussems and Vogt (2020) that interleaved presentations are more challenging than consecutive presentations and it is possible that memory demands have played a role.

The studies in dissertation have tasked participants to hold varying types of information across the auditory and visual modality to perform across phonetic and lexical tasks. It has also challenged them to discriminate between small differences at the phonetic level, retain novel lexical forms, and store co-occurrence frequencies for multiple possible word-object pairings. While these tasks are linguistic in nature, they do require support from associated cognitive processes such as memory. Further work is required to describe the relationships that exist between CSWL as a process, language, and other cognitive systems required to implement it.

5.5 Considerations during internet-based testing

All experiments in this dissertation have been conducted online and show that internet-based studies involving auditory discrimination tasks and webcam-based eye-tracking are convenient and scalable. Despite the limitations placed by the pandemic, the target number of adult participants was easily met within two weeks of recruitment. Supervised testing sessions with children meant that recruitment was slower, but all the families who signed up completed the experiments until the end. There were no dropouts due to technical errors in the middle of the study or children refusing to participate completely.

One concern of unsupervised testing sessions with adults was the reliability of data. Without watching the participants as they were completing the experiment, it is possible that they would not respond to the tasks to the best of their ability or with their full attention. Including attention checks using filler items, like those used in in-person experiments, was effective in ensuring that this was not the case. Adult participants in studies 2 and 3 performed at ceiling for these items. There were also immediate checks as they were performing the experiment, such as being automatically excluded for an excessive number of missed trials, but no one was excluded this way.

There were no attention checks in study 3, but the eye tracking design of the study ensured participants were looking at the screen at all times. The experiment was configured in such a way that if the camera failed to detect the participant in the frame, the experiment would be paused until they returned to their position at calibration. Predicted patterns emerging such as the increase in looks to the target during the learning phase and during the same test trials were also a sign that attention was maintained throughout the task. Overall, reliable data can be obtained from participants without an experimenter present as long as the experiment is designed to check for reliability of data, which is also something in-person experiments should take into account as well.

Supervised testing sessions were implemented for children as an opportunity to collect an online measure, but also to rule out the possibility that parents would respond for their child. Generally, parents had minimal to no issues regarding the technical setup of the experiments. There was a pre-experiment setup that all parents completed successfully and no reports were received regarding errors or difficulties with the setup. Four videos were excluded due to poor video quality, but only one of them was due to a connection problem where the video of the participant froze for a part of the experiment. The remaining three had issues that could be resolved by asking parents to change the webcam they were using to a front-facing one or move to a room with better lighting conditions. These could have been avoided by including video demonstrations or diagrams in the set-up instructions.

The remainder of the videos excluded were due to children not being visible in the frame. Interrater reliability on the direction of looks was at 73% which is lower than lab-based studies. One challenge in achieving a degree of interrater reliability comparable to that of lab-based studies is the lack of information on the size of the screens participants were using. Determining the looking direction is aided by knowledge of the general physical area in which stimuli is contained. As the setup is fairly uniform in the lab and experimenters have an opportunity to see the actual screen, this serves as a standard for raters to judge the looks. In this experiment, a standard screen size was not prescribed for participants. Even then this would be difficult to ensure given that the experimental platform only records screen resolution which can be shared among different monitor sizes. One way to address this would be to have raters train on videos taken in the lab and then require participants to use a similar size of screen at home. The eligibility of devices can be assessed by asking participants to provide the specifications or model of their computer or external monitors for experimenters to double check.

5.6 Limitations and Implications

The COVID pandemic has impacted this dissertation in terms of recruitment, methodology, and structure of the thesis. In terms of recruitment, the pandemic created significant barriers to in-person data collection due to social distancing measures leading to closure of labs. This resulted in the decision to pursue internet-based testing, but there were limited findings on how well this was suited to testing infants in children. It was decided that data from adults and pre-school aged children was more feasible, despite the questions in this dissertation having more significance in infants.

The difficulty in recruitment also affected the structure of the experimental chapters. There were three experiments dedicated to encoding novel phonetic information during CSWL and only one for CSWL and social cues. The original plan was to have one experiment each to investigate in infants how CSWL integrated with three key domains in early linguistic development: phonetic, syntactic, and social. There was an attempt to maintain as many of these original plans as possible but with the practical consideration that enough data needed to be gathered to produce three experimental chapters. This led to implementing variations of the first experiment in adults to collect enough data for the dissertation but with plans to eventually test infants once restrictions were lifted. By the time in-person testing was possible, time permitted for only one more experiment to be conducted. This coincided with the lab facing a backlog of recruitment for several studies also testing infants. Pursuing a younger population for the social cues experiment would have delayed data collection due to the competition. Thus, adults were again recruited for the final experiment. Therefore, most of the experiments tested adults and focused on the tracking of phonetic information during CSWL.

This dissertation placed a large importance on studying word learning as it integrates with phonetic and social domains due to the close relationship of these three in early development, but its findings do not directly generalize to early L1 language acquisition. Both adults and children in this experiment had to contend with the influences of their L1 to discriminate a novel contrast, whereas infants are not faced with such a large interference. By the time they are ready to engage in word learning, they have also just begun attuning to these categories. In terms of eye gaze as a social cue, adults already have knowledge of its referential intent while children below two years old seem to have difficulty using eye gaze to link words to referents (Hollich et al., 2000). There is a difference in sensitivity of adults versus infants to phonetic and social information derived from their respective linguistic experiences which allows limited generalization of findings to early development.

There are study-specific limitations within the phonetic learning experiments. For one, there was no comprehensive data on participants' L2 learning background. This dissertation only ensured that participants spoke German natively and self-reported no prior exposure to Hindi. Ideally, prior L2 learning experience would have been controlled for. However, the retroflex is uncommon in European languages and is concentrated in languages in the Indian subcontinent. This makes it unlikely that the participants have encountered it frequently. If any participant had experience with a language with phonology similar to Hindi, this could have offered them an advantage during learning. However, retroflex consonants are rare in European languages and are only found in 20% of the word's languages mostly in the Indian subcontinent.

On another note, there were no implicit measures of discrimination abilities collected for adults. Looking data from the children offered the advantage of obtaining their immediate responses to the stimuli. This would also have been useful for the adults considering that working memory resource limitations have been discussed as a factor affecting their responses. The studies do not have completely comparable looking data measures between the adults and children.

Since all of the studies are training studies, the next limitation applies to all of them. This is the effect of time during and after learning. All experiments trained participants and tested learning in less than an hour. This is relatively short compared to studies where participants are provided multiple rounds of training sometimes spread across days or weeks. The intensity of the training is a significant factor especially for the first two studies considering that the Hindi dental-retroflex is notoriously difficult to perceive (Tees & Werker, 1984, Werker & Tees, 1983, Werker & Tees, 1984, Werker et al., 1981).

Training at an additional intensity might also lead to more lasting improvements or can provide more descriptive findings on how learning takes shape across time instead of a binary outcome of whether participants have learned or not. The last study with social cues does do this to some extent by tracking eye movements during the learning phase. Introducing longer delays in between training and testing can also provide some insight on how periods of consolidation affect learning outcomes. As Fuhrmeister et al. (2020) found, children showed better discrimination outcomes in adults only after a 24-hour period of delay, but immediate testing showed that adults have the advantage. The immediate train and test design used in this dissertation does not offer insight on how the process and outcomes of CSWL change across time.

