# The role of interactive and cognitive biases in language use and language change

by

Sara Morales Izquierdo

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology
August 2023

## Table of Contents

## List of Tables

## List of Figures

## Acknowledgements

First of all, I would like to thank Dr. Olga Fehér, for her excellent supervision, for the long discussions and her feedback on each of the aspects of the research and, for her support and her encouragement to take the time to explore different ideas while reminding me to land them when I went on long derivatives. Also, for her patience, flexibility, and her honesty and transparency across this very bumpy road, adapting her supervision style to the circumstances at each time, and for helping me find ways of working that work for me. Equally, I would like to thank my second supervisor Professor Sotaro Kita for his insightful feedback on the direction of my thesis, and my mentor Dr. Hester Duffy, for her non-judgmental support in crucial moments of the program.

Thank you to Dr. Paula Rubio-Fernández and to Dr. Chiara Gambi for agreeing to act as external examiners for the dissertation and taking the time to read it.

I would also like to thank Carmen Saldaña, Jia Loy, and Kenny Smith for their collaboration in the illusion of causality, and category accentuation project for providing their expertise, materials, and guidance at every stage of the project.

Equally, I would like to thank the research assistants Máté Kövesdi, Clarisa García-Moreno Dora, Narmeen Sheikh, and Jade Partridge for their invaluable help in transcribing and coding data, and to Alexander Martin and Ben Shinkwin for volunteering their voice and recording all audio stimuli.

I am also very grateful for all my peers and friends at the department of psychology for their continuous support during the program. Sharing this journey with you made it infinitely easier.

Finally, I would like to thank my friends, Iraia, Jessi, Mina, and Aisha for their support, body-doubling, and encouragement across these almost five years. To my partner Rachel for her infinite support, and for keeping me focused and sane, particularly in the last few months of the writing process, and for making life easier and better in general. To my *ama*, for believing that I could and should finish this thesis and encouraging me to do it.

This work is far from perfect, but it wouldn´t even *be* if it wasn’t for all of you. Thank you.

## Declaration

This thesis is submitted to the University of Warwick in support of the application for the degree of Doctor of Philosophy in Psychology. It has been composed by the author and has not been submitted in any previous application for any degree. The work presented (including data collection and data analyses) was carried out by the author.

## Abstract

The richness and diversity of human languages is remarkable. Researchers have tried to understand how languages evolved to be how they are now, identifying how processes in language acquisition, transmission, interaction, and use shape their structure. This research comes from different disciplines and methodological approaches, such as typological research, cognitive science, pragmatics, or developmental psychology. This thesis attempts to integrate the learnings from these disparate fields using novel methodological approaches to understand the process and forces in language evolution. This first part of this dissertation, Chapters 2 and 3, explore the effect of interaction in language learning and evolution. We expand on the existing artificial language learning paradigms to allow a real-time observation and monitoring of language learning process through interaction. This novel paradigm allows as to observe how asking participants to guess the meaning of a word before producing it boosts the speed of language acquisition. We also discuss and propose different ways in which these paradigms can be used to directly observe the way in which sociolinguistic, pragmatic, and communicative processes affect language structure while it is being acquired through interaction. Aside from interactive biases, individual cognitive biases also have been shown to affect language evolution. The second part of the dissertation, Chapters 4, 5, and 6, address how domain-general biases in processing can affect language change. We find evidence that illusion of causality and category accentuation biases shape language acquisition, leading to language change and interacting with other communicative and cognitive pressures for language evolution. In summary, this dissertation bridges the gaps between the different disciplines working on understanding language evolution. It offers methodological innovations for the study of language learning through interaction and it shows how domain-general biases can explain some of the variability and observations in language evolution and interact with other better-studied pressures.

# Chapter 1: Introduction

Language acquisition is not a straightforward process. It involves many different processes such as identifying the phonemes of the language (phonemic categorisation), separating the stream of speech into words (segmentation), identifying what words refer to (mapping), and how they can be combined to express complex meanings (learning the grammar). All of this is done in a noisy environment by learning from multiple sources and types of interaction. Disentangling how these processes happen and through what learning mechanisms has been the focus of study of biologists, linguists, psychologists, and etiologists for decades (Clark, 2009).

One of the first attempts to understand the process of language acquisition was by Skinner (1957), who claimed that it could be fully explained by operant conditioning mechanisms. The proponents of this view argued that linguistic behaviour was learnt through the same mechanisms as any other behaviour (Skinner, 1957), through operant and associative learning. They emphasised the effect that operant mechanisms such as positive reinforcement (Skinner, 1938) had on language acquisition. This sparked one of the biggest debates in psycholinguistics.

Skinner's proposal elicited a strong response from Chomsky (1959), who argued that children's linguistic abilities were beyond what could be explained by domain-general learning mechanisms. He developed instead the nativist hypothesis (Chomsky, 1959), which posited that children are born with an innate notion of how languages should work and adapt the input they receive to those notions. The main arguments laid out in support for this theory were: a) the structural similarities between existing languages, referred to as Universal Grammar (Chomsky, 1986; Greenberg, 1963), b) the observation that children's language skills were beyond what they could acquire from the input they were receiving, what came to be known as *poverty of stimulus* (Chomsky, 1959), and c) the observation that children, despite big differences in their linguistic input, seemed to converge on the same grammar (Chomsky, 1986).

This discussion was followed by research gathering evidence for both approaches and the generation of new accounts to understand the origin of languages and the process of their acquisition (Ambridge & Lieven, 2011; Dabrowska, 2015; Dabrowska & Lieven, 2005; Evans & Levinson, 2009; Tomasello, 2005). Christiansen and Chatter (2008) argued that universal patterns were not due to an innate language device, but were the outcome of domain-general cognitive and learning biases that led some structures to be more likely to be selected and transmitted, shaping languages to converge on these common structures over time. Evans and Levison (2009) took this argument further and concluded that language universals did not exist. Through a thorough review of existing typological research, they found that, across languages in the world, there were several exceptions for each Universal Grammar principle. They argued that the so-called universal features stemmed from an ethnocentric analysis based only on a few languages with links to English, and they showcased the vast diversity of language structures that existed in the world. They argued that Universal Grammar features were better characterised as trends which were an outcome of domain-general biases for better communication, and sociocultural evolutionary processes. Finally, usage-based approaches (Tomasello, 2005) argued that linguistic structures were an outcome of instances of language use in interaction and communication.

Notably, a common thread between the arguments against the nativist approach is that they link the process of language acquisition to the one of language evolution (Smith, 2022 for a review). Patterns of linguistic change are interpreted as a window to language transmission processes, and ultimately, to language acquisition processes. That is, if people do not reproduce the language exactly as it is presented to them, the modifications that we observe at the local level (in the context of the language use of an individual), and at language level (observing how a particular language changes over time) must reflect the innate biases that shape language processes (Culbertson, 2012).

Over a decade of research has worked to identify the cognitive, transmission and interaction pressures that give way to these cross-linguistic patterns and how they interact with each other (see Smith, 2022 for a review), using a multitude of methodological and theoretical approaches. In the next section, we will cover the current domain-specific and domain-general theories of language learning and transmission and some of the most important findings. Next, we will turn our attention to interaction, covering both its impact on the process of acquisition, and on language transmission and language change. Finally, we identify the methodological and theoretical gaps that make it hard to bring these fields together.

## Observations of language change

### Universal features

As mentioned earlier, one of the arguments in support of the idea of nativism was the identification of common trends between languages in the world, which led to the development of the concept of Universal Grammar (Chomsky, 1965, 1975; 1986; Greenberg, 1963; O'Grady et al., 1996). The commonalities in aspects such as the presence of nouns, verbs and adjectives in most languages led researchers to conclude that people must have been predisposed to structure languages in a particular manner. However, typological research – i.e., research on the structural characteristics of the world's languages, has revealed that the picture is much more complicated, with exceptions for almost every principle described by Universal Grammar (Dabrowska, 2015; Evans & Levinson, 2009; Newmeyer, 2008; Tomasello, 2005). In addition, the lack of phylogenetic independence between languages, and the long history of contact between languages, make it hard to draw inferences about the origin of structural similarities (Dabrowska, 2015; Evans & Levinson, 2009). The identification of these flaws gave way to alternative approaches to explain how languages came to be how they are. For example, Evans and Levinson (2009) emphasised the sociocultural aspects of language in its evolution, while Christiansen and Chater (2008) proposed that the structural commonalities across languages derived from domain-general cognitive biases. In any case, the identification of these general trends constituted a useful source of information about the underlying mechanisms (Culbertson, 2012). For example, the fact most languages either have a fixed word order or case marking system (Greenberg, 1963) led researchers to hypothesise that this was due to a bias for communicative efficiency (Bentz & Christiansen, 2013), which was then tested experimentally (Kurumada & Jaeger, 2015; Fedzechkina & Jaeger, 2020). Equally, the harmonic bias identified in typological research, that is, the tendency to place all syntactic dependents at the same location relative to the head (e.g., both numerals and adjectives before the noun), led Culbertson et al. (2020) to hypothesise that it was due to a bias for more efficient processing, which they demonstrated experimentally.

## Language emergence

Most of the languages we know have existed for centuries and hence contain a set of grammatical rules that have been transmitted across several generations, changing in a gradual and slow-paced manner. Hence, we could not observe how these languages were generated and converged into their relatively stable structure. There are however a few exceptions with newly emerged languages: mainly *creole* languages and new sign languages (Hudson Kam & Newport, 2005).

Creole languages are born when communities that speak different languages are brought into sudden intense contact with each other (Bickerton, 1984; DeGraff, 2007; Kocab et al., 2016; Mufwene, 2007; Sankoff & Laberge, 1978; Senghas & Coppola, 2001). Usually, a group of speakers need to accommodate to the language of the dominant group, from which they adopt lexicon and functional words. This was often the outcome of colonisation and slavery, in which people with different linguistic backgrounds were enslaved, forcefully transported away from their native territories, with the need of communication between each other and with the colonisers leading to the generation of new languages. The resulting languages called *pidgin*, a simplified version of the dominant language that includes features of the native languages of the speakers. *Pidgin* languages contain a high level of inter- and intra-individual variation in their rules: the lexicon, the syntax, etc. often lack structures to convey complex meanings. When this *pidgin* language is transmitted over generations, it often transforms into a *creole* language. Creoles have clearly defined grammatical rules, lexicon and native speakers whose first language is the creole language itself. An example of this is Kreyòl or Haitian Creole, which emerged in Haiti in the 17th and 18th centuries as a result of linguistic contact between West African people brought to Haiti as slaves and their French colonisers. The language contains a lexicon mainly based in French, but its syntax and grammar have their roots in languages from the Niger-Congo family (DeGraff, 2007).

The process of *creolisation* has often been cited as evidence for the existence of a tendency to generate structured grammar from inconsistent input. Adults whose second language is a *pidgin* or the initial variety of a creole tend to use it inconsistently and unpredictably, alternating between a variety of structures and forms to express a meaning (Bickerton & Givón, 1976). Children who are learning the creole as their native language, however, are thought to play a significant role in the creolisation process, by creating new consistent rules that give structure to the language, thus transforming the unstructured variable input from their parents (Bickerton, 1984; DeGraff, 2007; Kocab et al., 2016; Mufwene, 2007; Sankoff & Laberge, 1978; Senghas & Coppola, 2001).

A similar thing happens in the case of deaf children who are born to families that do not use any sign language. Children seem to use structures in a more consistent way than their hearing parents, creating home sign languages that have levels of structure similar to existing sign languages (Goldin-Meadow, 2003, Goldin-Meadow & Mylander, 1983, Goldin-Meadow et al., 1995, Haviland, 2013). Equally, studies with deaf children of hearing parents who learnt sign language later in life, and thus produce highly variable signs with a large number of mistakes, show that children produce more accurate output than their parents, correcting their mistakes and producing much fewer variable forms (Ross, 2001; Singleton & Newport, 2004) with a lower level of mistakes.

The emergence of consistent grammatical structure has also been observed in the development of Nicaraguan sign language (Goldin-Meadow, 2014; Senghas & Coppola, 2001; Senghas et al., 2004). This language started to develop in Nicaragua when the deaf community became more connected through the creation of a school for deaf children in 1977. Over new generations of speakers, cohorts of students who joined the school in successive years, the language started acquiring structural elements such as compositionality of the signs. Interestingly, the change seemed to be driven by the younger speakers of the language, those who had started the acquisition when they were younger. These changes were then transmitted up to the older speakers, which was interpreted as a sign of the importance of children as drivers of language evolution (Senghas et al., 2004). However, the level of structure in the younger cohort of learners of Nicaraguan Sign Language was still slightly lower than that in American Sign Language, which is more established (Goldin-Meadow et al., 2015), highlighting the impact of the introduction of new learners on the structural characteristics of a language.

## Pressures for change

### Language learning

Although typological research, corpus analysis, and the observation of natural languages have high ecological validity, they have a limitation in common: they do not allow us to manipulate the language that is learnt and transmitted, making it hard to establish the reasons behind the observations. To overcome this, evolutionary linguistics have long relied on experimental methods. These studies use reduced, experimenter-designed languages, created ad hoc for every study. This allows researchers to observe language acquisition, use and transmission processes in a controlled context, minimising the influence of participants' linguistic knowledge coming from the languages they speak (Hudson Kam, 2019; Hudson Kam & Newport, 2005, 2009; Kirby et al., 2008, 2015; Newport, 2020; Smith & Wonnacott, 2010).

In these studies, experimenters tried to replicate observations in real life in the laboratory. For example, to explore the reasons behind the changes observed in the *creolisation* process, researchers trained participants in artificial languages containing inconsistencies similar to those observed in adult non-native speakers (Bickerton & Givón, 1976) and observed whether the child and adult learners introduced changes in the language (i.e., Hudson Kam & Newport, 2005). One of the ways of replicating this inconsistency is the use of unpredictable variation, that is, the alternation of two or more forms to express the same meaning (such as three words that all mark plurality) that vary independently from any other linguistic or contextual element (Fedzechkina et al., 2012; Feher et al., 2016, 2019; Ferdinand et al., 2019; Hudson Kam, 2019; Hudson Kam & Newport, 2005, 2009; Hudson Kam & Chang, 2009; Samara et al., 2017; Smith et al., 2017).

One of the well-known mechanisms for language acquisition is that of *statistical learning* (see Romberg & Saffran, 2010, for a review), a process by which humans, from infancy, are able to detect the frequency distribution of different elements in language and use this information to learn different aspects of the language, from finding boundaries between words (Aslin et al., 1998; Saffran et al., 1996) to learning syntax (Gomez & Gerken, 1999). Given that people have been shown to be good at detecting these regularities, we would expect them to reproduce patterns in their linguistic output, and not to introduce any changes to the language. However, experimental research using languages containing inconsistencies show that, while people are able to detect and reproduce existing regularities governed by clear rules, they are not as good detecting and reproducing probabilistic detecting and reproducing the absence of these regularities (e.g., Hudson Kam & Newport, 2005).

When presenting children and adults with an artificial language that contained unpredictable variation, children were more likely than adults to use a single form more frequently than the others, reducing the amount of variation in the artificial language they were presented with, a process referred to as "regularisation" (see Newport, 2020 for a review). For example, Hudson Kam and Newport (2005) observed that, when provided with a language that contained unpredictable variation, children tended to stick to the most frequent variant (regularisation behaviour). On the contrary, adults tended to replicate the variability that they observed in the language they were taught by using each of the variants with the frequency in which they encountered them during training, a behaviour referred to as *probability-matching*. Taken together, these findings gave rise to Newport's Less is More hypothesis (Newport, 1988), which argues that children simplify the language during acquisition to make it more learnable. This would explain why, when presented with unpredictable variation they choose a favourite form and use it more frequently than the alternative(s) and why they were more likely to acquire grammatical rules in a deterministic manner. The Less is More hypothesis (Newport, 1988) was also used to explain the observations that children introduced change in emergent languages, proposing qualitatively different mechanisms by which children and adults acquire language. Austin et al. (2022) found evidence for a gradual change from the regularisation behaviour found in children to the probability-matching behaviour found in adults and argued that it is due to a switch from deterministic rule-learning in children to probabilistic acquisition by adults. In addition, Hudson Kam and Newport (2009) also found that regularisation behaviour appeared in adults when the input language was made more complex, for example, by increasing the number of possible variants. According to the authors (Hudson Kam & Newport, 2009), the difference between children and adults could be based on their memory capacity. A study by Hudson Kam and Chang (2009) tested whether the difference in regularisation between children and adults stemmed from memory limitations in the retrieval of information. They replicated the study by Hudson Kam and Newport (2009), but they manipulated the cognitive load of the retrieval task. They found that, when providing participants with the nouns and verbs they required to form sentences, the regularisation seen in adults disappeared, and they returned to match the probabilities in their input, as they had done when provided with simpler artificial languages (Hudson Kam & Newport, 2005).

Perfors (2012) followed these results up and investigated whether a disruption in the process of encoding, that is, during language learning, led to an increase in regularisation. Across seven studies combining experimental and computational methods, he did not find this hypothesis to be true, and concluded that a previous bias for regularisation was a requirement for this to happen. However, Hudson Kam (2019) followed up by conducting a study using an artificial language that contained unpredictable variation in the determiners used for different lexical items. Two majority determiners were used for a set of nouns each, and some noise determiners were used equally frequently for nouns within both categories. Hudson Kam (2019) manipulated interference during learning, as Perfors (2012) did, and added a condition manipulating interference during retrieval. Using a more complex language, she found that, as in Perfors (2012), interference during encoding did not affect regularisation, and if anything, it reduced it. However, partially in line with Hudson Kam and Chang (2009), she found that, when incorporating interference in the retrieval phase of the task, participants imposed structure on the language, not by regularising to the most common form, but by creating rules that reduced unpredictability, such as using each of the noise markers with a particular lexical item. We will return to this behaviour, referred to as conditioning (Smith & Wonnacott, 2010), in a later section.

Ferdinand et al. (2019) also studied the effect of cognitive load in regularisation, this time in the moment encoding. They manipulated the level of cognitive load by varying the number of words participants were learning information from. In this study, an equivalent condition using non-verbal stimuli was also included. The increase in cognitive load led to higher regularisation behaviour, which was particularly prominent when using linguistic stimuli, in contrast to non-linguistic stimuli. They argued that the bias for regularisation might stem from an interaction between domain-general processes of cognitive load and language specific processes (Ferdinand et al., 2019).

Finally, Perfors (2016) explored the effect of *pragmatic assumptions* about the experimental task in regularisation behaviour. His hypothesis was that participants completed the studies with two main assumptions in mind: that any variation has significance and that the goal of the study was to produce the correct language. When providing participants with a premise that challenged these assumptions, participants tended to regularise rather than matching the probabilities in their input. Brooks and Kempe (2019) suggested that the difference in pragmatic assumptions about the task might contribute to the difference between children and adults: children, having less prior knowledge, would consider a broader range of hypothesis than adults and apply different assumptions to language learning and to the experimental task in hand, leading to a different behaviour.