Finally, the studies in this dissertation are limited in how they reflect the naturalistic word learning setting. The first issue is in terms of co-occurrence. In these experiments, the target was assigned a 100% co-occurrence with the word but experience tells us that this is not likely the case in real life. Speakers are likely to reference objects not in the immediate setting. As Frank et al. (2013) have found in encoding videos of mother-infant interactions, co-occurrence frequencies between words and targets were not perfect. Distractors also had very close co-occurrence frequencies to the target, creating much competition that a computational model failed to extract the correct pairings based on co-occurrences alone. It is likely that in naturalistic settings, learners have to make do with imperfect co-occurrence frequencies when attempting word mapping.

The second is in how objects in these experiments are neatly placed into equal-sized sections devoid of any other contextual information. This of course is done to achieve experimental control, but it is very far from how learners encounter objects when learning their labels. Objects appear in scenes with other contextually relevant objects and would vary in their spatial arrangement. Some would appear more in the foreground, appearing more salient in the field of view, while other visible objects would be in the background. This in some way would affect how different objects would be considered as likely referents.

While experiments would seek to control for extraneous cues like these, it is important to acknowledge that this comes at the cost of making it more difficult to generalize findings to naturalistic settings. The more control is exercised in experimental settings, the less comparable it is to what learners are facing in their environment. While the findings of this dissertation show that phonetic information can be tracked alongside co-occurrence frequencies and that social cues support CSWL, it is limited in its ability to conclude that learners are engaging in this strategy in real-life word learning situations.

5.7 Future Directions

From its inception, the research questions in this dissertation were formulated in reference to first language acquisition. One recommendation would be to adapt the phonetic and social cue CSWL experiments for 12-month-old infants. This is the youngest age that CSWL learning has been observed in (Smith & Yu, 2008) and at this age infants are already past the perceptual narrowing window. At this age, infants might still be flexible in learning a non-native contrast and could be able to combine with their statistical learning skills. It could provide further insight on how plausible it is for statistical learning to kickstart word-referent mapping by investigating if it can take place alongside the essential task of attuning to native phoneme categories.

An extension of this would be to recruit 20-month-olds to compare their performance to 12-month-olds. Infants younger than 20 months old were found unable to discriminate a native contrast in a word learning task, despite being able to discriminate it in a perception task (Stager & Werker, 1997). It has not yet been tested if this limitation would extend to non-native contrasts and, if so, will it be overcome at the same age.

Investigating the developmental change in how infants use eye gaze during CSWL could give insight on how effective a social cue is in supporting early word learning development. In Hollich et al.'s (2000) work, 12-month-olds assigned a novel label to an object with salient perceptual characteristics rather than an object where a speaker was looking to. It was only when a combination of touching, handling, looking, and labeling the object that infants at this age assigned a label to that item. One question to ask would be whether 12-month-old infants' attention would be focused on perceptual salience over co-occurrence frequencies. It would also be interesting to compare if pairing co-occurrence frequencies with a complementary eye gaze cue creates faster or improved learning outcomes in infants across 12, 18, and 24 months. This provides a longitudinal aspect to knowledge on the topic. Not only do these ages almost exactly align with those in Hollich et al.'s (2000), but they also represent other milestones in word learning. The "vocabulary spurt" was observed to occur at around 18 months (Bloom, 1973; Goldfield & Reznick, 1990) and might reflect some changes in how infants further improve their ability to discover word meanings. Meanwhile, the findings that support how infants leverage understanding of referential intent to assign labels to objects was mostly observed in 24-month-olds (Tomasello & Akhtar, 1995). These suggestions emerge from the idea that while the mechanism of CSWL is a plausible way to get word learning started, infants are eventually able to make use of more cues to further their word learning abilities. This then leads to the question of how the use of these cues emerge and change across development.

There could also be some benefit to make CSWL experiments more naturalistic. Many of the current CSWL studies, including those in this dissertation, present objects neatly isolated across quadrants against monochromatic backgrounds. While sufficient experimental control is achieved and participants do show learning in these paradigms, it becomes harder to conclude that this is similar to what is occurring in naturalistic development since the learning situations are so different. Infants encounter objects in an environment with multiple objects and a speaker present providing the input.

One way to make CSWL experiments more natural would be to present learning trials as videos of a visible speaker naming objects in front of them placed on either side of the screen. Even in the screen recordings that we did in study 2, it was still possible to tell the direction of a child’s look more than half of the time. Currently, more naturalistic versions of CSWL are those which use vignettes from video corpora of parent-child interactions to present learning instances. Instead of observing the learning of another, head mounted eye tracking could be useful in gathering actual first-person point of view during live learning situations. It needs to be considered how a balance between experimental control and a representative learning environment can be achieved to further support CSWL as the mechanism for early word learning.

One can also make learning situations more naturalistic by manipulating co-occurrence frequencies specifically of the referent so that it does not have a perfect correspondence with the label. Analysis of parent-child interaction videos show that a correct co-occurrence cannot be established for 30% of a parent’s utterance to a child, because the referent was absent when a label was spoken (Kyger, 2013; Harris, Jones, & Grant, 1984). One of the issues not directly answered by any account of CSWL is what degree of support from co-occurrence frequencies or referential cues could be considered enough for a learner to decide on a correct referent. Alternatively, this also concerns when does a learner start dropping proposed referents that are not supported by the current input from the environment. Investigating learning situations with more competition from distractors due to less consistent co-occurrences between the correct referent and label could provide some insights to this.

There is also an opportunity to adapt a more qualitative method to measuring speech perception in L2 learning in adults. Measures of speech perception are currently binary: it is either the participant hears the difference or not. However, there is little information on what they hear and how this changes with learning. This is a key proposal in the L2LP learning: that learners will adapt their L1 category to resemble the features of the L2 but the current sound discrimination tasks do not capture this adaptation. The task used by Joseph et al. (2015) to measure verbal working memory capacity could be adapted as a test of speech perception. In their study, participants were played a sound and were asked to replicate the sound they just heard. This is done by turning a dial that plays a sound token and adjusts its F1 and F2 dimensions as one turns it. This provides a quantitative measure to describe what acoustic properties participants are perceiving in a non-native contrast and test if it is systematically approaching features of the L2 token. This can also be adapted for use with children if introduced as a game. Using a test of speech perception like this could provide a descriptive trajectory of L2 speech learning.

All the studies in this dissertation had the goal of capturing the results of learning but was limited to an immediate test. More data on learning outcomes and retention by implementing experiments that last across multiple sessions would be useful for ensuring that robust learning and not temporary associations are being formed. This is especially important for Studies 1 and 2 considering that the advantage that children have over adults in phonetic learning appears in the long-term (Piske et al., 2001; Flege et al., 1999). Gathering more longitudinal data suggested above can be more possible through the next recommendation which is the use of internet-based testing.

The studies in this dissertation support the reduced time dedicated to data collection both for researchers and participants afforded by internet-based testing. Based on this experience, it is recommended that internet-based testing not be considered as a temporary phase that was required by the pandemic. Rather, it is a valid and viable option for paradigms that this dissertation has used where millisecond reaction time-based measures were not required or areas of interest on the screen were large enough for a webcam to accurately track eye movements of participants who can sit still. Internet-based testing paves the way for recruiting larger groups of participants, people who are not in geographical proximity to labs or universities, or those who do not have the time or resources to participate in in-person experiments. Internet-based testing improves accessibility for both researchers and participants alike.

The recommendations put forth by this dissertation center around looking at the bigger picture in several ways. It suggests looking at how different cues present in the input combine to produce a successful word learner rather than examining each one separately. It also places incorporating the nuances of a naturalistic learning environment in equal consideration as achieving experimental control. The suggestions also account for how language learning is a process that happens across time and suggests ways that the trajectory of learning could be more descriptively captured. Finally, it also encourages expanding the knowledge of what methods are available to researchers as internet-based testing broadens the opportunities for more people to participate in science.

5.8 Conclusion

One of the main questions of this dissertation is regarding the ability to learn novel phonetic information when this is embedded in minimal pair words presented during a CSWL experiment. The findings showed that adults to some extent can simultaneously track novel phonetic information while learning word-referent mappings in a CSWL experiment. Adults showed the ability to correctly identify the referents of minimal pair words containing a non-native contrast after being presented co-occurrence frequencies of these pairings after approximately three minutes of training. The first study found a discrepancy between measures of pure discrimination and the more referential identification task. Adults were performing at chance on the discrimination tasks but managed to perform above chance on the identification task. While evidence does exist that the ability to perceive phonetic contrasts affects success in learning during CSWL experiments, the findings of this dissertation suggest that the lexical and referential contexts assist adults in learning to attune to these fine-grained phonetic differences.

In contrast, the four-year-old group did not appear to benefit from the CSWL-based training. Results indicated that the difficulty in perceiving the novel contrast have blocked their ability to learn word-referent pairings. However, previous work lab-based training studies with naïve listeners found that improvements in children's performance begin to emerge over multiple training sessions where they might even surpass adults. Therefore, more longitudinal investigations would be needed to confirm the effectiveness of CSWL experiments in training non-native speech perception in children.

The second focus of this research is to manipulate the difficulty of the learning context to determine whether the associative learning account or hypothesis testing account better describes observed performance in adults during CSWL. While previous work has focused on creating more challenging learning contexts through increasing number of possible referents, this study used trial spacing to manipulate the complexity of learning. In addition to this, it explored the interaction of trial spacing with eye gaze, a supportive cue that has strong ties to referential intent and is available in naturalistic language learning situations.

The results show that participants did not show memory for any other referents apart from the target during the test phase regardless of spacing condition. Previous worked would have interpreted this finding as fully in support of hypothesis testing, however looking behavior during the learning phase showed a gradual increase in the target across conditions. The lack of immediate preference would be more in line with associative learning accounts. The only exception to this was the interleaved spacing condition with a social cue wherein participants developed a preference of the target in earlier trials in the learning phase. While participants appear to be conservative in choosing a referent, the combination of a more challenging learning condition paired with a strong indicator of a referential relationship was sufficient for participants to propose a target earlier. An implication for future CSWL research based on these findings is to consider not only the challenges that stem from a noisy learning space but equally consider how supportive cues could affect strategies learners employ.

References

Akhtar, N. (2005). Is joint attention necessary for early language learning? In B. D. Homer & C. S. Tamis-LeMonda (Eds.), *The development of social cognition and communication* (pp. 165–179). Lawrence Erlbaum Associates Publishers.

Akhtar, N., & Gernsbacher, M. A. (2007). Joint Attention and Vocabulary Development: A Critical Look. *Language and Linguistics Compass*, 1(3), 195–207. https://doi.org/10.1111/j.1749-818X.2007.00014.x

Akhtar, N., & Tomasello, M. (1998). Intersubjectivity in early language learning and use. In *Intersubjective communication and emotion in early ontogeny* (pp. 316–335). Cambridge University Press.

Akhtar, N., & Tomasello, M. (2000). The Social Nature of Words and Word Learning. In R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Becoming a Word Learner: A Debate on Lexical Acquisition* (pp. 115–135). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195130324.003.005

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020a). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Anwyl-Irvine, A., Dalmaijer, E., Hodges, N., & Evershed, J. (2020b). *Online Timing Accuracy and Precision: A comparison of platforms, browsers, and participant's devices*. PsyArXiv. https://doi.org/10.31234/osf.io/jfeca

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2003). *Foreign Accent in English Words Produced by Japanese Children and Adults*. 4.

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602

Aslin, R. N., & Pisoni, D. B. (1980). Effects of Early Linguistic Experience on Speech Discrimination by Infants: A Critique of. *Child Development*, *51*(1), 107–112.

Au, T. K., & Glusman, M. (1990). The Principle of Mutual Exclusivity in Word Learning: To Honor or Not to Honor? *Child Development*, 61(5), 1474–1490. https://doi.org/10.1111/j.1467-8624.1990.tb02876.x

Aussems, S., & Vogt, P. (2020). Adults Use Cross-Situational Statistics for Word Learning in a Conservative Way. *IEEE Transactions on Cognitive and Developmental Systems*, *12*(2), 232–242. https://doi.org/10.1109/TCDS.2018.2870161

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Baker, W., Trofimovich, P., Mack, M., & Flege, J.E. (2006). The effect of perceived phonetic similarity on non-native sound learning by children and adults. *BUCLD 26 Proceedings*, 36-47.

Baldwin, D. A., & Moses, L. J. (2001). Links between Social Understanding and Early Word Learning: Challenges to Current Accounts. *Social Development*, *10*(3), 309–329. https://doi.org/10.1111/1467-9507.00168

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' Reliance on a Social Criterion for Establishing Word-Object Relations. *Child Development*, *67*(6), 3135–3153. https://doi.org/10.2307/1131771

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Benitez, V. L., Yurovsky, D., & Smith, L. B. (2016). Competition between multiple words for a referent in cross-situational word learning. *Journal of Memory and Language*, *90*, 31–48. https://doi.org/10.1016/j.jml.2016.03.004

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., Pulsifer, P., Beaver, D. I., Chelliah, S., Dubinsky, S., Meier, R. P., Thieberger, N., Rice, K., & Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, *56*(1), 1–18. https://doi.org/10.1515/ling-2017-0032

Best, C. T. (1995) A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research* (pp. 167– 200). York Press.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345–360. https://doi.org/10.1037/0096-1523.14.3.345

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language Learning & Language Teaching* (Vol. 17, pp. 13–34). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.17.07bes

Bloom, L. (1973). *One word at a time: The use of single word utterances before syntax*. Mouton.

Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.10, retrieved 24 May 2023 from http://www.praat.org/

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and Me: Familiar Names Help Launch Babies into Speech-Stream Segmentation. *Psychological Science*, *16*(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Boxtel, S. van, Bongaerts, T., & Coppen, P.-A. (2003). Native-like attainment in L2 syntax. *EUROSLA Yearbook*, *3*(1), 157–181. https://doi.org/10.1075/eurosla.3.10box

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

Brekelmans, G. (2020). *Phonetic vowel training for child second language learners: the role of input variability and training task*. [Doctoral dissertation, University College London]. UCL Discovery.