In summary, natural observations and experimental research show that language learners, particularly children, tend to reduce any unpredictable variation that they encounter, often by eliminating the least frequent variants (Newport, 2020 for a review). Several factors have been identified to contribute to this effect, such as the complexity of the variation (Ferdinand et al., 2019; Hudson Kam & Newport, 2009), the cognitive load during retrieval (Hudson Kam, 2019; Hudson Kam & Chang, 2009) and certain pragmatic assumptions about the task and the properties of the language (Perfors, 2012, 2016). The difference in executive functioning abilities between children and adults (Hudson Kam & Chang, 2009), and the difference in their previous biases and assumptions (Brooks & Kempe, 2019; Gopnik et al., 2015) could be behind the differences between children and adults in their regularisation behaviours.

So far, we have covered two ways in which participants act when presented with unpredictable variation: probability-matching, that is, reproducing the unpredictable variation that they encounter, and regularising, that is, overproducing one of the variants, hence reducing variability. However, there is a third behaviour that participants often show in these contexts: conditioning (Smith & Wonnacott, 2010). In this case, participants retain the variability, but they make it fully or partially predictable. For example, when trained on two randomly appearing words with the same function (e.g., two plural markers), they condition their use on certain lexical items, that is, they use one of the markers with a set of nouns, and the other with the alternative set. This behaviour is quite common in experimental contexts, with authors reporting on it even if it was not the main focus of the study (Hudson Kam &

Newport, 2005, 2009; Hudson Kam, 2019; Smith et al., 2017; Feher et al., 2016; Perfors, 2016; Wonnacott, 2011).

The sort of conditioned variation found in experimental studies is also common in natural languages (Givón, 1985; Labov, 1986, 2006). As stated before, linguistic variation is abundant but seldom unpredictable (Givón, 1985). There are often multiple ways to express an idea or a meaning, but grammar provides us with rules on the situation in which each of the variants is correct. For example, the plural in English is usually marked with an "-s" (e.g., *cat* to *cats*) for most nouns, but with an "-es" if they finish with an "o" (e.g., *tomato* to *tomatoes*). The final sound of a noun (with some exceptions) reliably predicts its plural form; the plural marker is conditioned to the phonetic properties of the final allophone. Sociolinguistic factors also condition the use of variants (Labov, 1986, 2006). For example, although *alcoholic drink* and *booze* have the same meaning, the formality of the context determines which of the expressions is most likely to be used.

Given how prominent it is in natural language and how participants tend to impose it when presented with unpredictable variation, conditioning has been proposed as a potential mechanism for language evolution (Samara et al., 2017). However, few studies have looked at this phenomenon directly, so the conditions under which learners introduce conditioning in the language are not well known. Hudson Kam (2019) found conditioning to be more frequent when the cognitive load in a production task was increased, and she argued that, given the lower availability of cognitive resources in production, participants in this condition tended to repeat association they had previously used, leading to the formation of strong connections, and hence to conditioning. Equally, Samara et al. (2017) found that both children and adults, when they were presented with a language that contained sociolinguistic conditioning, they were able to learn and replicate these patterns. Nevertheless, when provided with a language containing unpredictable variation, children showed a tendency to show regularisation behaviour and eliminate or reduce the use of the least frequent variants, whereas adults tended to introduce conditioning by using each of the variants with a set of words in the lexicon. Understanding the roots and implications of the introduction of conditioning in language, and how it affects linguistic structure is one of the main aims of this dissertation.

## Transmission

Another source of conditioning identified by researchers is language transmission. Aside from observing how participants treat variation individually, there is recent set of studies observing how generational transmission has shaped it. These studies, based on an experimental paradigm introduced by Bartlett (1932), used diffusion chains, in which the artificial language that the previous generation of participants produce is used as the input language for the following generation, also called *iterated learning* (Kirby & Hulford, 2002). These studies allow us to observe the evolution of artificial languages, and how the biases exhibited by individual learners affect the language structure over generations. One of the main observations is that when providing participants with an unstructured artificial language, small individual biases can get amplified generation by generation leading to fully conditioned compositional languages (Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Kirby et al., 2014 for a review).

For instance, Smith and Wonnacott (2010) found that when using iterated learning, , small statistical biases in individual participants gradually amplify and converge into deterministic rules. Further studies have replicated this gradual increase in structure using different mechanism such as an increase in the size of the community of speakers (Fay et al., 2010; Raviv et al., 2021) or the meaning space (Raviv et al., 2019). Over generations, unstructured languages containing unpredictable variation evolved into compositional linguistic systems that followed deterministic grammatical rules rather than probabilistic rules. Smith et al. (2017) extended this work by looking at the effect of transmission when learning from multiple speakers, instead of just one, and found that the speed in which transmission amplified was slowed down. These studies show how the mere process of transmission can affect linguistic structure (Kirby et al., 2014; Smith, 2022).

## Communication

In addition to language acquisition and transmission processes, we cannot forget that language is a means for communication, and that it is used in interaction. Evans and Levinson's (2009) and Christiansen and Chater's (2008) initial approaches focused on the link between general cognitive biases in language acquisition and language evolution rather than communicative processes.

However, in the last few decades, research has shown that the constraints imposed by communication have a vital role on what language looks like. Communication can even explain some of the patterns that are frequent across languages (Christiansen & Chater, 2016; Culbertson et al., 2012, 2013, 2016; Culbertson & Newport, 2015; Fedzechkina et al., 2012, 2018; Fedzechkina & Jaeger, 2020; Kurumada & Jaeger, 2015; Smith & Culbertson, 2018). One of the main theories accounting for the effect of interaction is that of communicative efficiency (Kirby et al., 2015). According to this theory, structure in language come from a combination of two forces: learnability and expressivity. The force for learnability was first explained through the bottleneck theory (Christiansen & Chater, 2016). Given the limited cognitive ability to process language while it is received, participants reconstruct the structure of the language from the elements they have retained from the limited exposure they have had, which leads to an imperfect reproduction of the language (see Christiansen & Chater, 2016 or Fedzechkina et al., 2018 for a review). This effect biases the languages towards learnability, which stems from simplicity and structure, as learnable languages are more likely to be reproduced accurately. However, learnability is not the only pressure that shapes languages. Successful communication requires language to express meanings unambiguously. The combination of these two forces leads languages to be highly efficient, and nearly perfect in the balance between learnability and expressivity.

Artificial language learning studies have gathered support this hypothesis. Kirby et al. (2008) showed that eliminating the pressure for communication, and hence for expressivity, led a randomly generated language to evolve into one that was high in learnability (as it was composed of a small number of short words) but low in expressivity (the words did not allow meanings to be unambiguously distinguished, with a high number of homonyms). However, when including a pressure for communication, like in Kirby et al.'s 2015 study, randomly generated languages evolved into efficient languages, with a higher learnability than the input language whilst retaining expressivity. Numerous studies since have found results supporting the theory of communicative efficiency (Gibson et al., 2019 for a review). In addition, this account has been used to explain some of the cross-linguistic patterns. For example, Fedzechkina et al., (2017) found that the inverse relationship between word order flexibility and case marking found in languages could be explained by a bias towards communicative efficiency.

As with regularisation, children have been shown to be more susceptible than adults to the cognitive biases related to communicative efficiency. In line with the approach that children drive language evolution, Culbertson and Newport (2015) explored how children treated variability in structures. Concretely, the tested the principle of harmonic bias, first described by Greenberg (1963) which states that languages are biased towards structures in which all modifiers are located either before or after a noun. English, for instance, follows that principle, locating numeral modifiers and adjectives all before the noun they modify (e.g. Two orange cats), whereas Spanish would not follow this principle, as numerals are usually located before the noun they modify and adjectives after it: e.g. *Dos* (two) *gatos* (cats) *naranjas* (orange). When trained on a language containing both possibilities, children choose the harmonic structures, showing a stronger harmonic bias than adults (Culbertson & Newport, 2015). Interestingly, even when receiving an input in which the harmonic structure is not present, children transform their productions to make them harmonic (Culbertson & Newport, 2017). This tendency for regularisation, though arguably more pronounced in children, has also been observed in adults. For example, Fedzechkina et al. (2012) found this same harmonic bias in adults.

So far, we have seen how cognitive constraints such as memory (Hudson Kam, 2019), pragmatic assumptions (Perfors, 2016), communicative pressures (Christiansen & Chater, 2016), and transmission processes (Kirby et al., 2015) can explain the evolution patterns that we observe in language.

## Interaction

Social interaction does not only affect linguistic structure by imposing communicative constraints. Usage-based theories of language acquisition (Bybee, 2010; Ellis, 2006; Tomasello, 2000, among others) argue that language itself is generated through interaction and is an outcome of language use and the cognitive processes involved in it. For example, Tomasello stated that one of the bases for first language acquisition was the understanding that others have communicative intentions and establish joint attention with interlocutors (Tomasello & Farrar, 1986, Tomasello, 2000). An essential part of language acquisition was for children to understand that language was a means to communicate intentions, and trying to decode what these intentions were, based on both the analysis of the linguistic utterance and its pragmatic function (see Clark, 2018 for review).

This led researchers to explore how interaction affected the process of language acquisition, with research showing that, even if infants are able to learn some language through overhearing (Gampe et al., 2012), most of their learning comes from interacting with carers and peers, and that interaction has a boosting effect on their learning (Ataman-Devrin et al., 2023; Anderson & Pembek, 2005; Clark, 2018; Kartushina et al., 2022; Strouse & Samson, 2021; Tomasello & Todd, 1983; Weisleder & Fernald, 2013).

But, if interaction is an essential part of language acquisition, how does it affect language evolution? One of the potential interactive mechanisms involved is linguistic *alignment*. In day-to-day conversations, people tend to imitate their interlocutor's lexicon, grammatical structures, tone, or speed; they align with each other (Branigan et al., 2005; Pickering & Garrod, 2004). Pickering and Garrod (2004) hypothesised that largely automatic structural priming mechanisms explained this effect: hearing a certain linguistic structure leads to the activation of its representation, making it more likely that interlocutors will subsequently use it. In other words, people tend to repeat what they just heard, which makes linguistic processing in real-time interaction faster.

A few recent studies have explored whether interactive processes could also affect the elimination of unpredictable variation. Feher et al., (2016) trained participants on a language containing unpredictable variation. Then, they asked participants to take part in a referential game. They found evidence of structural priming in the interaction phase of the game. Also, when comparing the regularisation behaviour before, during, and after the interaction task, they observed that interaction led to a higher degree of regularisation, which persisted in part after interaction. Following up on this, Feher et al., (2019) trained participants on different versions of an artificial language containing unpredictable variation. Versions differed in the frequency with which one variant appeared in participants' training languages. After training, participants were paired with other participants who had learnt a different version of the language (i. e. with a different frequency of variants) and asked to play a referential game. Due to structural priming processes, participants aligned in their productions and converged on using the variant with a frequency that fell between their individual input frequencies. When participants who had been trained on probabilistic use of a variant (variable users) were paired with participants who had been trained on a categorical language (their training only contained one of the variants), the variable users accommodated to the categorical users, whereas categorical users did not change their behaviour to align with the variable users. In both cases, the effects observed during the interaction phase persisted in a later individual recall test. These studies showed that interaction could play a role in the reduction of unpredictable variation, through priming processes, and in combination with other forces arising from transmission and acquisition.

Another field exploring how communication affects linguistic structure is *experimental semiotics* (see Galantucci & Garrod, 2011 or Nölle & Galantucci, 2022 for recent reviews). Studies in this field are similar to the interaction experiments using artificial language learning in that participants are asked to communicate using a novel system. The evolution of the system is the matter of study and researchers manipulate aspects such as the number of participants interacting with each other (Fay et al., 2010) or the parameters of communication (Garrod et al., 2010). However, in this field, participants are not provided with an initial language whose evolution is observed. Instead, they are often asked to communicate through graphic means, generating the communication system from scratch

(Garrod et al., 2007, 2010; Fay et al., 2010). For example, Motamedi et al. (2019) conducted a study in which they asked participants to use gestures to describe a series of concepts, and they tested how interaction and transmission processes affected the structure of the gestures. They found that, when combining pairwise interactions and transmission, the gestures gained systematicity, evolving into a system that shared features with existing sign languages. The findings in this field strongly align with those using artificial language learning: the emergence of a grammar is encouraged by communicative pressure (Little et al., 2017; Motamedi et al., 2019), transmission (Garrod et al., 2010; Motamedi et al., 2019), the size of the community (Galantucci et al., 2012), or the increase in the meaning space (Nölle et al., 2018).

These results complement the findings using artificial language learning replicating the emergence of grammar in communicative systems, and also show that these processes are not exclusive to language-based systems but common to communication in other types of systems too (Nölle & Galantucci, 2022), and hence, cannot be explained by an innate grammar, unless that grammar is multimodal.

In summary, language change does not seem to be exclusively guided by universal principles. Communicative constraints, interaction, sociolinguistic factors, learning biases and transmission all have been shown to play an integral role in language evolution.

Although the processes of language acquisition, interaction, and language change have been related to each other, the research lines investigating the links between them have remained relatively independent. Studies exploring the effect of interaction in language evolution to date have either used communicative system created by the participants themselves during interaction (e.g., Fay et al., 2010; Garrod et al., 2007, 2010) or have taught participants an artificial language before starting to interact (e.g., Feher et al., 2016, 2019).

However, natural language is usually acquired through interaction with a speaker that knows it. In this thesis, we develop a novel experimental paradigm in which participants acquired an artificial language implicitly, through interaction, and use it to investigate the effects of interaction on language acquisition.

Aside from the differences in methodology between these research lines, the definition of what constitutes interaction also varies greatly. The studies cited here all involved some level of communication between two parties, but the conditions under which this happened varied greatly, with some studies focusing on the verbal exchange of messages with communicative intention vs. the observation of those exchanges (Strouse & Samson, 2021) and other looking at written exchanges in a communicative context vs.

decontextualised production of language (e.g. Feher et al., 2016, 2019). This has made it hard to pinpoint the specific mechanisms involved in the effect of interaction in language acquisition and change. This thesis attempts through the careful and systematic manipulation of the different aspects of interaction, to shed light on this issue.

### Statistical learning and associative learning

So far, we have discussed how different forces in language acquisition, language use, linguistic transmission, and interaction can affect language structure. As discussed earlier, an important domain-general process affecting language acquisition is *statistical learning.* An unfamiliar language to an untrained ear is an unintelligible stream of sounds, and once we know it has communicative intention, starting to decode the different elements in the language requires, amongst many other processes, identifying patterns. Humans have been shown to be quite proficient at identifying and acquiring those patterns in language, through a process called statistical learning, the ability to implicitly extract statistical patterns from complex data (see Romberg & Saffran, 2010, for a review). A classic example of this is Saffran et al.'s 1996 study with infants, in which 8-month-old babies were exposed to streams of meaningless syllables and were able to identify the boundaries between pseudowords based on the transitional probabilities between syllables. This process has also been shown to aid the detection of grammatical rules (Gerken, 2005; Gomez & Gerken, 1999; Reeder et al., 2013; Wonnacott et al., 2008) or phonotactic rules (Chambers et al., 2003). Even probability-matching behaviour in linguistic tasks shows how accurately (adult) participants can reproduce the patterns they encounter with people's output languages closely matching the probability structure of their input (e.g., Hudson Kam, 2019; Hudson Kam & Newport, 2005; Perfors, 2012).

Proficiency in statistical pattern identification is not exclusive to language, with studies finding it with many other types of stimuli, from basic perceptual stimuli to complex reasoning (see Schapiro & Turk-Brown, 2015 for a review).

A closely related process is associative learning (Thiessen & Erickson, 2015). Associative learning is the process by which we perceive the association between certain cues and outcomes in the environment by evaluating their contingency (Shanks, 1995). For example, we may observe that whenever there is dark cloud in the sky (a cue), rain is likely to follow (the outcome). In other words, the probability of the outcome (raining) is affected by the presence of the cue (dark clouds): they are contingent on each other. This process of associative learning has aided humans to predict outcomes very relevant to their survival, such as identifying the cues the predict danger or to find food or shelter (Matute et al., 2015; Shanks, 1995).

When we talk about binary cues and outcomes, there are four natural possibilities, represented in Table 1.1.: Both the cue and the outcome to be present (cell a), both the cue and the outcome to be absent (cell d), the cue to be present while the outcome is absent (cell b) and the outcome to be present while the cue is absent (cell c).

Table 1.1.

*Possible cue-outcome events*

| | Outcome present | Outcome absent |
|---|---|---|
| Cue present | a | b |
| Cue absent | c | d |

We say that a cue and an outcome are contingent on each other when the probability of the outcome depends on the presence/absence of the cue. The measurement of contingency is usually formalised as ΔP (see Equation 1, Allan, 1980), where P(O|C) represents the probability of an outcome when a cue is present (in Table 1.1. a/(a+b)), and P(O|¬C) represents the probability of an outcome when the cue is absent (in Table 1.1. c/(c+d)).

$$\Delta P= P(O|C)-P(O|\neg C). \qquad (1)$$

The contingency (ΔP) value ranges from – 1 to 1. A negative contingency represents a situation in which the presence of a cue reliably predicts the absence of an outcome (e.g., a vaccine that prevents the development of a disease), whereas a positive contingency represents a situation in which the presence of the cue predicts the presence of the outcome (e.g., a toxin causing disease). A value 0 represents a null contingency, that is, a situation where the presence or absence of a cue does not affect the probability of an outcome.

Though humans are quite good at perceiving these contingencies, we have a bias towards Type I errors (perceiving a contingency where there is none) (Blanco, 2017). This has been explained by the relatively higher risk for survival that missing an existing pattern have over perceiving non-existent ones, for example when it comes to perceiving dangers in the environment or finding new sources of nourishment (Haselton & Buss, 2000; Haselton & Nettle, 2006; Blanco, 2017). This bias is referred to as *illusion of causality*, and despite its evolutionary advantage, it comes with drawbacks. For instance, is it has been shown to be strongly related to belief in pseudoscience and conspiratorial thinking (Blanco et al., 2011; Matute et al., 2011; Rodriguez-Ferreiro & Barberia, 2021; Torres et al., 2022).

There are two widely studied circumstances under which illusion of causality is prevalent: when the probability of the outcome is high (*outcome-density bias*, Allan & Jenkins, 1980; Alloy & Abrahamson, 1979; Msetfi et al., 2005) and when the probability of the cue is high (*cue-density bias,* Allan & Jenkins, 1983). This is particularly accentuated when both the cue and the outcome are frequent (Blanco et al., 2013). Relatedly, in the field of social psychology, a similar effect has been found, labelled as *illusory correlation* (Hamilton & Gifford, 1963). Hamilton and Gifford showed that people tend to associate majority groups with positive traits and minority groups with negative traits. The phenomena of illusory correlation and illusion of causality have long been linked (McArthur, 1980) and show how the bias to perceive a non-existent relationship between two elements when either (or both) are frequently observed extend beyond the perception of natural events. Aside from the bias to perceive contingency where there is none, there is also a well-documented bias to exaggerate the extent of existing contingencies, a phenomenon known as *category accentuation* (Tajfel & Wilkes, 1963). Given the importance of categorisation in language acquisition processes such as phonemic categorisation (Maye et al., 2002, 2008) or the perception of grammatical categories (Frost et al., 2016), these biases could also be in operation in language.