Brooks, R., & Meltzoff, A. N. (2002). The importance of eyes: How infants interpret adult looking behavior. *Developmental Psychology*, *38*(6), 958–966. https://doi.org/10.1037/0012-1649.38.6.958

Brooks, R., & Meltzoff, A. N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modeling study. *Journal of Child Language*, *35*(1), 207–220. https://doi.org/10.1017/S030500090700829X

Bunce, J. P., & Scott, R. M. (2017). Finding meaning in a noisy world: Exploring the effects of referential ambiguity and competition on 2·5-year-olds' cross-situational word learning. *Journal of Child Language*, 44(3), 650–676. https://doi.org/10.1017/S0305000916000180

Bunce, J. P., Gordon, C. L., Abney, D. H., Fleming, M. M., Greenwood, M. D., Chiu, E., Spivey, M. J., & Scott, R. M. (2016). *Mouse Tracking Reveals Knowledge of Multiple Competing Referents During Cross-situational Word Learning*.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social Cognition, Joint Attention, and Communicative Competence from 9 to 15 Months of Age. *Monographs of the Society for Research in Child Development*, *63*(4), i–174. https://doi.org/10.2307/1166214

Çetinçelik, M., Rowland, C. F., & Snijders, T. M. (2021). Do the Eyes Have It? A Systematic Review on the Role of Eye Gaze in Infant Language Development. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.589096

Chen, C., & Yu, C. (2015). The Effects of Learning and Retrieval Contexts on Cross-situational Word Learning. IEEE International Conference on Development and Learning and Epigenetic Robotics: [Proceedings]. IEEE International Conference on Development and Learning and Epigenetic Robotics, 2015, 202–207. https://doi.org/10.1109/DEVLRN.2015.7346141

Cooper, R. P., & Aslin, R. N. (1990). Preference for Infant-Directed Speech in the First Month after Birth. *Child Development*, *61*(5), 1584–1595. https://doi.org/10.2307/1130766

Curtin, S., Goad, H., & Pater, J. V. (1998). Phonological transfer and levels of representation: The perceptual acquisition of Thai voice and aspiration by English and French speakers. *Second Language Research*, *14*(4), 389–405. https://doi.org/10.1191/026765898674095369

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics*, *34*(2), 269–284. https://doi.org/10.1016/j.wocn.2005.06.002

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 892–903. https://doi.org/10.1037/a0035657

Davidson, L., Shaw, J., & Adams, T. (2007). The effect of word learning on the perception of non-native consonant sequences. *The Journal of the Acoustical Society of America*, *122*(6), 3697–3709. https://doi.org/10.1121/1.2801548

Delaney, P. F., Spirgel, A. S., & Toppino, T. C. (2012). A deeper analysis of the spacing effect after “deep” encoding. *Memory & Cognition*, *40*(7), 1003–1015. https://doi.org/10.3758/s13421-012-0207-3

Delle Luche, C., Durrant, S., Poltrock, S., & Floccia, C. (2015). A methodological investigation of the Intermodal Preferential Looking paradigm: Methods of analyses, picture selection and data rejection criteria. *Infant Behavior and Development*, *40*, 151–172. https://doi.org/10.1016/j.infbeh.2015.05.005

Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2014). Prelinguistic Infants Are Sensitive to Space-Pitch Associations Found Across Cultures. *Psychological Science*, *25*(6), 1256–1261. https://doi.org/10.1177/0956797614528521

Dye, C., Kedar, Y., & Lust, B. (2019). From lexical to functional categories: New foundations for the study of language development. *First Language*, *39*(1), 9–32. https://doi.org/10.1177/0142723718809175

Eglington, L. G., & Kang, S. H. K. (2017). Interleaved Presentation Benefits Science Category Learning. *Journal of Applied Research in Memory and Cognition*, *6*(4), 475–485. https://doi.org/10.1016/j.jarmac.2017.07.005

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science (New York, N.Y.)*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

ELAN (Version 6.5) [Computer software]. (2023). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

Escudero, P., & Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition*, *133*(2), 408–413. https://doi.org/10.1016/j.cognition.2014.07.002

Escudero, P., Mulak, K. E., & Vlach, H. A. (2016a). Cross-Situational Learning of Minimal Word Pairs. *Cognitive Science*, *40*(2), 455–465. https://doi.org/10.1111/cogs.12243

Escudero, P., Mulak, K. E., & Vlach, H. A. (2016b). Infants Encode Phonetic Detail during Cross-Situational Word Learning. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01419

Escudero, P., Mulak, K. E., Fu, C. S. L., & Singh, L. (2016). More Limitations to Monolingualism: Bilinguals Outperform Monolinguals in Implicit Word Learning. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01218

Esteve-Gibert, N., & Muñoz, C. (2021). Preschoolers benefit from a clear sound-referent mapping to acquire nonnative phonology. *Applied Psycholinguistics*, *42*(1), 77–100. https://doi.org/10.1017/S0142716420000600

Feldman, N. H., Myers, E. B., White, K. S., Griffiths, T. L., & Morgan, J. L. (2013). Word-level information influences phonetic learning in adults and infants. *Cognition*, *127*(3), 427–438. https://doi.org/10.1016/j.cognition.2013.02.007

Flege, J. E. (1992). Speech Learning in a Second Language. In Ferguson, L. Menn, and C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research, Implications* (pp. 565–604). York Press).

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 92, 233-277.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age Constraints on Second-Language Acquisition. *Journal of Memory and Language*, *41*(1), 78–104. https://doi.org/10.1006/jmla.1999.2638

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. *Psychological Science*, *20*(5), 578–585. https://doi.org/10.1111/j.1467-9280.2009.02335.x

Frank, M. C., Tenenbaum, J. B., & Fernald, A. (2013). Social and Discourse Contributions to the Determination of Reference in Cross-Situational Word Learning. *Language Learning and Development*, *9*(1), 1–24. https://doi.org/10.1080/15475441.2012.707101

Freeman, M. R., Blumenfeld, H. K., Carlson, M. T., & Marian, V. (2021). First-language influence on second language speech perception depends on task demands. *Language and Speech*, 0023830920983368. https://doi.org/10.1177/0023830920983368

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448–EL454. https://doi.org/10.1121/1.5009688

Fuhrmeister, P., Schlemmer, B., & Myers, E. B. (2020). Adults Show Initial Advantages Over Children in Learning Difficult Nonnative Speech Sounds. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2667–2679. https://doi.org/10.1044/2020_JSLHR-19-00358

Gangwani, T., Kachergis, G., & Yu, C. (2010). Simultaneous Cross-situational Learning of Category and Object Names. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *32*(32). https://escholarship.org/uc/item/738979fp

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & Psychophysics*, *66*(3), 363–376. https://doi.org/10.3758/BF03194885

Gerson, S. A., Simpson, E. A., & Paukner, A. (2017). Drivers of social cognitive development in human and non-human primate infants. In *Social cognition: Development across the life span* (pp. 98–128). Routledge/Taylor & Francis Group.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176. https://doi.org/10.1016/S0010-0277(99)00036-0