Associative learning has been proposed as one of the mechanisms for language learning (Ellis, 2006), and some of the findings in domain-general learning appear to explain some of the phenomena in language learning (see Ramscar et al., 2013). However, when describing the effect of pattern detection in language learning, perceptual biases are not usually accounted for. As described earlier, one of the behaviours that is frequently observed in the treatment of unpredictable variation is that of conditioning (Hudson Kam & Newport, 2005, 2009; Hudson Kam, 2019; Smith et al., 2017; Feher et al., 2016; Perfors, 2016; Wonnacott, 2011). However, while there are many studies devoted to understanding regularisation and the conditions under which it appears (Feher et al., 2016, 2019; Ferdinand et al., 2019; Hudson Kam, 2019; Hudson Kam & Newport, 2005, 2009; Hudson Kam & Chang, 2009; Perfors, 2012, 2016), the predictors of conditioning behaviour are not well known (Samara et al., 2017).

The illusion of causality, illusory correlation and conditioning imply imposing structure where there is none. Similarly, category accentuation implies an increase in the structure. Given that associative learning process have shown to be involved in language (Ellis, 2006), we hypothesise that a biased perception of statistical regularities could explain conditioning behaviour in language, leading to a gradual increase of linguistic structure.

In this thesis, we use a series of artificial language studies to test whether the biases we find in domain-general associative learning can explain some of the patterns that we observe in language change in the face of unpredictable variation.

## Thesis overview

Across this dissertation, we use artificial language learning in novel and flexible ways to explore questions regarding the effect of basic cognitive processes and interaction on how languages are acquired and used. The first part of the dissertation focuses on the effect of interaction in language acquisition and language change, developing a paradigm to investigate the effect of interaction on language learning. The second part of the dissertation explores the impact of associative learning and domain-general biases on language change in language evolution.

Chapter 2 presents a series of experiments in which participants acquire a miniature artificial language implicitly, that is, without direct instruction, through an adaptation of the director-matcher paradigm. The aim of this study is two-fold: on the one hand, we test the paradigm, stimuli, and the different measures of language learning, and on the other hand, we test the effect of interactivity on language acquisition, separately for semantic, phonological and grammatical accuracy.

Chapter 3 presents a further development of the paradigm in which participants learn an artificial language through online interaction with a confederate and another participant. We present a design for a study that would detangle the effect of being involved in an interaction and the effect of learning through observation vs. through interaction, which is one of the shortcomings of the existing research on the effect of interaction over language acquisition. In addition, we propose a multitude of manipulations that this paradigm would allow, permitting the simultaneous manipulation of language structure, mode of exposure, and social variables, and how this could advance the existing theoretical accounts for language acquisition and evolution. The empirical testing of this paradigm was not possible due to the restrictions imposed by the COVID-19 pandemic.

Chapter 4 presents an experiment in which we test whether the predictions from associative learning biases can account for conditioning behaviour in the face of unpredictable variation. In a related study, Chapter 5 tests the predictions from category accentuation literature in the treatment of probabilistic conditioning. In both cases, we instruct participants on an artificial language and manipulate the statistical distribution of different elements, observing the change from the input language to the output language and comparing them to the predictions from domain-general and linguistic theories.

The experiments in Chapter 4 and 5 both included questions that explicitly asked participants to report their perception of the statistical distribution of different elements. Chapter 6 examines these data in relation with the results from Chapters 4 and 5, answering questions about how awareness affects the processes of language change.

# Chapter 2: Effect of interactivity and ambiguity in language learning

Children acquire language in a rich social context, through social interactions with their carers, siblings, extended family, and other children of similar ages. As such, social interaction in first language learning appears not only the most common source of learning, but an essential one (Clark, 2018). Observational studies show that child-carer interactions can have a significant impact on language learning, with direct interaction resulting in better learning than overhearing conversations of others (Weisleder & Fernald, 2013) or passive exposure to language on a screen (Kartushina et al., 2022). In an effort to directly explore the conditions under which social interaction impacts language learning, experiments have compared language acquisition live vs on-video, finding that young infants struggle to learn linguistic information, such as novel words or structures, in the absence of real-life interaction, a phenomenon known as "video deficit" (Anderson & Pembek, 2005; Roseberry et al., 2014; see meta-analysis by Strouse & Samson, 2021). However, this effect can be mitigated against by adding a social element to the video presentation: the inclusion of a social support figure (Roseberry et al., 2009), the company of a peer (Lytle et al., 2018), or using a videocall instead of pre-recorded material (Roseberry et al., 2014) all led to better learning. Different mechanisms that have been hypothesised to contribute to the effect of interaction on language learning include the use of eye gaze as a cue (Baldwin, 1993; Tomasello, 1995), joint attention (Tomasello & Todd, 1983; Pruden et al., 2006; Ataman-Devrim et al., 2023), the contingency between the infants' actions and the learning source (Roseberry et al., 2014), and motivation (Walton et al., 2012), with a combination of several of them possibly providing the explanation (Lytle & Kuhl, 2018).

Social interaction facilitating language learning is not limited to infants' first language acquisition. Although adults can learn from non-interactive contexts (Barr & Wyss, 2008), studies in second language acquisition in adult learners also point to the positive impact of interaction. Most studies addressed this issue in the context of real-life interaction in the classroom. Different approaches to teaching, such as collaborative language learning, a teaching style that focuses on promoting interaction between peers and with the teacher (Long & Porter, 1985; Pica, 1992, 1994; Thorne & Lantolf, 2006), or collaborative writing, a technique in which different students interact with each other to write a text in their second language (see Elabdali, 2021 for a meta-analysis), have been shown to promote language learning. However, a recent review on the topic showed that, although some form of interaction generally benefits the acquisition of a second language, the field fails to agree on basic definition of interaction, and consequently, on methods to explore this question, making it difficult to look at the mechanisms involved (Hiver et al., 2021). Most of these studies are conducted in the classroom and focus on learning outcomes, and thus do not include manipulations that would easily allow us to pinpoint the specific mechanisms that make communicative approaches work. However, some interactive mechanisms have been identified as potentially playing an important role.

One suggested mechanism is corrective feedback (Pica, 1992, 1994). With the development of virtual learning environments, some recent studies attempted to observe if mere corrective feedback in word and grammar language tasks can be beneficial for adult second language learning. Findings show that consistent feedback leads to a better recall of new linguistic structures (Dale & Christiansen, 2004; Krishnan et al., 2018), particularly when it comes to the word-referent pairing aspect of language learning (Krishnan et al., 2018). Relatedly, recent studies that investigated learning in an online environment have concluded that increased interactivity helps the acquisition of new factual information in young adults (De Felice et al., 2021) as well as novel words and action patterns in young children (Myers et al., 2017).

These studies focus on and manipulate very different aspects of interaction, with some understanding interaction as a verbal communicative exchange between two or more parties (Anderson & Pembek, 2005; Ataman-Devrim et al., 2023; Kartushina et al., 2022; Tomasello & Todd, 1983; Pruden et al., 2006; Roseberry et al., 2014; Weisleder & Fernald, 2013), whilst others define it as collaboration in a task (Elabdali, 2021; Long & Porter, 1985; Pica, 1992, 1994; Thorne & Lantolf, 2006), and others focus on some level of responsiveness and feedback in online learning (De Felice et al., 2021; Myers et al., 2017). Equally,, the contrast between the interactive condition and the non-interactive one involves complex subtle simultaneous differences that make it hard to isolate the individual contributions of different variables. These limitations make it hard to identify and test what the contribution of different elements in the effect of interaction in language learning (Hiver et al., 2021).

Another variable that is very hard to control for in language learning experiments is participants' prior exposure to the language. Artificial language learning paradigms can overcome this hurdles and have indeed been used to investigate mechanisms of language learning (e.g., Bernard & Onishi, 2023; Fedzechkina et al., 2012, 2017, 2018, 2020; Marcus et al., 1999; Saffran et al., 1996). Studies using online language learning paradigms however have rarely used interactive contexts. The few exceptions have concentrated on long term language change or language evolution (Feher et al., 2016, 2019; Kirby et al., 2014; Raviv et al., 2019) rather than acquisition. Across the present three experiments, we explore the effect of task interactivity on language learning by developing and testing a novel artificial language learning paradigm in which learning occurs in an interactive context.

Our paradigm is based on a referential communication task developed by Krauss and Weinheimer (1966), also known as "Pictionary" (Garrod et al. 2007) or "director-matcher" task (Clark & Wilkers-Gibbs, 1986). As in previous studies, participants play a game in which they are asked to describe an item to, or receive a description from, their partner, in this case a computer programme. In previous studies using artificial language learning, participants receive direct training on the nouns and/or grammar before using them in the director-matcher task (i.e. Feher et al. 2016, 2019; Saldaña et al., 2019a; Smith et al., 2014, 2017), or they are given a set of linguistic components from which they have to generate their own language through interaction (Kirby et al., 2014; Raviv et al., 2019). Similarly, in studies in experimental semiotics using this paradigm participants develop their own non-linguistic communication systems through interaction (Fay et al., 2010; Garrod et al., 2007; Healey et al., 2002).

In contrast, in this series of experiments, participants learn an artificial language directly through the director-matcher task, with no previous instruction. This learning environment mirrors natural language learning, in which people learn through interactions with native speakers with the aim of achieving successful communication. This paradigm, therefore, bridges the gap between second language acquisition studies, which usually take place in contexts of explicit instruction (Anderson & Pembek, 2005; Philip et al., 2013; Roseberry et al., 2014), and experimental pragmatics (Clark, 1996; Noveck & Sperber, 2006), opening a new set of possibilities for investigating cognitive processes behind implicit language acquisition through interaction, by allowing the experimental manipulation of different aspects of the learning environment and the direct and detailed observation of trial-by-trial learning outcomes. In addition to the advancement of basic research on cognitive mechanisms, this study has implications for the improvement of language learning in virtual environments, an area of large popularity in recent years with the advent of digital language learning apps like Duolingo.

To investigate the mechanisms behind the effect of interaction on language learning, in this study we focused on interactivity, which we defined as the reception of feedback after the production of an utterance. In all three experiments, we manipulated the level of interactivity of the task in two conditions: in both the non-interactive and interactive conditions, participants learnt an artificial language through playing an interactive game with a computer, however, the task for participants in the interactive condition contained an additional element in which they were asked to generate a response and received evaluative feedback, whereas this element was absent in the non-interactive condition. Based on the literature on interaction in language learning, as well as from more domain-general literature, such as the generation effect on memory research (Bertsch et al., 2007),  we predicted that participants in the interactive condition, by the end of the training, would 1) produce more accurate words for the target objects, 2) produce more accurate word forms overall, and 3) produce words with a more accurate combinatorial structure.

Given the novelty of our paradigm, Experiments 1 and 2 aimed to pilot the specific methodological features and identify shortcomings that would lead to noisy data and undesired differences between conditions. Experiment 1 aimed to test the materials, artificial language, the paradigm, and the measures, to see if participants were able to learn a simple language without explicit instruction. In this experiment, aside from manipulating interactivity, we also manipulated the level of ambiguity (LoA) of the distractors that were presented with the target object in screen. This variable had three levels: maximum LoA, medium LoA, and minimum LoA. Experiment 2 further develops the paradigm, increasing the salience of its interactive aspects and implementing an optimal distractor structure based on the results in Experiment 1. Experiment 3 adjusts the timings, and included a delayed retest to examine the longer term effects of interactivity. The use of artificial languages in an experimental setting allowed us to test the effect of interaction at its most basic level and have greater control of the variables affecting it and perform a closer analysis of the output.

## Experiment 1

Given that this was the first study in which participants learnt an artificial language through interaction and without explicit instruction, and that these specific materials and measures were used, we aimed to test whether it was possible to learn the artificial language within a set number of trials, and whether our measures were sensitive. We manipulated interactivity and distractor structure.

Interactivity had two levels: interactive and non-interactive. In both conditions, participants took part in a director-matcher style game. In the director trials, they were shown an object and asked to retype its name. In the matcher trials, they were shown a name and presented with a group of four objects, one of which was the one corresponding the name. Then, they were told which of the four objects matched the name. The crucial difference between conditions was that, in the matcher trials, participants in the interactive condition were asked to predict which out of four objects the one corresponding with the name would be before being told, whereas participants in the non-interactive condition simply were asked to press “Continue” to see the right answer. The aim was to test whether adding an element of interactivity would affect participants’ learning. We expected participants in the interactive condition to learn faster and have a higher accuracy on their productions than those in the non-interactive condition.

Distractor structure referred to the characteristics of the objects in the array of possible objects in the matcher trial, and had three levels: Minimum LoA, Medium LoA and Maximum LoA. These conditions differed in a few parameters, which as further developed in the “Design” section, but the main difference was on the probability of participants guessing the right object by recalling the meaning of only part of the word. This ranged from 25% in the maximum LoA condition to 100% in the minimum LoA condition. We expected that, compared with the minimum LoA condition, participants in the maximum LoA would show a lower score in matching, due to a lower likelihood of selecting the right answer, but would, overall, learn at higher rate, as keeping one factor constant increases the salience of the modified factor. We also included a medium ambiguity condition containing different types of distractors that would allow us to test whether participants’ awareness of the linguistic structure predicted later performance.

Overall, we hoped this first study would help us: 1) test the materials and the paradigm, 2) test whether interactivity had an effect on language learning, and 3) test whether this effect was dependent on distractor structure.

## Methods

### *Participants*

148 first-year undergraduate students at the University of Warwick took part in the study in exchange of course credit. We used the following exclusion criteria: Participants who had an average accuracy score of less than .875 across all trials in the interaction trials, that is, those had not retyped the words, were excluded from the sample (a total of 13). Also, participants who had not learnt the phonological forms of the words by the last block, that is, those who had an average accuracy score of less than .5 in Block 4, were excluded from the sample (an additional three participants). This yielded a final sample of 132 participants. This research was approved by Ethics Committee of the Department of Psychology at the University of Warwick on the 25th of April 2019.

### *Stimuli And Language Structure*

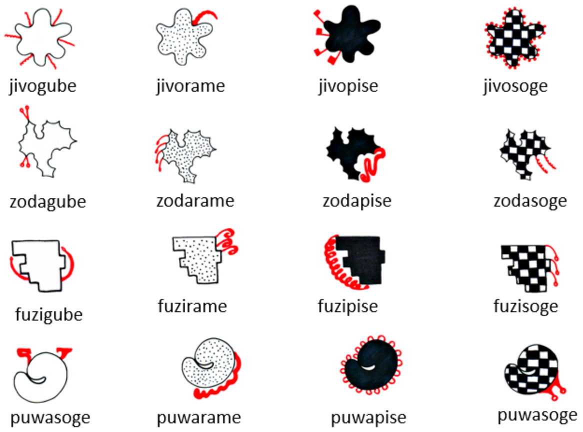
The language consisted of sixteen nouns associated with sixteen different artificial objects. These objects varied in two parameters: shape and filling. There were four distinct shapes and four possible fillings, yielding the sixteen objects. The objects can be observed in Figure 2.1. The nouns were compound, containing two lexemes: the first associated with the shape of the object (“jivo-“, “zoda-“, “fuzi-“, or “puwa-”), and the second with the filling (“-gube”, “-rame”. “-pise”, or “-soge”), which were combined to form sixteen individual labels for the objects (e.g. “jivogube”).

All the components, regardless of whether they referred to shape or filling, were the same length and had the same consonant-vowel structure and were designed to be distinctive and easily readable. That was achieved by using a CVCV structure in which consonants were monographs that had a single main mapping in English in the position they were included in (Brooks, 2015). The only structural difference between shape and filling components was that the end vowel was always “e” for filling components, and never for shape components.

The association between the eight noun components and the eight shapes and fillings was randomised for every participant. Figure 2.1 shows an example of the mapping between these nouns and objects. The set of artificial objects was obtained from the study by Kirby et al., (2015).

Figure 2.1

*Set of artificial objects with an example of the pseudowords they could be assigned.*



### *Design*

We used a mixed design with two between-subject independent variables: level of ambiguity of the distractors and interactivity, and one within-subject independent variable, block, with four levels. The variable interactivity had two levels: interactive or non-interactive. The difference between these two conditions is developed in the Task section. The variable level of ambiguity (LoA) of the distractors had three levels. We defined ambiguity as the inverse probability of participants to guess the right answer when knowing one of the parameters: the higher the probability, the lower the ambiguity. The distractor objects had either: zero parameters in common with the target object (neither shape nor filling) in the minimum LoA condition, one parameter in common with the target object (either the shape or the filling) in the maximum LoA condition or a varying number of parameters in common in the medium LoA condition (one of the distractors had one of the parameters in common, the second distractor had the other parameter in common, and the third distractor did not have any of the parameters in common, neither with the target nor with the remaining distractors).

These conditions also varied on two other key aspects: proportion of the meanings space they were presented with, and the salience of the structural characteristics of the language. The meaning space, here was composed by four shapes and four fillings. The proportion of the meaning space, therefore, referred to how many of the possible shapes and how many of the possible fillings participants saw in each of the trials. For example, participants in the minimum LoA condition saw the 100% of the meaning space for both parameters (shapes and fillings) in each of the trials. This is because each of the trials contained an example of each of the four possible shapes and each of the four possible fillings. In contrast, participants in the Maximum LoA condition, saw in each trial a 25% of the meaning space of one of the parameters and the 100% of the other, as one of the parameters was kept stable across the target and the distractors. For example, they saw four images with the same shape and a different filling. The salience of the structural characteristics referred to whether each trial provided participants with visual cues that would help them understand the structure of the meaning space. For example, structural salience was high for participants in the Maximum LoA condition: by sometimes showing the four possible fillings and keeping the shape stable and other time keeping the fillings stable and showing the four possible shapes, participants could easily infer that the sixteen images arose from the combination of four possible fillings and shapes. In contrast, the salience of structural characteristics for participants in the Minimum LoA condition was low, as in each of the trials, no shape or filling were presented more than once, making it hard for participants to infer that different shapes and fillings could be combined. Table 2.1 shows the differing parameters between distractor condition, and Figure 2.2 presents an example of the different distractor conditions.

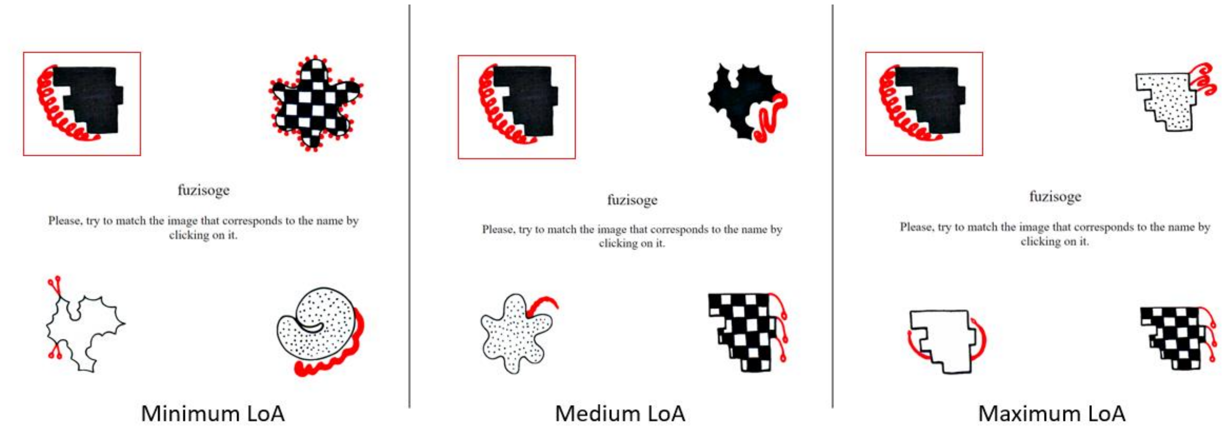
Table 2.1.

*Distractor condition*

| | Minimum LoA | Medium LoA | Maximum LoA |
|---|---|---|---|
| **Probability to guess when knowing one parameter** | 100% | 50% | 25% (chance) if the stable one, and 100% if variable one |
| **Percentage of the meaning space shown** | 100% parameter 1<br>100% parameter 2 | 75% parameter 1<br>75% parameter 2 | 25% parameter1<br>100% parameter 2 |
| **Salience of structural characteristics** | Low | Medium | High |

Figure 2.2.

*Examples of trials by level of ambiguity of the distractors.*



fuzisoge

Please, try to match the image that corresponds to the name by clicking on it.

fuzisoge

Please, try to match the image that corresponds to the name by clicking on it.

fuzisoge

Please, try to match the image that corresponds to the name by clicking on it.

*Note*. The target objects for the trial are framed in red.

Participants were randomly assigned to one of the six conditions. The number of participants for each of the conditions can be observed in Table 2.2. Even if there was variation in the number of participants per condition, the difference was not significant, χ2 = 2.64, p = .267.

Table 2.2

*Number of participants by condition*

| **Interactivity** | **Level of ambiguity** | **N** |
|---|---|---|
| Interactive | Maximum | 12 |
| Interactive | Medium | 21 |
| Interactive | Minimum | 23 |
| Non-Interactive | Maximum | 24 |
| Non-Interactive | Medium | 30 |
| Non-Interactive | Minimum | 22 |

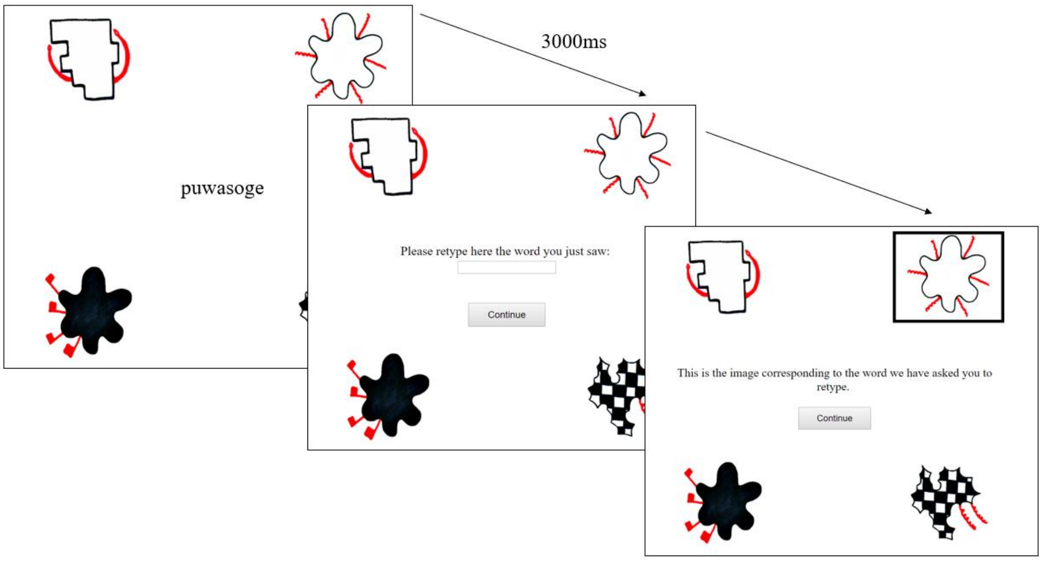
***Task***

Participants learned a novel language (above) by playing a director-matcher style communication game with L3arn, a computer partner, to whom they had to describe objects and who described objects for them in the language. As opposed to previous experiments involving director-matcher tasks (e. g. Clark & Wilkers-Gibbs, 1986; Feher et al., 2019), we did not train the participants prior to the task, so they were all trained during the interaction with the computer agent.

**Interaction trials.** Participants saw two different types of trials: director trials and matcher trials. In director type trials, they were shown four objects forming a square and a noun in the middle. After 3000 ms, the noun disappeared, and participants were asked to retype what they could remember. Participants were required to retype at least four characters to make sure that they were performing the task. Once participants had retyped the noun, the object corresponding to it was framed in a black square. Figure 2.3 shows the structure of the director trials.

Figure 2.3.

*Structure of the director type trials.*



Please retype here the word you just saw:

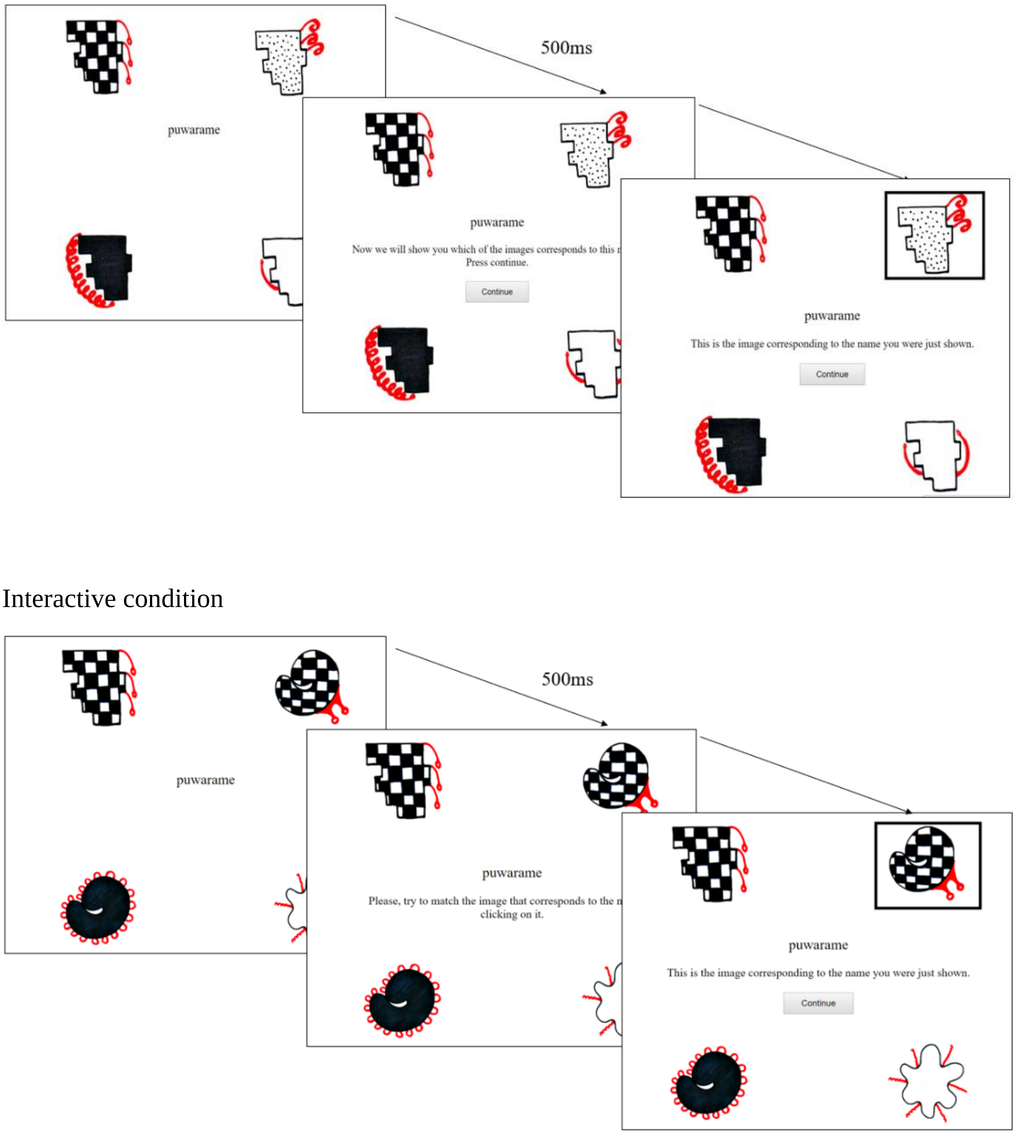
This is the image corresponding to the word we have asked you to retype.

In the matcher type trials, participants were shown four figures forming a square and a noun in the centre of the square. If the participants had been placed in the non-interactive condition, they were shown a prompt asking them to press ‘Continue’ in order to see which image corresponded with the noun in the centre of the screen. When they pressed ‘Continue’, the target object was framed in a black square. If participants had been placed in the interactive condition, after 500ms they were shown a prompt asking them to press on the object they thought corresponded with the noun in the centre of the screen. When they pressed one of the objects, the target object was framed in a black square. The structure of the matcher trials for each of the conditions can be observed in Figure 2.4.

Figure 2.4.

Structure of the matcher type trials by condition of interactivity

Non-interactive condition



puwarame

puwarame

Now we will show you which of the images corresponds to this r
Press continue.

puwarame

This is the image corresponding to the name you were just shown.

Interactive condition

puwarame

puwarame

Please, try to match the image that corresponds to the n
clicking on it.

puwarame

This is the image corresponding to the name you were just shown.

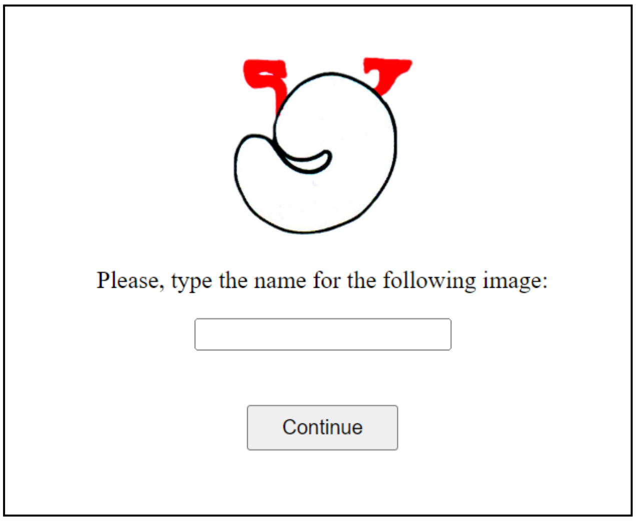
*Note*. The upper diagram shows an example of a non-interactive matcher trial. The lower diagram shows an example of an interactive matcher trial.

Participants alternated between director and matcher trials. Trials were separated by a blank screen that prompted them to press the spacebar to proceed.

**Production test.** Participants were presented one by one, each of the sixteen objects and asked to type the correct label for each of them, as shown in Figure 2.5. The order of presentation was randomised for each participant, and they were not allowed to proceed to the next object until they had produced a response of at least four characters.

Figure 2.5.

*Production test trial*



Please, type the name for the following image:

### *Procedure*

Participants completed the study using their own laptop or desktop, accessing the experiment online, which was hosted on a university server, to which the data was sent automatically upon completion. They were presented with a consent form, which they had to approve to continue to the study. They were informed that they could leave the study by closing the browser, and no cookies were collected.

Participants were asked not to take notes during the study and to complete it in one go, as if they left the window, they would not be able to return at a later time. After reading the instructions, participants proceeded to the Interaction trials of the study. Participants went through four blocks of trials, consisting of 32 trials each (sixteen director type trials and 16 matcher type trials). Across these trials, they were presented with the mapping between each the sixteen possible objects and their corresponding nouns twice per block (once as a director and once as a matcher). The order in which each of the objects was presented as a target was randomised for each participant and the set of distractors for each trial was randomly selected from an array with all the possible sets of distractors given the target object and the distractor condition. The type of trial they started with (director or matcher) was randomised for every block. After every block of trials, participants were presented with a Production Test, in which they were shown each of the sixteen objects, one by-one and asked to produce the correct noun for each of them. The experiment took on average 39.21 minutes (S.D.=18.51).

### *Pre-processing and measures*

As all the participants were required to produce at least a four-character string in every trial, we needed to eliminate those trials in which participants had not attempted to produce the right noun. We eliminated all symbols and spaces and calculated the Levenshtein distance between participants' productions and the target. Levenshtein distance (Levenshtein, 1966) produces an integer value representing the minimum number of operations (addition, deletion, and replacement) required to transform a test string into its target string. A value of 0 represents that the two strings are identical. The higher the value, the less similar the test string and the target string are.

We established five main dependent variables that we calculated for each trial: retyping accuracy (for director-type trials), total accuracy, phonological accuracy, grammatical accuracy, and matching accuracy (only for participants in the interactive condition). All four of them range from 0 (completely inaccurate) to 1 (completely accurate).

Retyping accuracy aimed to explore whether participants in all conditions had followed the instructions to retype the target nonword in the director trials. We run this test to check for any unexpected differences between conditions that could help explain further results. We coded the productions that had a LV distance of 1 or less as correct and all the others as incorrect. We decided to use this dichotomic variable because the distribution of Levenshtein distances was very skewed towards 0, and because any production that differed from the target by more than one error (deletion, replacement, or addition of a character) could be interpreted as a sign of lack of engagement.

Total accuracy measured the level of overlap (i. e. inverse of normalised Levenshtein distance) between a production and the target word. We first normalised the measure by converting all values above 8 into 8, and then divided the distance by 8. Finally, we inverted the scale. Hence, 8 was converted to 0, 4 to .5, 0 to 1, etc., ranging from 0 to 1 in .125 steps. Phonological accuracy measured wordform learning by looking at the level of overlap between a production and its closest existing word in the language. These were the steps we followed: 1) we separated each production in two components, the first contained the first four the second the last four characters, 2) we calculated the LV distance between each of those four characters and the eight word components in the language, 3) we selected the lowest LV value to each of the components and if it was over 4, we transformed it to 4, 4) we averaged this value between both components of the word, 5) we divided the outcome by 4, and subtracted the result from 1. The outcome was a measure that ranged from 0 (no overlap with any of the possible wordforms) to 1 (the word is composed by two of the existing word components, whether they are used in the right order or not), with steps of .125.
Grammar accuracy measured the accuracy of component ordering. The correct order of components was SHAPE + FILLING, therefore, productions that followed that order were

coded as accurate (score of 1), whereas those in any other (e.g. SHAPE + SHAPE) were coded as inaccurate (score of 0). Any other productions were excluded from these analyses. Finally, matching accuracy measured the accuracy of the matching responses for participants in the interactive condition in each trial (0 if incorrect and 1 if correct).

### *Statistical analyses*

Using R 4.2.2 (R Core Team, 2022), we analysed the effect of interactivity condition, distractor condition, and block on our five measures. We used the glmmTMB package (Brooks et al., 2017) to run mixed effect models and applied Satterthwaite's approximation of the degrees of freedom, in order to obtain individual p-values for each of the contrasts of interest (lmerTest package, Kuznetsova et al., 2017). We established a threshold of p < .05 for significance. For all models, we sum coded interactivity condition, and applied repeated contrast coding to block, as we were interested in the block-to-block change in the dependent variables as a measure of learning. We used Helmert contrast coding for distractor condition. The first contrast compared Minimum LoA to Medium LoA, and second, the average of the first two to the Maximum LoA condition. We also conducted nested analyses of the effect of interactivity condition and distractor within each of the blocks, applying Type III Sum of Squares formulas to estimate the effects within the model ("joint tests" function in the emmeans package; Lenth et al., 2022) and Tukey HSD method to perform nested pairwise comparisons to further explore the significant interactions. If any of the models did not converge, we eliminated the random factor with the lowest variance until the model converged, as suggested by Barr et al. (2013).

To explore whether participants from all conditions had followed the instructions, we ran a mixed effects logistic regression on the productions in director-type trials, with retyping accuracy as the outcome, fixed effects for Block, interactivity condition, distractor condition and their interactions, random intercepts for participant, position of the trial in the block, and by-participant random slopes for each of the fixed effects. We did not include the target word as a random intercept, because the model did not converge, and this had the lowest variance. For the analysis of the production trials, we included interactivity condition, distractor condition, and block (1-4) as fixed effects. We also included the triple interaction between all fixed factors, as well as the subsequent two-way interactions as fixed effects, random intercepts for participants and for the target label in that trial, and by-participant random slopes for interactivity condition, distractor condition, and block. We ran separate models for total accuracy and phonological accuracy.  We followed the same structure for grammatical accuracy, but applying a binomial logistic mixed effect model, since grammar accuracy could only adopt two values (correct or incorrect) for each trial.

We also run a logistic mixed effects model for matching accuracy, but exclusively with participants in the interactive condition, and excluding all fixed and random effects relating to interactivity condition.

## Results

### *Retyping behaviour*

The average accuracy of retyping was very high for all blocks and condition (see Table 2.3). We found a main effect of block. The proportion of accurate responses improved from Block 1 to 2, β = .646, z =4.98 , p <.001 , but not from Block 2 to 3, β =.081 , z = .534, p =.594 , nor from Block 3 to 4, β =.129, z =.892 , p =.372. We did not find an effect of interactivity condition, β = -.078, z = -.852, p =.394, nor distractor condition (β =-.025, z = -.257, p = .797, for the first contrast, and β =.015, z = .203, p = .839, for the second contrast), or any interaction between these individual variables and Block, all ps >.134. However, we did find a significant interaction of the second contrasts of distractor condition with interactivity condition, β =.145, z =2.03 , p =.042. A post hoc nested test by distractor condition showed that this was due to an effect of interactivity condition over accuracy in the Medium LoA condition, F(1) = 6.135, p = .013. Participants in the non-interactive – Medium LoA condition produced a higher proportion of accurate responses than those in the interactive-Medium LoA condition.

Table 2.3.

*Mean proportion of correctly retyped trials by condition and block.*

| | | **Block 1** | **Block 2** | **Block 3** | **Block 4** |
|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Interactive** | **Minimum LoA** | .937 (.166) | .961 (.135) | .959 (.137) | .970 (.119) |
| | **Medium LoA** | .914 (.189) | .945 (.157) | .952 (.147) | .967 (.124) |
| | **Maximum LoA** | .937 (.166) | .971 (.111) | .971 (.117) | .971 (.117) |
| **Non-interactive** | **Minimum LoA** | .943 (.159) | .967 (.124) | .977 (.104) | .967 (.124) |
| | **Medium LoA** | .953 (.146) | .971 (.117) | .972 (.115) | .981 (.095) |
| | **Maximum LoA** | .939 (.164) | .964 (.128) | .957 (.140) | .956 (.142) |

***Total accuracy***

We found that participants improved their accuracy from one block to the next, with the biggest increase taking place from Block 1 to 2, β =.191 , z = 10.96 , p<.001, after which it became more gradual, Block 2 to 3, β =.101, z = 6.39, p <.001, Block 3 to 4, β =.064, z =6.75 , p <.001). However, we did not find any effect of interactivity condition across all blocks, β = .004, z =.22, p = .823, nor on the increase between any of the blocks, all ps >.276, nor within any of the blocks, all ps >.314.

Equally, the effect of distractor condition was not significant across all blocks (β = .002, z = ,11, p = .911 for the first contrast, and β =.009 , z =.66. , p = .507 for the second contrast). We did not find an effect over the increase within any of the blocks, all ps >.153, nor on the total accuracy within any of the blocks, all ps> .411.