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, *1*(1), 3–55. https://doi.org/10.1207/s15327817la0101_2

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard Words. *Language Learning and Development*, *1*(1), 23–64. https://doi.org/10.1207/s15473341lld0101_4

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, *17*(1), 171–183. https://doi.org/10.1017/S0305000900013167

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166

Golinkoff, R. M., & Hirsh-Pasek, K. (2000). *Becoming a Word Learner*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195130324.001.0001

Golinkoff, R. M., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-Five Years Using the Intermodal Preferential Looking Paradigm to Study Language Acquisition: What Have We Learned? *Perspectives on Psychological Science*, *8*(3), 316–339. https://doi.org/10.1177/1745691613484936

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. *Speech Communication*, *50*(2), 109–125. https://doi.org/10.1016/j.specom.2007.07.003

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, 87(1), B23–B34. https://doi.org/10.1016/S0010-0277(02)00186-5

Hall, D. G., Waxman, S. R., & Hurwitz, W. M. (1993). How two- and four-year-old children interpret adjectives and count nouns. *Child Development*, 64, 1651–1664. https://doi.org/10.2307/1131461

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88. https://doi.org/10.1016/j.cognition.2019.03.011

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877. https://doi.org/10.3758/s13421-018-0808-6

Harris, M., Jones, D., & Grant, J. (1984). The social-interactional context of maternal speech to infants: An explanation for the event-bound nature of early word use? *First Language*, *5*(14), 89–99. https://doi.org/10.1177/014272378400501401

Hartley, C., Bird, L.-A., & Monaghan, P. (2020). Comparing cross-situational word learning, retention, and generalisation in children with autism and typical development. *Cognition*, *200*, 104265. https://doi.org/10.1016/j.cognition.2020.104265

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

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1), 65–94. https://doi.org/10.1177/0267658307071601

Heeren, W. F. L., & Schouten, M. E. H. (2010). Perceptual development of the Finnish /t-tː/ distinction in Dutch 12-year-old children: A training study. *Journal of Phonetics*, *38*(4), 594–603. https://doi.org/10.1016/j.wocn.2010.08.005

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., Hennon, E., Rocroi, C., & Bloom, L. (2000). Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning. *Monographs of the Society for Research in Child Development*, *65*(3), 85-100.

Horst, J. S. (2013). Context and repetition in word learning. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00149

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*, 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language Experience Shapes the Development of the Mutual Exclusivity Bias. *Infancy*, *15*(2), 125–150. https://doi.org/10.1111/j.1532-7078.2009.00009.x

Höhle, B., & Weissenborn, J. (2000). The origins of syntactic knowledge: Recognition of determiners in one old German children. https://www.semanticscholar.org/paper/The-origins-of-syntactic-knowledge-%3A-recognition-of-H%C3%B6hle-Weissenborn/a1788d67279eee64df34b443199cc8d550026158

Hu, C.F. (2017). Resolving referential ambiguity across ambiguous situations in young foreign language learners. *Applied Psycholinguistics*, *38*(3), 633–656. https://doi.org/10.1017/S0142716416000357

Jaswal, V. K., & Hansen, M. B. (2006). Learning words: Children disregard some pragmatic information that conflicts with mutual exclusivity. *Developmental Science*, 9(2), 158–165. https://doi.org/10.1111/j.1467-7687.2006.00475.x

Joseph, S., Iverson, P., Manohar, S., Fox, Z., Scott, S. K., & Husain, M. (2015). Precision of working memory for speech sounds. *Quarterly Journal of Experimental Psychology*, *68*(10), 2022–2040. https://doi.org/10.1080/17470218.2014.1002799

Jusczyk, P. W., Hohne, E. A., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, *61*(8), 1465–1476. https://doi.org/10.3758/BF03213111

Kabakoff, H., Go, G., & Levi, S. V. (2020). Training a non-native vowel contrast with a distributional learning paradigm results in improved perception and production. *Journal of Phonetics*, *78*, 100940. https://doi.org/10.1016/j.wocn.2019.100940

Kachergis, G., & Yu, C. (2014). Continuous measure of word learning supports associative model. *4th International Conference on Development and Learning and on Epigenetic Robotics*, 20–25. https://doi.org/10.1109/DEVLRN.2014.6982949

Kachergis, G., Shiffrin, R., & Yu, C. (2009). Frequency and Contextual Diversity Effects in Cross-Situational Word Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31). https://escholarship.org/uc/item/1jh968zh

Kachergis, G., Yu, C., & Shiffrin, R. M. (2014). Cross-situational word learning is both implicit and strategic. *Frontiers in Psychology*, *0*. https://doi.org/10.3389/fpsyg.2014.00588

Kalashnikova, M., Mattock, K., & Monaghan, P. (2016). Mutual exclusivity develops as a consequence of abstract rather than particular vocabulary knowledge. *First Language*, *36*(5), 451–464. https://doi.org/10.1177/0142723716648850

Kedar, Y., Casasola, M., & Lust, B. (2006). Getting There Faster: 18- and 24-Month-Old Infants' Use of Function Words to Determine Reference. *Child Development*, *77*(2), 325–338. https://doi.org/10.1111/j.1467-8624.2006.00873.x

Kedar, Y., Casasola, M., Lust, B., & Parmet, Y. (2017). Little Words, Big Impact: Determiners Begin to Bootstrap Reference by 12 Months. *Language Learning and Development*, *13*(3), 317–334. https://doi.org/10.1080/15475441.2017.1283229

Kibbe, M. M., & Leslie, A. M. (2013). What's the object of object working memory in infancy? Unraveling 'what' and 'how many.' *Cognitive Psychology*, *66*(4), 380–404. https://doi.org/10.1016/j.cogpsych.2013.05.001

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. https://doi.org/10.1177/01427237211066405

Kim, S. K., & Webb, S. (2022). The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis. *Language Learning*, *72*(1), 269–319. https://doi.org/10.1111/lang.12479

Koehne, J., Trueswell, J. C., & Gleitman, L. R. (2013). Multiple Proposal Memory in Observational Word Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *35*(35). https://escholarship.org/uc/item/5gp8j5qs

Koehne, J., & Crocker, M. W. (2015). The Interplay of Cross-Situational Word Learning and Sentence-Level Constraints. *Cognitive Science*, *39*(5), 849–889. https://doi.org/10.1111/cogs.12178

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science (New York, N.Y.)*, *255*(5044), 606–608. https://doi.org/10.1126/science.1736364

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early Speech Perception and Later Language Development: Implications for the “Critical Period.” *Language Learning and Development*, *1*(3–4), 237–264. https://doi.org/10.1080/15475441.2005.9671948

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Kyger, M. F. (2013). Parents Provide Children with Social Cues for Word Learning. *UCLA*. ProQuest ID: Kyger_ucla_0031D_11482. Merritt ID: ark:/13030/m5cz4kqd. Retrieved from https://escholarship.org/uc/item/27h4d4t8

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, 3(3), 299–321. https://doi.org/10.1016/0885-2014(88)90014-7

Lazaridou, A., Chrupala, G., Fernández, R., & Baroni, M. (2016). Multimodal Semantic Learning from Child-Directed Input. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 387–392. https://doi.org/10.18653/v1/N16-1043

Lenneberg, E. H. (1967). *Biological foundations of language*. Wiley.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368. https://doi.org/10.1037/h0044417

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology*, *94*, 67–84. https://doi.org/10.1016/j.cogpsych.2017.02.003

Markman, E. M. (1989). *Categorization and Naming in Children: Problems of Induction*. MIT Press.