Finally, we explored the interaction between distractor condition and interactivity condition. We did not find an interaction in total accuracy scores across blocks for the first contrast of distractor condition, β = -.007, z =-.33 , p = .745, but we found a marginally significant interaction between the second contrast of distractor condition and interactivity condition, suggesting the difference between Maximum LoA and the average of the remaining two conditions (Minimum LoA and Medium LoA) could be different for each of the interactivity conditions, β =.027, z =1.92 , p =.051. In order to explore this interaction further, we run a posthoc nested analysis of the effect of interaction condition within each of the blocks and distractor conditions. This showed that the effect of interactivity was only present in Block 4, F = 4.843, p =.027, and marginally in Block 3, F = 3.012, p = .083, for the Maximum LoA condition, but not for any of the other distractor conditions, nor in any of the other blocks, all ps >.189. Similarly, the interaction between interactivity condition and distractor condition over the block-to-bock increase was not significant for any of the blocks, all ps>.259. Table 2.4 and Figure 2.6 show total accuracy change across blocks by condition.

Table 2.4.

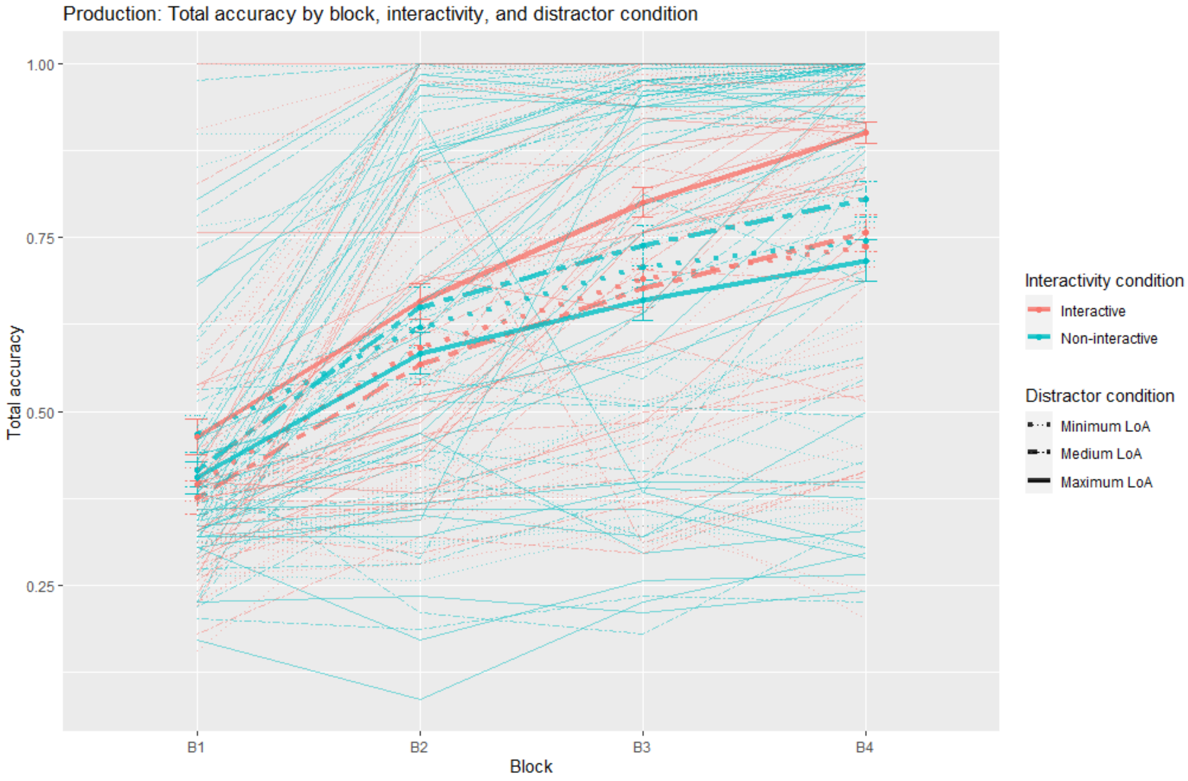
*Average total accuracy by condition and block.*

| | | **Block 1** | **Block 2** | **Block 3** | **Block 4** |
|---|---|---|---|---|---|
| | | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) |
| **Interactive** | **Minimum LoA** | .395 (.272) | .592 (.322) | .692 (.332) | .737 (.327) |
| | **Medium LoA** | .377 (.271) | .567 (.316) | .677 (.306) | .757 (.300) |
| | **Maximum LoA** | .464 (.293) | .658 (.296) | .800 (.247) | .901 (.174) |

| | | | | | |
|---|---|---|---|---|---|
| **Non-interactive** | **Minimum LoA** | .467 (.305) | .620 (.321) | .707 (.321) | .745 (.323) |
| | **Medium LoA** | .417 (.289) | .650 (.335) | .739 (.334) | .805 (.294) |
| | **Maximum LoA** | .405 (.269) | .584 (.348) | .660 (.341) | .717 (.348) |

Figure 2.6.

*Average total accuracy in the production test by block and condition*



*Note*. Error bars represent standard error. Colour represents interactivity condition, and linetype distractor condition. Semitransparent lines represent individual trajectories.

### *Phonological accuracy*

As with total accuracy, the phonological accuracy of participants' production increased from one block to the next in a logarithmic fashion, with the biggest increase taking place from Block 1 to 2, β =.879 , z = 84.43, p<.001, and then gradually from Block 2 to 3, β =.101, z = 9.76, p <.001, and Block 3 to 4, β =.027, z =4.10 , p <.001).

We did not find any effect of interactivity condition across all blocks (β <.001, z =-.92, p = .356). However, the block-by-block analysis showed that, within Block 1, the phonological accuracy of the productions from participants in non-interactive condition was higher than that of those in interactive condition (t = -2.017, p = .044). Subsequently, the increase in phonological accuracy from Block 1 to Block 2 was higher for those participants in the interactive condition, in comparison with participants in non-interactive condition (β = .021, z = 2.06, p = .039), which lead to no difference in the bock-by-block increase from Block 2 onwards, all ps >.251, and no difference between conditions within any of the remaining blocks, all ps>. 575.

The effect of distractor was not significant across blocks, (β = .001, z = .15, p = .879 for the first contrast, and β =-.005 , z =-.71, p = .477 for the second contrast). We did not find an effect over the increase within any of the blocks, all ps >.181, nor on the total accuracy within any of the blocks, all ps> .533.

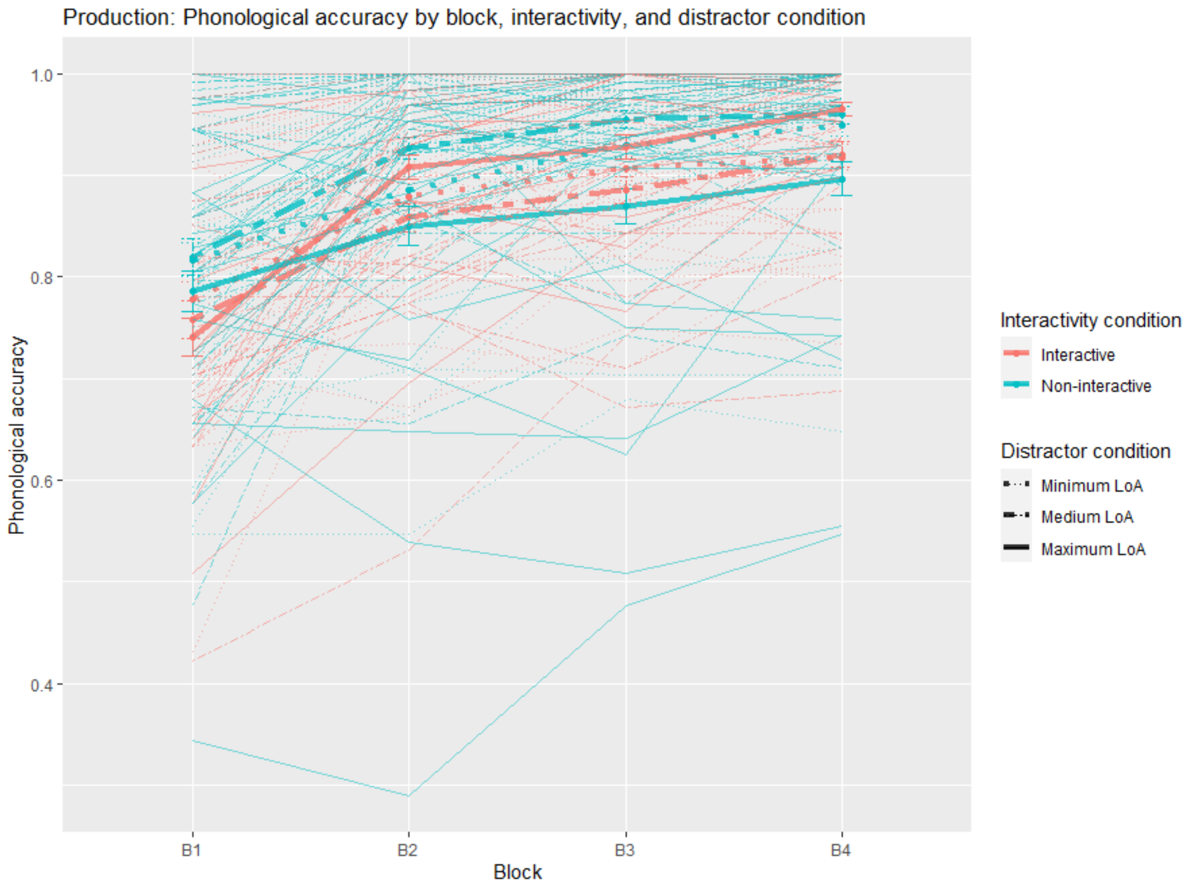
Finally, we explored the interaction between distractor condition and interactivity condition.

As with total accuracy, we did not find an interaction in total accuracy scores across blocks for the first contrast of distractor condition (β = -.087, z =-.103 , p = .303), but we found a marginally significant interaction between the second contrast of distractor condition and interactivity condition, suggesting the difference between the average of Medium LoA and Minimum LoA and Maximum LoA conditions could be different for each of the interactivity conditions (β =.013, z =1.71 , p =.088). Further to that, we found a marginally significant interaction between the second contrast of distractor condition and interactivity condition on the change from Block 1 to Block 2, (β = 0.152, z = 1.92, z = .053), and a marginally significant interaction between the first contrast of distractor condition and interactivity condition in the change from Block 3 to Block 2, (β = .010, z = 1.84, p = .067). All other interactions in the block-to-block change were not significant, p > .423.

In order to explore this interaction further, we run a posthoc nested analysis of the effect of interaction condition within each of the blocks and distractor conditions. The results suggested that the marginal interaction was led by a significant difference between interactivity conditions within the Medium LoA distractor condition within Blocks 2, F = 6.345, and 3, F = 6.546, p = .010, and a significant difference between distractors conditions within the non-interactive condition in those blocks (F = 3.234, p = .039 for Block 2, and F = 3.676, p =.025 for Block 3). All other within-block interactions were not significant, all ps > .073. Participants in the non-interactive – Medium LoA condition obtained significantly higher phonological accuracy scores in these blocks than participants in any other conditions, leading to the marginal interaction by distractor condition from Blocks 1 to 2, that conduced to a significant within block difference in Blocks 2 and 3. We did not find, however, any significant difference within any of the blocks to follow-up from the marginal interaction from Block 3 to 4. Figure 2.7 shows the average values in phonological accuracy by condition across the four blocks.

Figure 2.7.

*Average phonological accuracy in the production test by block and condition.*



*Note*. Error bars represent standard error. Colour represents interactivity condition, and linetype distractor condition. Semi-transparent lines represent individual trajectories.

***Grammatical accuracy***

Grammatical accuracy only showed an increase from Block 2 to Block 3 (β =1.243 , z = 2.702, p = .007). The increase between the rest of the blocks was not significant, all ps >. 986. Similarly, we did not find an effect of interactivity condition overall (β =.309, z =.002, p =.998), nor in the block-by-block increase, all ps > 319, nor within each of the blocks, all ps>.126. The same was the case for distractor condition. We did not see an effect in either of the contrasts (β = -.336, z = -.914, p =.361 for contrast 1, β = .396, z = .002 , p =.998 for contrast 2), nor in the block-by-block increase, all ps>.728, nor within any of the blocks, all ps>.666. Interactivity condition and distractor condition did not interact with each other overall nor in the block-by-block increase, nor within any of the blocks, all ps>.361.

***Matching accuracy***

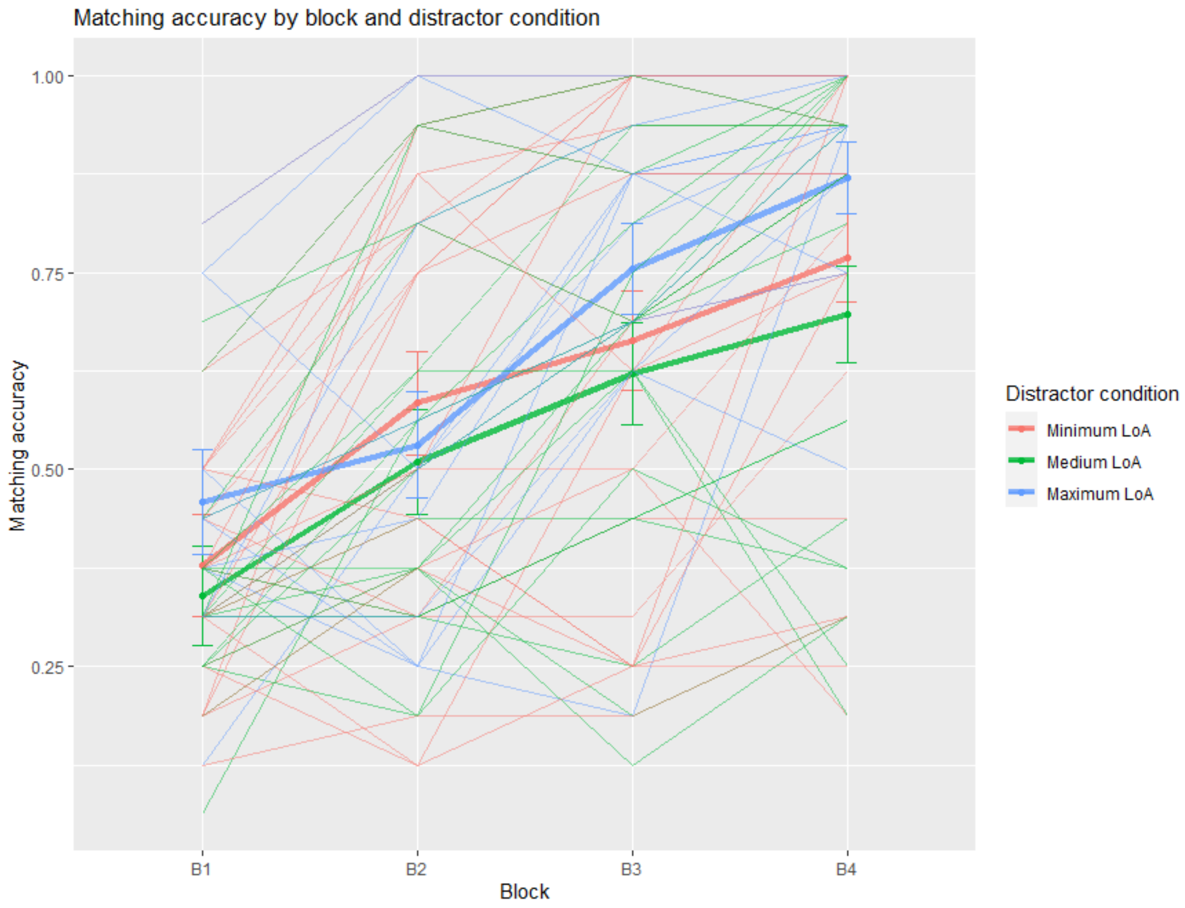
We explored whether there was any significant difference by distractor condition on matching accuracy. We found that accuracy increased block-by-block (Block 1 to 2, β = .753, z = 4.795, p <.001, Block 2 to 3, β =.771, z =4.994, p <.001, and Block 3 to 4, β = .799, z = 4.055, p <.001).

The first contrast explored the difference between “Minimum LoA” condition and “Medium LoA" condition. We did not find an overall difference in accuracy between these two conditions, β =-.169, z = -.830, p =.406, nor in the increase between conditions, all ps>.457.

The second contrast explored the difference between “Maximum LoA” condition and the average of “Minimum LoA” and “Medium LoA” conditions. We did not find a significant overall effect (β = .150, z = 1.328, p = .184). Regarding the block-to-block change, we found a marginally significant difference in the change from Block 2 to 3, suggesting that participants in the “Maximum LoA” condition had increased their matching score more than participants in the other two conditions (β = .207, z = 1.710, p = .087). However, we did not find this effect on the change between any of the other pairs of blocks, all ps>.125.
Finally, within-block pairwise contrasts showed no significant difference between any of the conditions within any of the blocks, all ps>.117. Figure 2.8 shows change in matching accuracy across blocks by distractor condition.

Figure 2.8.

*Matching accuracy within the interactive condition, by distractor condition.*



*Note.* Thicker lines represent group averages whereas thinner lines represent individual trajectories. Error bars represent standard error.

### *Exploratory analyses of the errors and the relationship between indices*

In an attempt to better understand the mechanism behind the impact of the distractor, we looked at the data from participants in the interactive condition and medium LoA distractor condition. Within this condition, participants could choose between the target, two distractors that had one dimension each in common with the target, and a distractor that had no dimension in common. If participants had perceived the compositional nature of the language, we would expect them to have been more likely to choose either of the distractors with one dimension in common with the target when they made a mistake. This was indeed the case: participants in the medium LoA condition were more likely to select the distractor that had the shape in common, then the distractor that had the filling in common, and then the distractor that nothing in common. In addition, we found a marginally significant negative correlation between the proportion of incorrect trials in which participants had chosen the distractor that had no dimensions in common with the target, and total accuracy in the production tests, rho(21) = -.416, p = .062. This suggests that those participants who perceived the structure of the language may have been more likely to produce accurate responses.

Across all blocks, retyping behaviour showed a significant moderate correlation with total accuracy in the production tests, rho(132) = .500, p =<.001, and with phonological accuracy, rho(132) =.476, p =.001. These correlations were stronger for participants in the non-interactive condition (rho(132) =.575, p <.001, and rho(132) <.536, p =.001, respectively) than for those in the interactive condition (rho(132) =.299, p =.006, and rho(132) =.365, p =.025, respectively).

Matching accuracy strongly predicted total accuracy across all blocks, rho(56) = .835, p <.001, and moderately typing behaviour, rho(56) =.451, p <.001 . Finally, in order to explore the effect of corrective feedback on learning, we examined the correlation between matching accuracy in Block 1, and total accuracy, rho(56) = .454, p <.001.

### Discussion

This first pilot allowed us to test whether participants could implicitly learn an artificial language through this new paradigm, without direct instruction. It also aimed to test specific features of this study, such as the difficulty of the artificial language we designed, or the number of trials required. Most of the participants had learned the language by block 4, especially those who had retyped correctly all the nouns throughout the task. We also observed an increase of not answered trials in Block 4, which probably reflects the effect of fatigue. This suggests that increasing the training might have a detrimental effect over performance. As the analysis of the errors in the Medium LoA condition, as well as the analysis of the grammatical errors and the total accuracy errors suggests (extended below), we also showed that participants are able to learn the structure of artificial language implicitly, without explicit instruction of the mapping between word elements and objects.

We found that the effect of interactivity over total accuracy was dependent on the distractor structure. We only found a positive effect of interactivity within the Maximum LoA distractor structure, which was the one that made the structure of the language most salient. The effect of interactivity over total accuracy was not visible within any of the other distractor structures. These results go in line with those from the matching task. Participants in the Maximum LoA were the quickest ones to reach ceiling performance, followed by those in the Minimum LoA condition. However, we had predicted that participants in the Maximum LoA condition would have the lowest matching score, as the probability of selecting the right option when knowing the name of one of the parameters was the lowest of the three distractor conditions. This could be due to the fact that the low salience of the structure in the Minimum LoA condition prevented participants from benefitting from the high probability of guessing the right answer when knowing only one of the parameters. This goes in line with the results of the exploratory analyses of the Medium LoA condition, which suggested that those participants who had perceive the structure of the language might have more accurate responses.

Overall, however, participants in the Medium LoA - interactive condition had the lowest performance, but this could be due to a lower engagement by participants in this condition, as their retyping behaviour was also overall worse than that of participants in all other conditions.

We also showed that this distractor structure itself was not enough to boost language learning, as participants in the non-interactive condition were also presented with trials in which these distractor structures were present. It was required for them to interact with these to get the benefit of the distractor structure.

Similarly, phonological accuracy seemed to be influenced by retyping behaviour, but not interactivity per se: we found an unpredicted difference in retyping between participants in the interactive and non-interactive condition within the Medium LoA distractor conditions, which later replicated on the phonological accuracy scores. Given the small number of participants on each of the distractor conditions, we cannot reach any conclusions on the reasons behind. It could be that participants in this particular condition were more engaged than participants in any of other conditions for reasons unrelated to the task. We also found that phonological accuracy started off higher in block 1 for participants in the non-interactive condition and that by block 2, participants in the interactive condition had caught up. This could be because the interaction task could have been distracting in the beginning, decreasing the accuracy scores, and later on boosted the phonological accuracy. Finally, grammatical accuracy was not affected by interactivity condition nor distractor condition.

There was a big limitation with one of the conditions in our study. Participants in the Medium LoA interactive condition could guess the right answer without paying attention to the wordform, as the correct answer would always be that item that contained two of the parameters that other items also contained. If that was the case, these participants would not pay attention to the wordform. However, this would only happen if they were aware of the structural characteristics of the meaning space. Given their lower scores in the matching task and the fact that their responses followed a normal distribution, we can argue that this was not the case. Furthermore, we would have expected a bimodal/skewed distribution on the scores, with those participants who had realised the limitation at ceiling, and those who had not, showing more average responses. In contrast, the number of participants at ceiling was in fact lower than in the two other conditions, across all of the blocks and within each of them.

From this study, we can see that interactivity, as defined in our study, may affect the mapping between meaning and wordform, but only when the structure of the language is made salient through the distractor structure, and it is not the probability of guessing the right answer or the presentation of the entire meaning space that affects it. Our exploratory analyses of the Medium LoA condition further reinforce this idea. Grammatical accuracy was high across conditions from the first block in this study. Hence, a ceiling effect does not allow us to identify any potential between-group differences. Phonological accuracy seems to be an independent linguistic feature, which is not affected by interactivity, but it is by retyping behaviour.

In our study, participants in both interactive and non-interactive condition were retyping the nouns and presented with possible distractors. Hence, participants in the non-interactive condition could be also trying to guess the right answer. In addition, the feedback for both of the groups was equal, that is, participants trying to match did not receive direct feedback on their performance, but were simply shown the correct answer. These aspects make the two conditions very similar in their level of interactivity. Experiment 2 attempted to address these limitations through some changes on the paradigm.

# Experiment 2

In Experiment 2, we adapted the paradigm so that the participants received explicit feedback on their matches and could keep a track of their performance in order to emphasise the interactive aspect of the task, with the aim of making it resemble a social interaction to a higher degree Equally, in an attempt to make the task more similar to the human interaction context in which languages are learnt, we framed the task as a game in which they played with a computer programme called L3arn, with whom they had to communicate in its language. In director-type trials, instead of retyping, participants in both conditions had to produce an answer, as it happens in interactions between people who do not speak the same language. In this study, we did not manipulate distractor-type. Instead, we used a variation of the Medium LoA structure, that retained the salience of the linguistic structure, but included a uniform ambiguity across trials. We did this by offering participants four options in which two of the distractors shared a shape and two of the distractors shared the filling. This way, participants could not guess the right answer only by knowing the compound for either the shape or the filling of the target object (like in the Minimum LoA condition of Experiment 1) nor only by looking at the distractor structure (like in the Medium LoA condition in Experiment 1), hence overcoming the limitations described for these distractor structures in the discussion of Experiment 1. This structure was based of Feher et al. (2016) and is developed further in the methods section.

## Methods

### *Participants*

78 first-year psychology undergraduate students took part in the study, all above the age of 18. As in the first study, we excluded those participants who had not learnt the phonological forms of the words by the last block. that is, those who had an average accuracy score of less than .5 in Block 4. Contrary to the case in Experiment 1, we did not exclude participants based on their score in interaction trials, as these trials involved free recall production, rather than retyping, and hence, they could not be used as a proxy measure for attention. Two participants from the interactive condition were excluded from the final sample, yielding a final sample of 76 participants, 37 in the interactive condition and 39 in the non-interactive condition.

### *Language structure and materials*

We used the same materials as in Experiment 1.

***Design***

We used a between-subjects design with one factor: interactivity. This factor had two levels: interactive or non-interactive. The difference between these two conditions is developed in the Procedure section. We also included block as a within-subject factor, with four levels.

***Task description***

As in Experiment 1, the task was an adaptation of the director-matcher task, composed of interaction trials (director- and matcher-type) and production tests. However, there were a few crucial differences on the framing of the task, the timing of the presentations, and the demands in the trials.

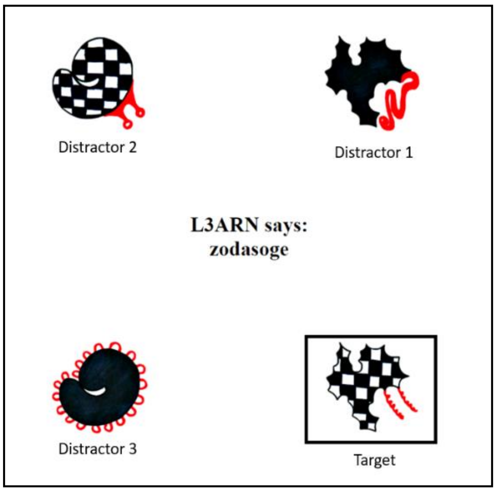
**Interaction trials.** In order to increase the salience of the interactive element of the task, we told participants that they would be interacting with a computer program called L3ARN, to whom they described objects (in the director-type trials), and who described objects for them (in the matcher-type trials). The director-type trials were equal for participants in both conditions, whereas the matcher-type trials were different for participants in the interactive and non-interactive conditions.

In matcher-type trials**,** participants were presented with four objects (the target object and three distractors) forming a square and they could read the prompt “L3arn says:” followed by a noun (the name of the target object) in the centre of the screen. Based on the data gathered in the first study, we designed a distractor-structure that emphasised the salience of the structural property of the language while keeping the level of ambiguity stable, so that the probability of guessing the right option was not determined by whether participants knew the mapping of the parameter that was kept stable or the one that was not.

The first distractor had the same shape as the target object and a different filling. The second distractor had the same filling as the target object and a different shape. The third distractor had the same filling as the first distractor and the same shape as the second distractor. Hence, it had no shape or filling in common with the target object (see Figure 2.9). The position of each object in the square was randomised for each trial. This design was based on Feher et al., (2016) and guaranteed that participants needed to be familiar with both components referring to shape and filling to be able to map the name to the object correctly, hence keeping the ambiguity high while increasing the salience of the language structure.

Figure 2.9.

*Distractor structure*

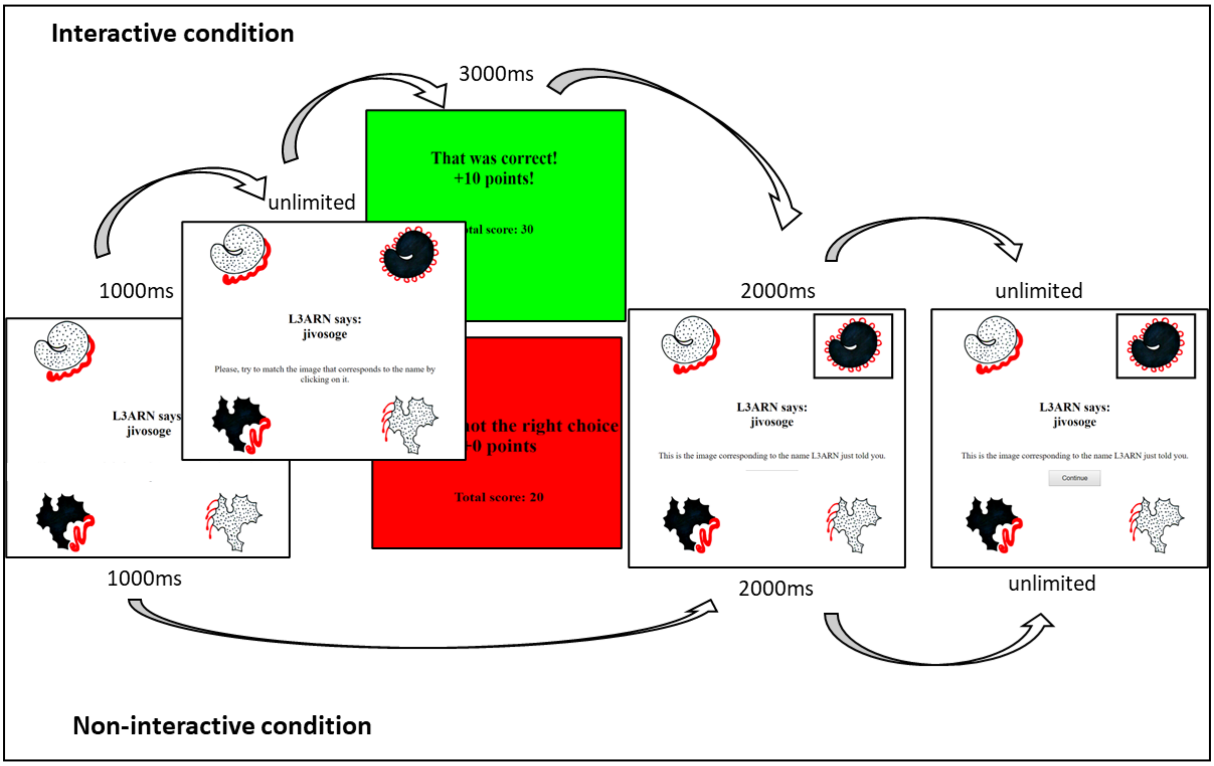


If they had been assigned to the interactive condition, after 1000ms, a prompt asking participants to press on the object they thought corresponded with the noun was added to the centre of the screen ("Please try to match the image that corresponds to the name by clicking on it"). They had unlimited time to click on one of the four images. If they had pressed the target object, they were presented with a green screen and a prompt saying: "That's correct! +10 points!". They could also see a counter showing how many points they had obtained in the ongoing block so far ("Total score: X"). The points were not exchangeable for any course credit, money or any other sort of compensation but were included to emphasise the feedback in the task and help participants keep a track of their performance. If participants pressed any other object than the target, they saw a red screen and a prompt saying: "That's not the right choice. +0 points.". They were still able to see to overall counter for the block. After 3000ms, participants returned to a screen with the four objects forming a square and the target noun in the centre preceded by "L3arn says:". This time, the target object was framed, and a prompt saying "This is the image corresponding to the name L3arn just told you" was added to the centre of the screen. After 2000ms, a "Continue" button was added to the screen, which they could press whenever they were ready to, in order to proceed to the next trial.

If participants had been placed in the non-interactive condition, they were presented with the screen containing the four objects in a square and the noun of the target object for 1000ms. After that, the target object was framed and the prompt "This is the image corresponding to the name L3arn just told you" was added to the centre of the screen. Finally, after 2000ms, a "Continue" button appeared on the screen, which participants could press whenever they felt ready to move on to the next trial. We included a 1000ms blank screen between trials (see Figure 2.10 for a visual representation of a matching trial).

Figure 2.10.

*Visual representation of the matching trials.*

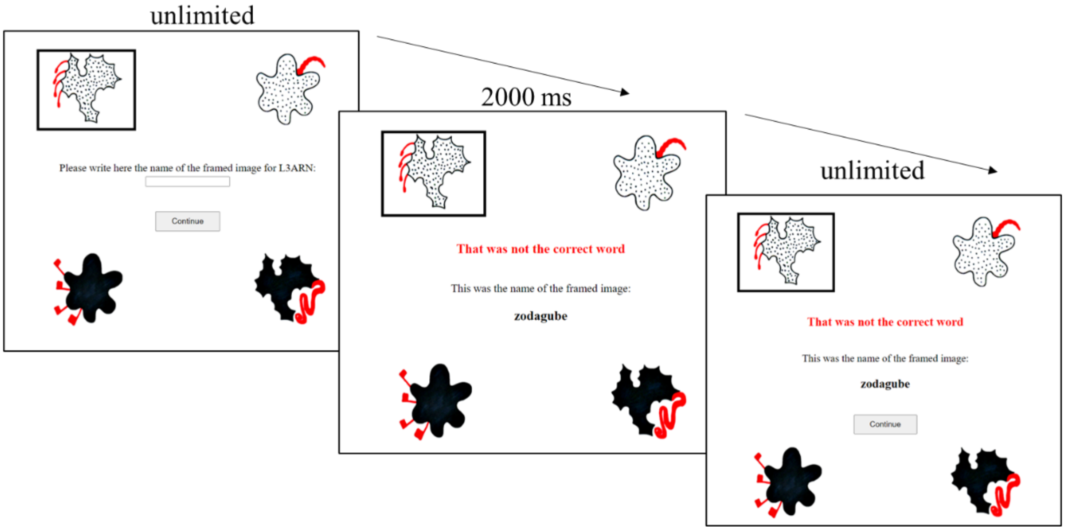


*Note.* Follow the upper arrows for the interactive condition and the lower arrows for the non-interactive condition.

In the director-type trials, participants were again presented with four objects (the target object and three distractors) forming a square. In contrast with Experiment 1, participants were asked to produce what they thought was the right answer instead of simply retyping the target noun. The target object was framed, and they could read a prompt asking them to write the name of the framed object “Please write here the name of the framed image for L3ARN:”. Participants had unlimited time to produce their response. When participants gave their response, they received feedback on their performance. If their response was incorrect, they could see a prompt saying “That was not the correct word” in red font in the centre of the square, whereas if their response was correct, the prompt said “CONGRATULATIONS! You typed the right word for L3ARN!” in green font. In either case, a prompt was also added indicating which was the correct production (i.e., “This was the name of the framed image:" followed by the correct noun). If they produced less than 4 characters, their response was not accepted, and they were asked to try to guess the answer. This was to prevent participants from skipping through director-type trials. After 2000ms, a "Continue" button was added to the screen, which participants could press whenever they were ready to proceed. We included a 1000ms blank screen between trials. Figure 2.11 shows a visual representation of director-type trials for both conditions.

Figure 2.11.

*Structure of the director-type trials.*



Please write here the name of the framed image for L3ARN:

Continue

That was not the correct word

This was the name of the framed image:

**zodagube**

That was not the correct word

This was the name of the framed image:

**zodagube**

Continue

**Production test.** Production test was identical to that in Experiment 1.

***Procedure***

Participants performed the task in individual computers in the lab, in sessions of between 2 and 15 participants. They were randomly assigned to a condition at entrance.[1] They were offered an information sheet and a consent form. After gathering the consent form from all the participants in the session, participants started the task at the same time. The number of

[1] The computers were prepared in the room for the number of invited participants. Every computer was randomly assigned a condition, ensuring a roughly equal number of computers in each condition per session. Computers in the same row were set up to the same condition to prevent participants from contrasting their condition with their partner's. However, the row that was set to a condition in a given session was set to the other condition in the next session. Participants came to the room and took a seat in any of the prepared computers, naturally spreading across the room.

trials and sequence of events was identical to that in Experiment 1. The average time for completing the study was 58.14 minutes (SD = 12.23).

***Pre-processing and measures***

The data pre-processing and the collected measures were similar to those in Experiment 1. We calculated total accuracy, phonological accuracy, and grammatical accuracy for Production trials, following the same procedure as in Experiment 1. However, we did not include a Retyping Accuracy measure. As director-type trials in Experiment 2 did not involve retyping but free recall, we treated them as we did with the production trials, calculating the total accuracy, grammatical accuracy, and phonological accuracy of the productions.

***Statistical analyses***

The analysis strategy followed the same logic as that of Experiment 1, with some differences based on the changes in design. The models for Production trials (with total accuracy, grammatical accuracy, and phonological accuracy as the outcome variables) had the same structure as in Experiment 1, excluding those fixed effects and random slopes related to distractor condition, and of course retaining interactive condition, block and their interaction as fixed factors. Similarly, we ran these same three models for the productions in the director-type trials, this time adding the position of the trial within the block as a fixed factor instead of a random factor, given that participants received feedback between trials and were hence expected to improve their scores across the block. We also included the interaction between trial position and block, trial position and condition, as well as the triple interaction between the three. Finally, as there was no manipulation within the interactive condition, in contrast with Experiment 1, we did not run a model in matching accuracy (the accuracy on the guesses of participants in the interactive condition across the matcher-type trials), using this variable exclusively for our exploratory analyses.

**Results**

***Total accuracy***

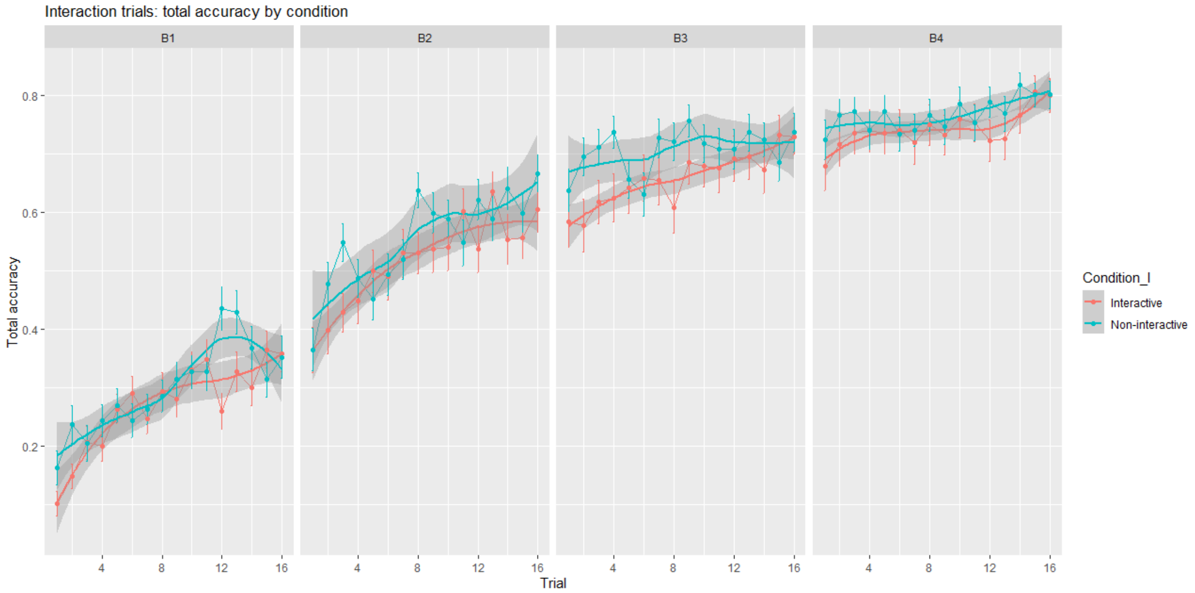
**Interaction trials.** As Figure 2.12 shows, total accuracy increased block-by-bock (β =.245, z = 10.386, p <.001, from Block 1 to Block 2, β = .216, z = 10.299, p <.001, from Block 2 to Block 3, and β = .087, z = 4.725 , p <.001 from Block 3 to Block 4), and trial-by-trial within each of the blocks, β = .010, z = 12.870, p<.001, though the trial-by-trial increase decreased from Block 2 to Block 3, β = -.008, z = -4.582, p<.001. That suggested that participants were reaching an asymptote around Block 3, after which the increase in total accuracy slowed down. We did not find any other interactions between block-by-block change and trial, all ps> .339.