Markman, E. M. (1992). Constraints on Word Learning: Speculations About Their Nature, Origins, and Domain Specificity. In *Modularity and Constraints in Language and Cognition*. Psychology Press.

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, 47(3), 241–275. https://doi.org/10.1016/S0010-0285(03)00034-3

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*(1), 15–35. https://doi.org/10.1016/0167-6393(83)90061-4

Mather, E., & Plunkett, K. (2009). Learning Words Over Time: The Role of Stimulus Repetition in Mutual Exclusivity. *Infancy*, *14*(1), 60–76. https://doi.org/10.1080/15250000802569702

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces: Language and Faces. *Developmental Psychobiology*, *56*(2), 154–178. https://doi.org/10.1002/dev.21177

Maye, J. & Gerken, L. (2000). Learning phonemes without minimal pairs. In *Proceedings of the 24th Annual Boston University Conference on Language Development*, 522-533.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134. https://doi.org/10.1111/j.1467-7687.2007.00653.x

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & Mcclelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(2), 89–108. https://doi.org/10.3758/CABN.2.2.89

McGillion, M. L., Herbert, J. S., Pine, J. M., Keren-Portnoy, T., Vihman, M., & Matthews, D. E. (2013). Supporting early vocabulary development: What sort of responsiveness matters? *IEEE Tranactions on Autonomous Mental Development*, *5*(3), 240–248. https://doi.org/10.1109/TAMD.2013.2275949

McGregor, K. K., Smolak, E., Jones, M., Oleson, J., Eden, N., Arbisi-Kelm, T., & Pomper, R. (2022). What Children with Developmental Language Disorder Teach Us About Cross-Situational Word Learning. *Cognitive Science*, *46*(2). https://doi.org/10.1111/cogs.13094

McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., & Van Engen, K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *Attention, Perception, & Psychophysics*, *80*(6), 1559–1570. https://doi.org/10.3758/s13414-018-1537-4

McRoberts, G. W., McDonough, C., & Lakusta, L. (2009). The Role of Verbal Repetition in the Development of Infant Speech Preferences From 4 to 14 Months of Age. *Infancy: The Official Journal of the International Society on Infant Studies*, *14*(2), 162–194. https://doi.org/10.1080/15250000802707062

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019. https://doi.org/10.1073/pnas.1105040108

Meltzoff, A. N., & Brooks, R. (2009). Social cognition and language: The role of gaze following in early word learning. In J. Colombo, P. McCardle, & L. Freund, *Infant pathways to language: Methods, models, and research disorders* (pp. 169–194). Psychology Press.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Mirman, D. (2016). *Growth Curve Analysis and Visualization Using R*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315373218

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics*, *18*(5), 331–340. https://doi.org/10.3758/BF03211209

Monaghan, P. (2017). Canalization of Language Structure From Environmental Constraints: A Computational Model of Word Learning From Multiple Cues. *Topics in Cognitive Science*, *9*(1), 21–34. https://doi.org/10.1111/tops.12239

Monaghan, P., & Mattock, K. (2012). Integrating constraints for learning word–referent mappings. *Cognition*, *123*(1), 133–143. https://doi.org/10.1016/j.cognition.2011.12.010

Monaghan, P., Brand, J., Frost, R., & Taylor, G. (2017). Multiple variable cues in the environment promote accurate and robust word learning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society, CogSci 2017, London, UK, 16-29 July 2017*. cognitivesciencesociety.org. https://mindmodeling.org/cogsci2017/papers/0164/index.html

Monaghan, P., Mattock, K., Davies, R. A. I., & Smith, A. C. (2015). Gavagai Is as Gavagai Does: Learning Nouns and Verbs From Cross-Situational Statistics. *Cognitive Science*, *39*(5), 1099–1112. https://doi.org/10.1111/cogs.12186

Mora, Joan C. (2005). Lexical knowledge effects on the discrimination of non-native phonemic contrasts in words and nonwords by Spanish/catalan bilingual learners of English. In *PSP2005*, 43-46.

Mundy, P., & Newell, L. (2007). Attention, Joint Attention, and Social Cognition. *Current Directions in Psychological Science*, *16*(5), 269–274. https://doi.org/10.1111/j.1467-8721.2007.00518.x

Nakata, T., & Suzuki, Y. (2019). Mixing Grammar Exercises Facilitates Long-Term Retention: Effects of Blocking, Interleaving, and Increasing Practice. *The Modern Language Journal*, *103*(3), 629–647. https://doi.org/10.1111/modl.12581

Nelson, K. (1988). Constraints on word learning? *Cognitive Development*, *3*(3), 221–246. https://doi.org/10.1016/0885-2014(88)90010-X

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), [aac4716]. https://doi.org/10.1126/science.aac4716

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*(3), 261–283. https://doi.org/10.1007/BF01067377

Pallier, C., Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*(3), B9–B17. https://doi.org/10.1016/S0010-0277(97)00030-9

Pan, S. C., Tajran, J., Lovelett, J., Osuna, J., & Rickard, T. C. (2019). Does interleaved practice enhance foreign language learning? The effects of training schedule on Spanish verb conjugation skills. *Journal of Educational Psychology*, *111*, 1172–1188. https://doi.org/10.1037/edu0000336

Peñaloza, C., Mirman, D., Cardona, P., Juncadella, M., Martin, N., Laine, M., & Rodríguez-Fornells, A. (2017). Cross-situational word learning in aphasia. *Cortex*, *93*, 12–27. https://doi.org/10.1016/j.cortex.2017.04.020

Pew Research Center. (2019). Mobile Connectivity in Emerging Economies. Retrieved from https://www.pewresearch.org/internet/2019/03/07/mobile-connectivity-in-emerging-economies/#fn-22186-1

Pi, Z., Zhang, Y., Zhu, F., Xu, K., Yang, J., & Hu, W. (2019). Instructors' pointing gestures improve learning regardless of their use of directed gaze in video lectures. *Computers & Education*, *128*, 345–352. https://doi.org/10.1016/j.compedu.2018.10.006

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. The MIT Press.

Pinker, S. (1989). Resolving a learnability paradox in the acquisition of the verb lexicon. In *The teachability of language* (pp. 13–61). Paul H. Brookes Publishing.

Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191–215. https://doi.org/10.1006/jpho.2001.0134

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *The Journal of the Acoustical Society of America*, *55*(2), 328–333. https://doi.org/10.1121/1.1914506

Pisoni, D. B., Aslin, R. N., Percy, A. J., & Hennessy, B. L. (1982). Some Effects of Laboratory Training on Identification and Discrimination of Voicing Contrasts in Stop Consonants. *Journal of Experimental Psychology. Human Perception and Performance*, *8*(2), 297–314.