A within-block nested analysis of the interaction between trial and condition showed that, within Block 3, the trial-by-trial increase of total accuracy in participants in the interactive condition was higher than the one for participants in the non-interactive condition, F =4.472, p = .0345. This was followed by a marginally significant difference in the block-by-block change from Block 3 to Block 4 (β =.031, z = 1.692, p = .091). There was some emerging evidence that participants in the interactive condition showed a higher increase in total accuracy than participants in the non-interactive condition.

We did not find any other effect of interactivity condition neither across all blocks (β = -.024, z = -1.057, p =.291), nor within any of the blocks (all ps>.326), nor in the change between any of the other blocks (all ps>.153). Finally, we did not find any interaction between the effect of trial, block-by-block change, and interactivity condition, all ps>.105.

Figure 2.12.

*Total accuracy in interaction trials by condition*

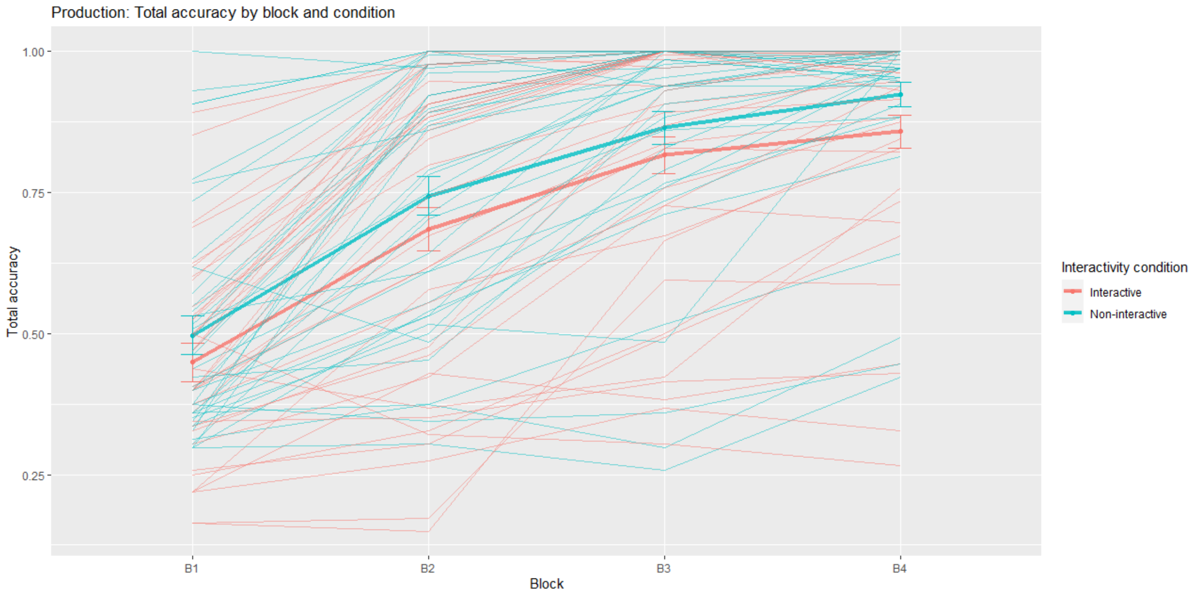


*Note*. Error bars represent standard error. Individual lines represent individual trajectories.

**Production tests.** As in the interaction trials, total accuracy increased block-by-block (β =.241, z = 11.74, p <.001, from Block 1 to Block 2, β = .126, z = 8.26, p <.001, from Block 2 to Block 3, and β = .050, z = 4.39, p <.001 from Block 3 to Block 4). We did not find an effect of interactivity over total accuracy on production trials overall (β = -.028, z =-1.33 , p =.185), nor within any of the blocks, all ps>.169,  nor in the block-by-block increase, all ps>.474. Figure 2.13 shows these results.

Figure 2.13.

*Total accuracy by block in the production trials*



***Phonological accuracy***

**Interaction trials.** As Figure 2.14 shows, phonological accuracy also increased gradually over training (β =.193, z = 13.925, p <.001, from Block 1 to Block 2, β = .058, z = 4.887, p <.001, from Block 2 to Block 3, and β = .032, z = 2.914, p =.003 from Block 3 to Block 4), and across trials within each of the blocks, β = .006, z = 16.903, p<.001. The increase across trials decreased from Block 1 to Block 2, β = -.010, z =-9.276, p<.001,  from Block 2 to Block 3, β = -.003, z = -2.508, p=.012, but not from Block 3 to Block 4, β = -.002, z = -1.596, p = .110, revealing a logarithmic pattern that plateaued in Block 3.

The effect of interactivity condition was significant on the increase from Block 1 to Block 2, β =.033, z = 2.372, p=.014, and from Block 3 to Block 4, β =023, z=2.096, p =.036, but not from Block 2 to Block 3, β = -.012, z = -1.003, p = .317. Participants in the interactive condition showed a higher improvement in their phonological accuracy score from Block 1 to Block 2 and from Block 3 to Block 4 than participants in the non-interactive condition. Equally, participants in the interactive condition increased their phonological accuracy score more than participants in the non-interactive condition across trials within each of the blocks, β <.001, z = 2.462, p =.014. In addition, we found a marginally significant interaction between block-by-block change, trial, and interactivity condition from block 1 to block 2, β =-.002, z = -1.934, p = .053, and from Block 3 to 4, β = -.002, z = -1.668, p =.095, but not from Block 2 to 3, β =- .002, z= -1.548, p =.122. These suggested that the within-block trial-by-trial increase in phonological accuracy might have been lower for participants in the non-interactive condition in comparison with the one for participants in the interactive condition in these pairs of blocks.

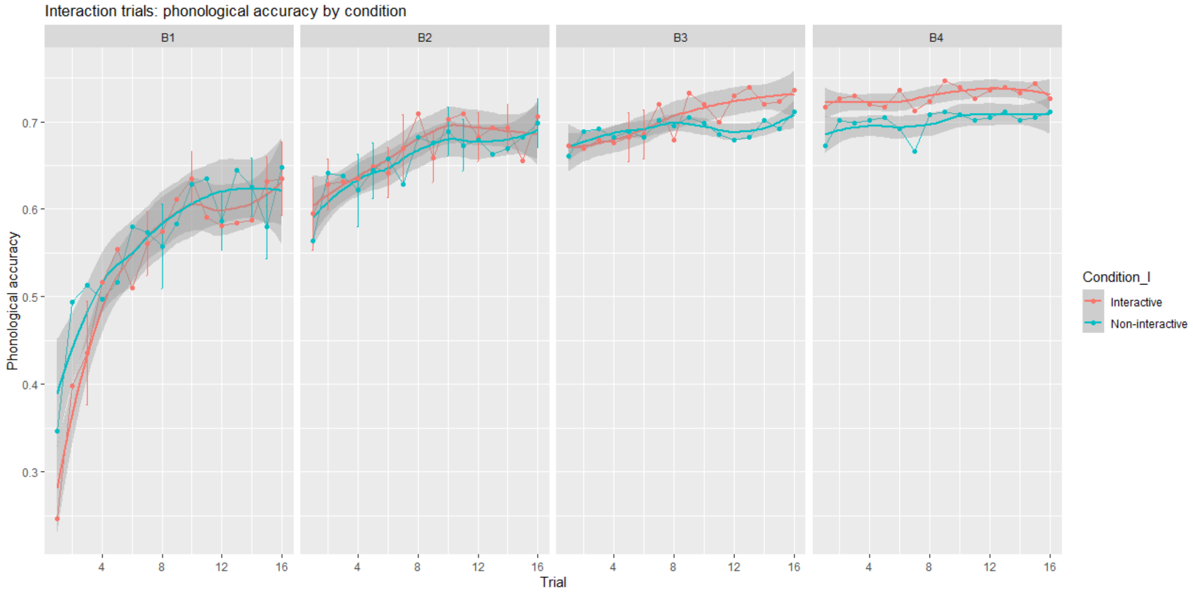
To better understand these interactions, we followed up with a by-block nested analysis of the interaction between trial and interactivity condition. We found an interaction between trial and condition within Blocks 1 and 3 (F = 7.711, p = .006, and F = 4.979, p = .026). Participants in the interactive condition increased their score more than participants in the non-interactive condition within these blocks. The interaction was not significant, however, within Blocks 2 and 4 (all ps>.899). Finally, condition did not show overall effect across blocks, β = -.004, z =-.154, p =.878, nor within any of the blocks, all ps>.591.

Taken together, these results suggest that phonological accuracy increased following a logarithmic pattern and reaching the asymptote in Block 3. Participants in the interactive condition however, continued improving their phonological accuracy score across Block 3, leading to a higher average increase from Block 3 to 4. Equally, they also improved their score faster across Block 1, leading to an average higher increase from Block 1 to Block 2. The effect of condition, however, was not significant across trial nor within any of the blocks.

In summary, the effect of interactivity was not present on the average final scores nor within any of the blocks, but their learning pattern was different. Participants in the interactive condition started from a lower baseline and increased their score more within Block 1, catching up with participants in the non-interactive condition by Block 2. Participants in the non-interactive condition reached an asymptote on Block 3, whereas participants in the interactive condition continued learning, until reaching an asymptote in Block 4.

Figure 2.14.

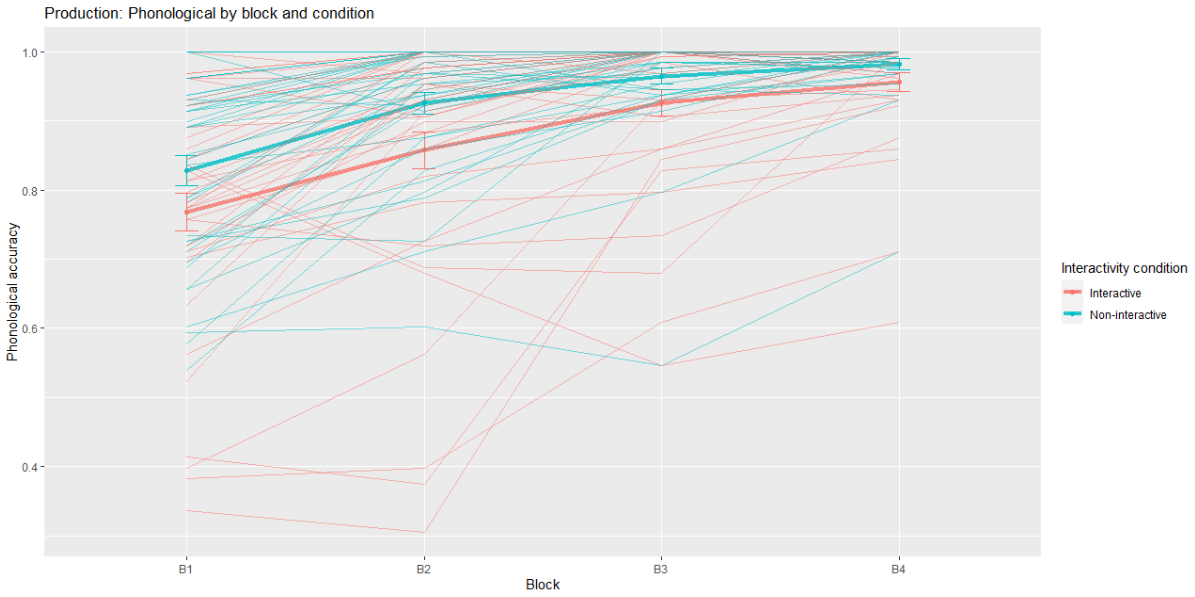
*Phonological accuracy by condition in the interaction trials*



**Production tests.** As Figure 2.15 shows, phonological accuracy increased block-by-block (Block 1 to Block 2, β =.094, z = 7.90, p<.001, Block 2 to 3, β =.054, z =4.50, p <.001, Block 3 to 4, β = .024, z = 3.71, p <.001). Contrary to what we predicted, participants in the non-interactive condition showed a higher phonological accuracy than those in the interactive condition across blocks, β = -.024, z = -2.11, p =.035. Within-block analyses showed that the effect was led by a significant difference between conditions in Blocks 1 and 2 (β = -.060, t = -1.963, p = .0497 and β = -.068, t = -2.577, p = .011, respectively), but not within Blocks 3 and 4 (all ps>. 145). There was no effect by condition in the block-by-block increase, all ps>.222. This suggests that participants in the non-interactive condition were more likely to produce phonologically accurate responses from the first block, and that the difference between conditions gradually became smaller until disappearing in Block 3.

Figure 2.15.

*Phonological accuracy by condition in the production trials*



### *Grammatical accuracy*

**Interaction trials.** Grammatical accuracy did not change block-by-block, all ps>.811, nor across trials within each of the blocks, β = .044, z = -.001, p = .999. Interactivity condition did not have any effect on it overall, nor within any of the blocks, nor in the block-by-block increase or in the increase across trials, overall or in the block-by-block increase, all ps>.341.

**Production trials.** Grammatical accuracy did not change block-by-block, all ps>.632. Interactivity condition did not have any effect on it overall, nor within any of the blocks, nor in the block-by-block increase, all ps>.823.

### *Exploratory analyses in the relationship between variables*

As it would be expected, total accuracy in the interaction trials strongly correlated with total accuracy in production tests, rho(78) =.806, p <.001. We also examined the correlations between matching accuracy and total accuracy for participants in the interactive condition. Matching accuracy and total accuracy showed a stronger correlation in the production test, rho(37) = .937, p <.001, than in the interaction trials, , and rho(37) = .828.

Finally, as in Experiment 1, in order to explore the effect of corrective feedback on learning, we examined the correlation between matching accuracy of participants in the interactive condition in Block 1, and total accuracy and found a moderate positive correlation both for production tests, rho(37) = .491, p = .002, and for interaction trials, rho(37) = .418, p = .010.

**Discussion**

We expected participants in the interactive condition to produce more accurate responses in the production tests than the participants in the non-interactive condition, however, overall, we did not find any significant differences on total accuracy between conditions in the interaction trials nor in the production test. However, the results in the interaction trials suggest that learning trajectory for participants in each condition was different, with participants in the non-interactive condition reaching an asymptote in the third block and participants in the interactive condition further improving their total accuracy score in the last block. We also found that, in the production tests, the productions of participants in the interactive condition had a lower phonological accuracy than those by participants in the non-interactive condition, but that the difference disappeared by the third block. We found a similar pattern for phonological accuracy in the interaction trials, where that of participants in the interactive condition overtook that of participants in the non-interactive condition from Block 3. In combination with the results in total accuracy, that indicates that, in the first two blocks, participants in the interactive condition were more likely to produce wordforms that did not exist in the artificial language when they made a mistake, whereas participants in the non-interactive condition were more likely to produce an incorrect but existing wordform. The results also show that grammatical accuracy was high and stable over time, and that it did not differ by condition.

One of the possible reasons for these results could be that the changes in the paradigm relative to Experiment 1 increased the cognitive load for participants in the interactive condition, as the trials for this condition contained a higher number of screens to process in every trial. This would explain the lower phonological accuracy score in the first two blocks, which could have delayed the learning of participants in the interactive condition. This is further supported by the difference in learning trajectories in the interaction trials, in which participants in the non-interactive condition reach an asymptote before those in the interactive condition, who seem to start overtaking them in Block 4, as suggested by the marginal difference between conditions.

Equally, literature from the field of memory research could give us an insight on the results. When participants are given a question and asked to generate a response, they are more likely to be able to recall it later. This phenomenon is referred to as the “generation effect" (Slamecka & Graf, 1978) and it is often found not immediately after learning, but in a delayed retest (Mulligan & Peterson, 2015). Given the similarities between this paradigm and ours, which asks participants in the interactive condition to generate a response, the expected effect may only appear after a delayed testing. Experiment 3 aims to address these limitations.

## Experiment 3

In order to overcome the limitations stated in the discussion, we run a follow-up experiment with 150 undergraduate students, simplifying the text in the interactive condition, and the extra screens that made it differ from the non-interactive condition, and correcting the trial timing so that it was equal between conditions. We also added a delayed test, 24 to 48h after the first one, in which we showed participants the pictures of the 16 objects and asked them to type their name. We asked them to report the estimated hours of sleep the night between the tests. Finally, we increased the sample size.

### Methods

#### *Participants*

165 undergraduate students took part on Day 1 of the study. We excluded 15 participants from Day 1 analyses (9 from the interactive condition and 6 from the non-interactive condition), as the mean of the Levenshtein distance of their productions in the testing phase of the fourth block was of more than 4 for all the possible words. The final sample for day 1 had 150 participants, 81 in the non-interactive condition and 69 in the interactive condition. Of those, 125 completed both Day 1 and Day 2. Here we analyse the data of those participants that completed both days, but analyses of the data for Day 1 with the full sample yielded similar results (see Appendix A).

#### *Language structure & materials*

Same as for Experiments 1 and 2

#### *Design*

The design was the same as for Experiment 2, but with Day (within-participants, 1 vs. 2) added as a factor.

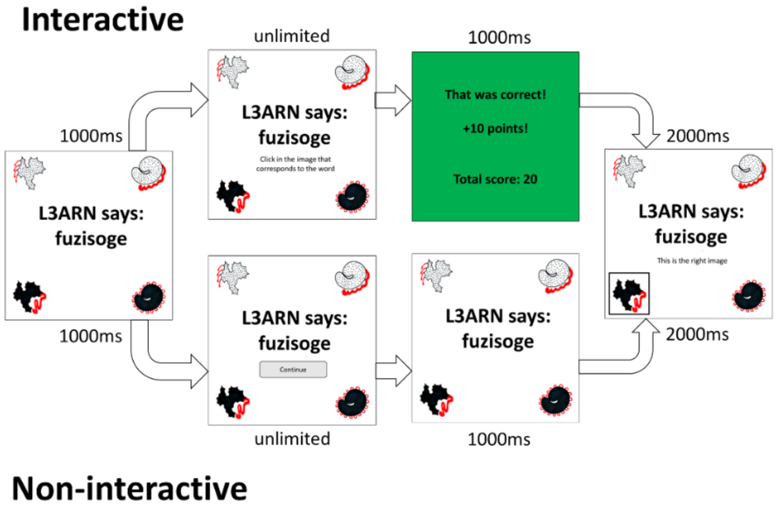
#### *Task*

The task was the same as for Experiment 2, but the timing was altered for the non-interactive condition to eliminate the confounding difference between conditions. The text was simplified for both conditions to reduce cognitive load and facilitate comprehension.