Poepsel, T. J., & Weiss, D. J. (2014). Context influences conscious appraisal of cross situational statistical learning. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00691

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *The Journal of the Acoustical Society of America*, *97*(2), 1286–1296. https://doi.org/10.1121/1.412170

Quine W. V. (1960). *Word and object*. Massachusetts Institute of Technology.

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

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). Appleton-Century-Crofts.

Rohrer, D., Dedrick, R. F., & Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, *107*(3), 900–908. https://doi.org/10.1037/edu0000001

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926–1928.https://doi.org/10.1126/science.274.5294.1926

Schwab, J. F., & Lew-Williams, C. (2016). Repetition across successive sentences facilitates young children’s word learning. *Developmental Psychology*, *52*, 879–886. https://doi.org/10.1037/dev0000125

Scofield, J., & Behrend, D. A. (2011). Clarifying the role of joint attention in early word learning. *First Language*, *31*(3), 326–341. https://doi.org/10.1177/0142723710395423

Scott, R. M., & Fisher, C. (2012). 2.5-year-olds use cross-situational consistency to learn verbs under referential uncertainty. *Cognition*, *122*(2), 163–180. https://doi.org/10.1016/j.cognition.2011.10.010

Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, *31*(1), 010502. https://doi.org/10.11613/BM.2021.010502

Shi, R., Marquis, A. R., & Gauthier, B. (2006). *Segmentation and Representation of Function Words in Preverbal French-Learning Infants*. https://www.semanticscholar.org/paper/Segmentation-and-Representation-of-Function-Words-Shi-Marquis/3e13447420d19621addad3db48210710b5bdc3f8

Singh, L. (2022). From information to action: A commentary on Kidd and Garcia (2022). *First Language*, *42*(6), 814–817. https://doi.org/10.1177/01427237221090024

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1), 39–91. https://doi.org/10.1016/S0010-0277(96)00728-7

Smith, L. B. (2000). Learning How to Learn Words: An Associative Crane. In R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Becoming a Word Learner: A Debate on Lexical Acquisition* (pp. 51–80). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195130324.003.003

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms. *Cognitive Science*, *35*(3), 480–498. https://doi.org/10.1111/j.1551-6709.2010.01158.x

Snow, C. E., & Hoefnagel-Höhle, M. (1978). The Critical Period for Language Acquisition: Evidence from Second Language Learning. *Child Development*, *49*(4), 1114–1128. JSTOR. https://doi.org/10.2307/1128751

Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition*, *38*, 179–211. https://doi.org/10.1016/0010-0277(91)90051-5

Sönning, L., & Werner, V. (2021). The replication crisis, scientific revolutions, and linguistics. *Linguistics*, *59*(5), 1179–1206. https://doi.org/10.1515/ling-2019-0045

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), Article 6640. https://doi.org/10.1038/41102

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346. https://doi.org/10.1037/bul0000169

Storkel, H. L. (2002). Restructuring of similarity neighbourhoods in the developing mental lexicon. *Journal of Child Language*, *29*(2), 251–274. https://doi.org/10.1017/s0305000902005032

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–145. https://doi.org/10.3758/BF03202673

Suanda, S. H., & Namy, L. L. (2012). Detailed Behavioral Analysis as a Window into Cross-Situational Word Learning. *Cognitive Science*, 36(3), 545–559. https://doi.org/10.1111/j.1551-6709.2011.01218.x

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. Journal of Experimental *Child Psychology*, 126, 395–411. https://doi.org/10.1016/j.jecp.2014.06.003

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. https://doi.org/10.1016/S0010-0277(00)00081-0

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2013). From Action to Interaction: Infant Object Exploration and Mothers' Contingent Responsiveness. *IEEE Transactions on Autonomous Mental Development*, *5*(3), 202-209. doi: 10.1109/TAMD.2013.2269905

Tees, R. C., & Werker, J. F. (1984). Perceptual flexibility: Maintenance or recovery of the ability to discriminate non-native speech sounds. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *38*(4), 579–590. https://doi.org/10.1037/h0080868

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*(1), 16–34. https://doi.org/10.1016/j.jml.2006.07.002

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716. https://doi.org/10.1037/0012-1649.39.4.706

Tomasello, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*, 61–82. https://doi.org/10.1515/cogl.2001.012

Tomasello, M., & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, *10*(2), 201–224. https://doi.org/10.1016/0885-2014(95)90009-8

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child Development*, *57*(6), 1454–1463.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First Language*, *4*(12), 197–211. https://doi.org/10.1177/014272378300401202

Tripathi, T., Dusing, S., Pidcoe, P. E., Xu, Y., Shall, M. S., & Riddle, D. L. (2019). A Motor Learning Paradigm Combining Technology and Associative Learning to Assess Prone Motor Learning in Infants. *Physical Therapy*, *99*(6), 807–816. https://doi.org/10.1093/ptj/pzz066

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. https://doi.org/10.1016/j.cogpsych.2012.10.001

Tuninetti, A., Mulak, K. E., & Escudero, P. (2020). Cross-Situational Word Learning in Two Foreign Languages: Effects of Native Language and Perceptual Difficulty. *Frontiers in Communication*, *5*. https://www.frontiersin.org/articles/10.3389/fcomm.2020.602471

Van Leussen, J.W., & Escudero, P. (2015). Learning to perceive and recognize a second language: The L2LP model revised. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01000

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382. https://doi.org/10.1016/j.cognition.2013.02.015

Vlach, H. A., & DeBrock, C. A. (2017). Remember dax? Relations between children's cross-situational word learning, memory, and language abilities. *Journal of Memory and Language*, *93*, 217–230. https://doi.org/10.1016/j.jml.2016.10.001

Von der Malsburg, T. (2015). Saccades: Detection of fixations in eye-tracking data. Retrieved from https://github.com/tmalsburg/saccades

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science*, *10*(2), 159–164. https://doi.org/10.1111/j.1467-7687.2007.00549.x

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107(2), 729–742. https://doi.org/10.1016/j.cognition.2007.08.007

Walley, A.C. (1993). The Role of Vocabulary Development in Children's Spoken Word Recognition and Segmentation Ability. (1993). *Developmental Review*, *13*(3), 286–350. https://doi.org/10.1006/drev.1993.1015

Wang, Y., & Kuhl, P. K. (2003). Evaluating the ``Critical Period'' Hypothesis: Perceptual Learning of Mandarin Tones in American Adults and American Children at 6, 10 and 14 Years of Age. In M. Solé, J., Recasens, D., & Romero, J. (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences, Barcelona, Spain 2015* (pp. 1537-1540).

Wang, F. H., & Mintz, T. H. (2018). The role of reference in cross-situational word learning. *Cognition*, *170*, 64–75. https://doi.org/10.1016/j.cognition.2017.09.006

Waxman, S., Fu, X., Arunachalam, S., Leddon, E., Geraghty, K., & Song, H. (2013). Are Nouns Learned Before Verbs? Infants Provide Insight into a Longstanding Debate. *Child Development Perspectives*, *7*(3), 10.1111/cdep.12032. https://doi.org/10.1111/cdep.12032

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *ournal of Memory and Language*, *50*(1), 1–25. https://doi.org/10.1016/S0749-596X(03)00105-0

Werker, J. F., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, *1*(2), 197–234. https://doi.org/10.1080/15475441.2005.9684216

Werker, J. F., & Hensch, T. K. (2015). Critical Periods in Speech Perception: New Directions. *Annual Review of Psychology*, *66*(1), 173–196. https://doi.org/10.1146/annurev-psych-010814-015104

Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. C. (1981). Developmental Aspects of Cross-Language Speech Perception. *Child Development*, *52*(1), 349–355. https://doi.org/10.2307/1129249

Werker, J. F., & Tees, R. C. (1983). Developmental changes across childhood in the perception of non-native speech sounds. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *37*(2), 278–286. https://doi.org/10.1037/h0080725

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Woodward, A. L. (2000). Constraining the Problem Space in Early Word Learning. In R. M. Golinkoff & K. Hirsh-Pasek (Eds.), *Becoming a Word Learner: A Debate on Lexical Acquisition* (pp. 81–114). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195130324.003.004

Woodard, K., Gleitman, L. R., & Trueswell, J. C. (2016). Two- and Three-Year-Olds Track a Single Meaning During Word Learning: Evidence for Propose-but-Verify. *Language Learning and Development*, *12*(3), 252–261. https://doi.org/10.1080/15475441.2016.1140581

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-Month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*(2), 234–243. https://doi.org/10.1016/j.cognition.2009.08.010

Yeung, H. H., Chen, L. M., & Werker, J. F. (2014). Referential Labeling Can Facilitate Phonetic Learning in Infancy. *Child Development*, *85*(3), 1036–1049. https://doi.org/10.1111/cdev.12185

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13–15), 2149–2165. https://doi.org/10.1016/j.neucom.2006.01.034

Yu, C., & Smith, L. B. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science*, 18(5), 414–420. https://doi.org/10.1111/j.1467-9280.2007.01915.x

Yurovsky, D., Yu, C., & Smith, L. B. (2012). Statistical Speech Segmentation and Word Learning in Parallel: Scaffolding from Child-Directed Speech. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00374

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive Processes in Cross-Situational Word Learning. *Cognitive Science*, 37(5), 891–921. https://doi.org/10.1111/cogs.12035

Zhang, Y., Chen, C., & Yu, C. (2019). Mechanisms of Cross-situational Learning: Behavioral and Computational Evidence. In J. B. Benson (Ed.), *Advances in Child Development and Behavior* (Vol. 56, pp. 37–63). JAI. https://doi.org/10.1016/bs.acdb.2019.01.001

Zhang, Y., Yurovsky, D., & Yu, C. (2021). Cross-situational Learning From Ambiguous Egocentric Input Is a Continuous Process: Evidence Using the Human Simulation Paradigm. *Cognitive Science*, *45*(7), e13010. https://doi.org/10.1111/cogs.13010

Appendix A. Model on proportion of fixations in the learning phase with main effects and interactions among areas of interest, spacing conditions, social cues, and linear and quadratic time term

Effect	Estimate	S.E.	t-value	p-value
(Intercept)	0.10	0.01	9.70	p < .001**
Interaction between area of interest and trial number				
target:trial 2	0.04	0.02	2.90	p < .05*
high frequency distractor:trial 2	0.03	0.03	1.01	p > .05
target:trial 3	0.06	0.02	4.20	p < .001**
high frequency distractor:trial 3	0.02	0.03	0.93	p > .05
target:trial 4	0.06	0.02	4.07	p < .001**
high frequency distractor:trial 4	0.01	0.03	0.46	p > .05
Interaction among time, area of interest, and spacing condition				
linear:target:interleaved	-0.07	0.04	-1.57	p > .05
linear:high frequency distractor:interleaved	-0.05	0.07	-0.64	p > .05
quadratic:target:interleaved	-0.01	0.04	-0.28	p > .05
quadratic:high frequency distractor:interleaved	-0.02	0.07	-0.26	p > .05
Interaction among time, area of interest, and social cue				
linear:target:with social cue	0.17	0.04	3.96	p < .001**
linear:high frequency distractor:with social cue	-0.07	0.08	-0.88	p > .05
quadratic:target:with social cue	0.11	0.04	2.50	p < .05*
quadratic:high frequency distractor:with social cue	-0.04	0.08	-0.55	p > .05

Effect	Estimate	S.E.	t-value	p-value
Interaction among time, area of interest, trial number, and social cue				
linear:target:trial 2:with social cue	-0.04	0.06	-0.62	p > .05
linear:high frequency distractor:trial 2:with social cue	0.11	0.11	1.02	p > .05
linear:target:trial 3:with social cue	-0.02	0.06	-0.31	p > .05
linear:high frequency distractor:trial 3:with social cue	0.06	0.11	0.60	p > .05
linear:target:trial 4:with social cue	0.00	0.06	0.08	p > .05
linear:high frequency distractor:trial 4:with social cue	0.03	0.11	0.27	p > .05
quadratic:target:trial 2:with social cue	-0.02	0.06	-0.37	p > .05
quadratic:high frequency distractor:trial 2:with social cue	0.00	0.11	-0.03	p > .05
quadratic:target:trial 3:with social cue	-0.17	0.06	-2.76	p < .05*
quadratic:high frequency distractor:trial 3:with social cue	-0.03	0.11	-0.32	p > .05
quadratic:target:trial 4:with social cue	-0.16	0.06	-2.51	p < .05*
quadratic:high frequency distractor:trial 4:with social cue	0.04	0.11	0.33	p > .05
linear:target:interleaved:with social cue	0.11	0.06	1.77	p > .05
linear:high frequency distractor:interleaved:with social cue	0.01	0.11	0.11	p > .05
quadratic:target:interleaved:with social cue	-0.05	0.06	-0.79	p > .05
quadratic:high frequency distractor:interleaved:with social cue	0.00	0.11	0.01	p > .05
Interaction among time, area of interest, trial number, spacing condition, and social cue				
linear:target:trial 2:interleaved:with social cue	-0.18	0.09	-2.03	p < .05*

Effect	Estimate	S.E.	t-value	p-value
linear:high frequency distractor:trial 2:interleaved:with social cue	-0.15	0.15	-0.97	p > .05
linear:target:trial 3:interleaved:with social cue	-0.13	0.09	-1.49	p > .05
linear:high frequency distractor:trial 3:interleaved:with social cue	-0.09	0.15	-0.62	p > .05
linear:target:trial 4:interleaved:with social cue	-0.08	0.09	-0.90	p > .05
linear:high frequency distractor:trial 4:interleaved:with social cue	0.02	0.15	0.12	p > .05
quadratic:target:trial 2:interleaved:with social cue	0.02	0.09	0.20	p > .05
quadratic:high frequency distractor:trial 2:interleaved:with social cue	0.06	0.15	0.38	p > .05
quadratic:target:trial 3:interleaved:with social cue	0.12	0.09	1.40	p > .05
quadratic:high frequency distractor:trial 3:interleaved:with social cue	0.08	0.15	0.52	p > .05
quadratic:target:trial 4:interleaved:with social cue	0.16	0.09	1.87	p > .05
quadratic:high frequency distractor:trial 4:interleaved:with social cue	-0.01	0.15	-0.03	p > .05