**Interaction Trials.** The distractor structure in the matcher trials was the same as in Experiment 2. Participants in the interactive condition were asked to click on the image that matched the label of the screen (Figure 2.16). If they clicked on the target object, they saw a green screen that said “That was correct! +10 points!”. If they clicked on any of the distractors, they saw a red screen that said “Not the right choice. +0 points”. This screen also indicated the total points so far in that block. Participants in the non-interactive condition were asked to press “Continue” to see the image that matched the name. When they clicked the button, the prompt and the button disappeared, leaving on screen the four objects and the label of the target object. As shown on Figure 2.16, the next screen for all participants contained the four objects with the target framed in a square and the correct label in the centre. After 2000ms, a button to proceed to the next trial was enabled.

Figure 2.16.

*Structure of the matcher-type trials.*



*Note.* Top panels illustrate the matching procedure in the interactive condition, while bottom panels show the steps in the non-interactive condition.

In the director trials, which were identical in the two conditions, participants were again presented with four objects forming a square. The target object was framed, and they were asked to type the label for it. To prevent participants from skipping through, they were not allowed to proceed until they produced at least 4 characters. They read "CONGRATULATIONS! Right word for L3arn!" in green font if their response was correct and "Incorrect word" in red font if it was not. In either case, they were shown the correct noun. After 2000ms, a button to proceed to the next trial was enabled.

**Production Test.** Same as in Experiments 1 and 2.

### *Procedure*

The study received ethical approval from the University of Warwick's Humanities & Social Sciences Research Ethics Committee on the 24th of January 2022, and was conducted online through a custom website, hosted on Warwick University servers. The general procedure was the same as that for Experiments 1 and 2, with participants being presented with four blocks of 32 interaction trials, each followed by a 8 trial production test.

Twenty-four hours after they had completed the first part of the study, participants were invited to complete the second part within the next 24 hours, after which they could no longer access the study. In the second part of the study, participants were asked how many hours they had slept the previous night and then completed a production test. Day 1 took participants 41.63 minutes to complete on average (s.d. = 11.3) whereas Day 2, which only involved 1 production test block, took an average of 3.14 minutes (s.d. = 2.10).

### *Measures*

The measures were the same as those in Experiment 2.

### *Statistical analyses*

We used the same analyses strategy as for Experiment 2. For the analysis of the data in Day 2, we run a linear mixed effects model with total accuracy in the production task as the outcome, fixed effects for day (Block 4 of Day 1 vs. Day 2) and condition (interactive vs. non-interactive), random intercepts for participant, hours between tests, hours of sleep, and target word, and by-participant random slopes for Day and condition.

## Results

***Total accuracy, Day 1.*** During interaction trials, all participants' productions became more *accurate* block-by-block (Block 1 to Block 2, $\beta = .398$, $z = 20.94$, $p <.001$, Block 2 to Block

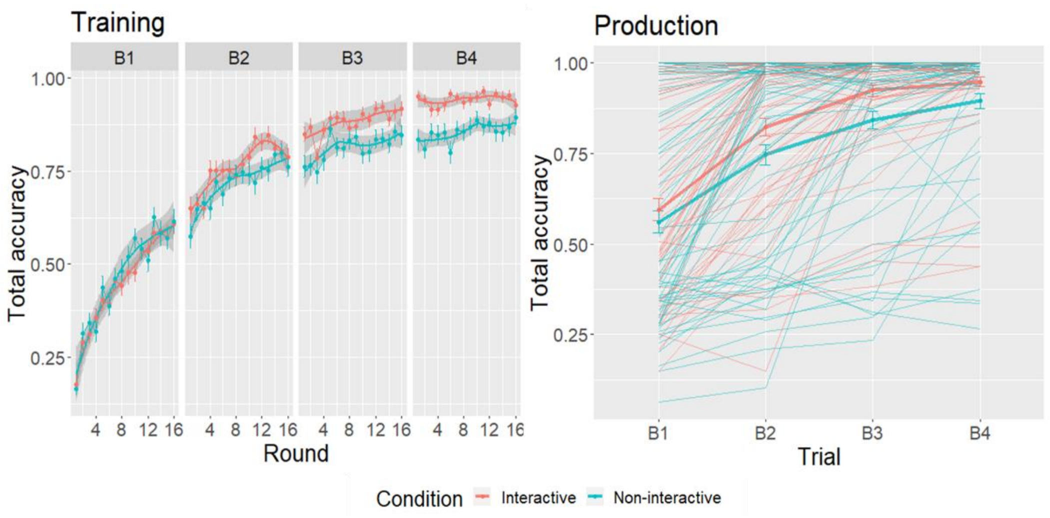
3, β = .159, z = 9.35, p <.001, and Block 3 to Block 4, β = .076, z = 4.51, p =.001), and trial-by-trial within each of the blocks (β = .011, z = 14.81, p <.001).

We found a marginally significant difference between conditions in total accuracy which suggested that participants in the interactive condition, overall, may have produced more accurate responses than those in the non-interactive condition (β = .026, z =1.75, p =.079). The nested block-by-block follow-up analysis showed that the difference was led by a significantly higher total accuracy score in the interactive condition than in the non-interactive condition within blocks 3 and 4 (significant difference between conditions within blocks 3 (F = 4.219, p = .046) and 4 (F = 5.266, p = .022), but not within block 1 and 2, all ps >.900. We did not find any significant differences between conditions in the block-to-block increase, all ps>.149, nor any triple interaction between any of the block-to-block changes, position of the trial and condition, all ps>.333. The within block-nested analyses did not show a significant interaction between trial position and condition within any of the blocks, all ps >.222.

In the trials in the production test, we also saw a block-by-block increase in total accuracy score, (Block 1 to Block 2, β = .207, z = 11.727, p <.001, Block 2 to Block 3, β = .099, z = 6.576, p <.001, Block 3 to Block 4, β = .039, z = 4.104, p <.001). When looking at the effect of interactivity, we found that, overall, participants in the interactive condition produced marginally more accurate responses than those in the non-interactive condition (β = .029, z = 1.74, p =.082). Further analyses showed that this effect was led by a within-block difference by condition in Block 2 (β = .074, t = 1.92, p = 0.054), and Block 3 (β = .080, t = 2.08, p = 0.038), despite no difference by condition in the block-by-block increase (all p's > .117), nor within Blocks 1 and 4 (all p's over .208). In summary, participants from the interactive condition showed a higher level of total accuracy within the last blocks of training, with no significant differences between conditions at the initial blocks of training.Figure 2.17 shows total accuracy scores by Block in interaction (Training) and production.

Figure 2.17.

*Total accuracy by interactivity condition and Block.*



*Note.* The left panel shows the average total accuracy by condition for each trial within each of the four interaction blocks. The right panel shows the total accuracy by condition and block, with thicker lines showing the average by condition for each block (and the error bars the standard error) and individual thinner lines showing individual averages for that block.

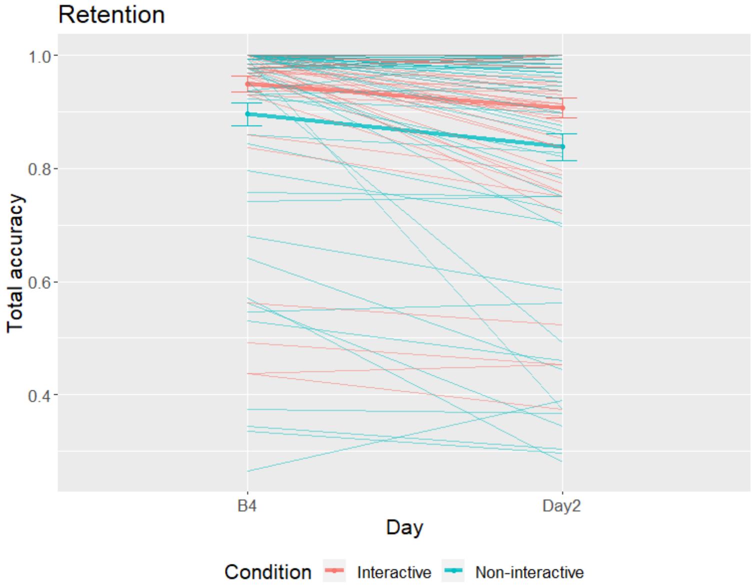
### *Day 1 vs Day 2, Production Test*

Next, we wanted to explore how persistent the effect of interactivity was, by comparing the total accuracy score in the production test of the fourth block on Day 1 with the test on Day 2. Day 2 test was completed between 24 and 48 hours after the initial test (Mean = 30.87, SD = 7.98) and after an average of 7.52 hours of sleep (SD = 1.56). Participants in the interactive condition waited an average of 3.72 hours more between tests than participants in the non-interactive condition, F(1, 112) = 6.57, p =.012, however, there was no correlation between hours between tests and total accuracy on Day 2, rho(123) = -.048, p = .594. There was no difference between conditions in the number of hours of sleep, F (1, 121) = 1.05, p = .309, and this did not correlate with total accuracy either, rho(125) = -.059, p = .514. Overall, the total accuracy of participants' productions decreased from Day 1 to Day 2 (β = -.025, z = 5.79, p < .001), but, across days, participants in the interactive condition had a higher accuracy score than participants in the non-interactive condition (β = .032, z = 2.08, p = .037), with a significant difference in accuracy on Day 2 (β = .0738 , t =2.29 , p =.022), and marginally significant difference on Day 1 (β = .056, t =1.72, p = .085).

However, there was no difference in the decrease in accuracy between conditions (β =- .004, z = -1.05, p =.292). Therefore, the forgetting rate was comparable in the interactive and non-interactive conditions (Figure 2.18).

Figure 2.18.

*Total accuracy in Production Test in the fourth block of Day 1 and in Day 2.*



*Note.* Thicker lines show the average total accuracy by condition for each and day. Error bars represent standard error of the mean, individual thinner lines showing individual averages for that day.

### ***Phonological accuracy***

In the interaction trials, we found a marginally significant effect of interactivity (β = .016, z =1.80, p =.072), that seemed to be led by a marginally significant interaction between trial position and interactivity within Block 1 (F = , 3.33, p = .068): participants in the interactive condition might have improved their phonological accuracy to a higher extent than participants in the non-interactive condition across the trials in Block 1. This interaction was not present within any other blocks (all ps >.262), and we did not find a general effect of condition within any of the blocks (all ps >.134), nor over the block-to-block increase (all ps. > 222).

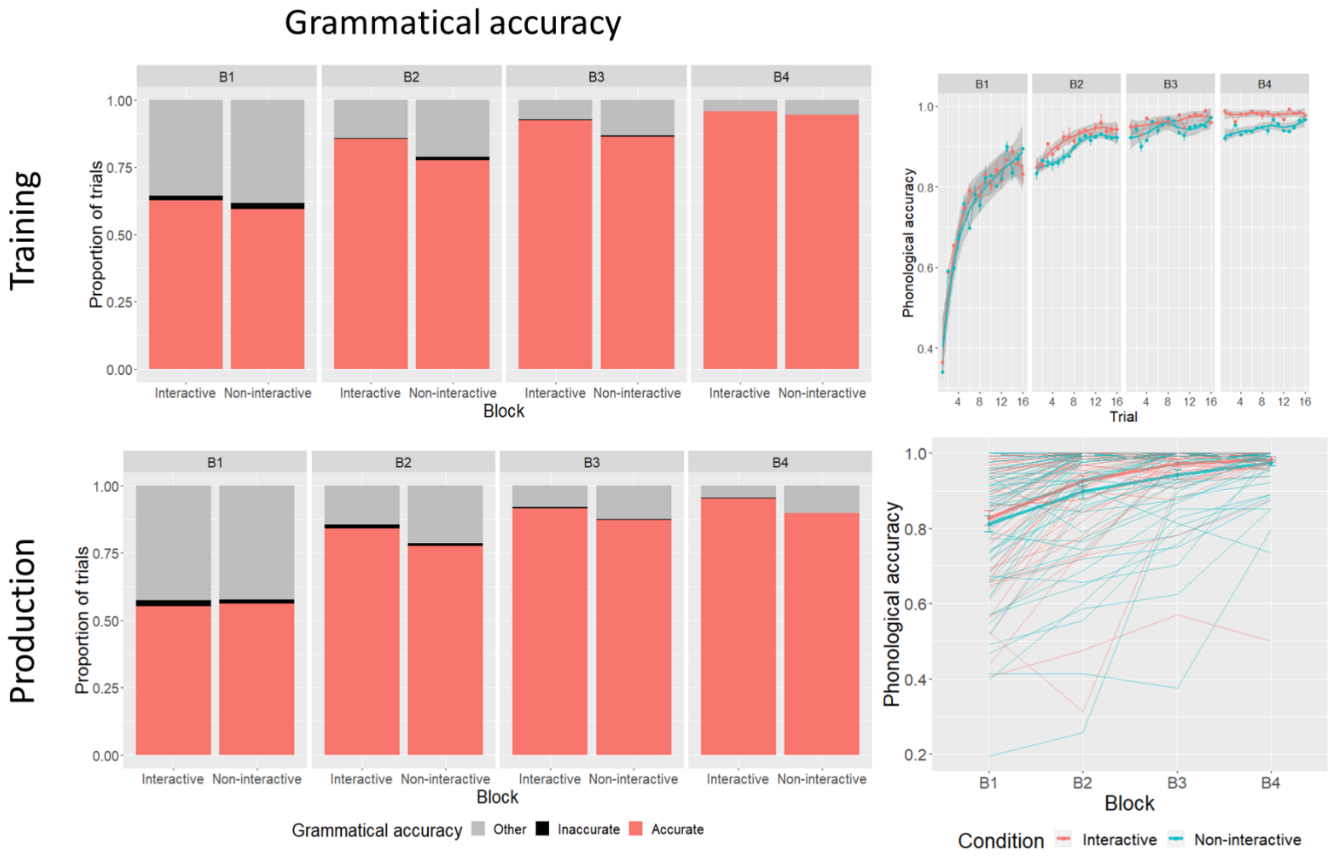
In the production trials, we did not find an overall effect of interactivity, nor block-to-block, nor within blocks, or trial-by-trial (all ps>.121). Once again, we did not find an effect of interactivity on overall learning or within any of the blocks (all ps >.219). However, the increase in phonological accuracy from Block 3 to Block 4 was higher for participants in the non-interactive condition than that of participants in the interactive condition (β = -.011, z=-1.98, p = .048). We did not see an effect of interactivity on the block-by-block increase for any of the other blocks (all ps >.524).

### *Grammatical accuracy*

Regarding the grammatical accuracy of participants' productions in the interaction trials, we did not find an overall effect of interactivity either, nor any interaction between interactivity and block, nor between interactivity and trial (all ps =.119). In the production test, grammatical accuracy did not change block-by-block, or by interactivity condition. The interaction between interactivity and block was not overall nor in the block-to-block change (all ps > .420), or within any of the blocks (all ps>.351). Figure 2.19 shows the results in Phonological and Grammatical accuracy.

Figure 2.19.

*Grammatical and phonological accuracy by interactivity condition*



### *Exploratory analyses in the relationship between variables*

As it would be expected, total accuracy in the interaction trials strongly correlated with total accuracy in production tests, rho(125) =.943, p <.001. Similarly, the correlation between matching accuracy and total accuracy for participants in the interactive condition was high, rho(60) = .844, p <.001 for production tests, and rho(60) = .893, p <.001 for interaction trials.

Finally, as in Experiments 1 and 2, in order to explore the effect of corrective feedback on learning, we examined the correlation between matching accuracy in Block 1, and total accuracy and found a moderate positive correlation both for production tests, rho(60) = .615, p <.001, and for interaction trials, rho(60)= .664, p = <.001

## Discussion

In Experiment 3, we observed an effect of interactivity over total accuracy towards the end of Day 1. More importantly, the effect persisted (and increased) when participants were retested a day later. However, both groups of participants learnt the structure of the language (grammar) and the actual wordforms to the same extent. In this experiment, we found some of the predicted effects. We discuss possible explanations, the relationship between these results and the ones of Experiment 1 and Experiment 2, and the implications of this study in the “General Discussion” section.

## General discussion

Our participants learnt an artificial language by interacting with a computer in a director-matcher game. We manipulated the level of interactivity of the game, observed the impact of participants being asked to guess the object that matches a wordform before the computer tells them over the total accuracy (whether they produce the right label for the right object), phonological accuracy (whether they produce an existing wordform in the artificial language), and grammatical accuracy (whether the structure of the wordform they produce is present in the artificial language).

We aimed to develop a paradigm that would allow us the observe the learning process while it happened. In Experiment 1, we first piloted the stimuli and paradigm, and observed that, the effect of interactivity was only present when the structure of the language was made salient by asking participants to interact with a distractor set that showcased the possible variations of one parameter within another parameter. In Experiment 2, we piloted a distractor structure that both increased the salience of the structure of the language and changed the paradigm mainly by 1) making learning fully implicit by asking all participants to produce an answer instead of retyping the right answer, 2) increasing the salience of interactivity on the interactive condition. We did not find the predicted results, possibly due to an undesired difference in cognitive load between conditions. Finally, in Experiment 3, we corrected the issues that we detected in Experiments 1 and 2 and found the predicted results: interactivity boosted language learning. However, we did not find any effect of interactivity over phonological accuracy and grammatical accuracy in either of the experiments (except for an effect of phonological accuracy in Experiment 1 led by differences in retyping accuracy).

Even though, with our paradigm, we were able to focus on a single aspect of interaction, clearly defining our manipulation, there are still several potential explanations to this effect. It could be the case that participants in the interactive condition were more engaged in the task leading to a higher level of motivation, as proposed by Walton et al., (2012). However, that would have arguably resulted in a difference in the overall learning of the wordforms, which we did not find. In addition, the time participants took to complete the task did not differ by condition, showing no evident difference on the time they spent looking at the stimuli, whether they had to select which image to click in the matching trials, or they could proceed to the next screen when desired.

Another hypothesis is that the feedback could have made participants more aware of the compositional structure of the language by encouraging a closer inspection of the different items to select the right option. This would in turn promote chunking (Miller, 1956) which is known to facilitate second language acquisition (Ellis, 2001). Results from Experiment 1 suggest that this might be the case, as the effect of interactivity was only present in the condition that emphasised the structural salience; similarly, the exploratory analysis of the Medium LoA condition suggested that a higher awareness of the structure of the language could predict accuracy. In contrast, grammatical accuracy of the participants' productions did not differ by condition in any of the three experiments, which we would have expected if participants in the interactive condition had an increased awareness of the grammatical structure of the language due to its higher salience in this condition. Nevertheless, grammatical accuracy exhibited very little change block-to-block and was at ceiling from Block 1 for most participants, regardless of the condition. Further studies using an artificial language with a more complex grammar could help understand the effect of interactivity over the learning of grammar.

Some research suggests that negative feedback increases the learning opportunities (Pica, 1992, 1994; Dale & Christiansen, 2004; Krishnan et al., 2018). Then, participants in the interactive condition could have benefitted from the additional opportunity for negative feedback stemming from the matching trials. However, follow-up analyses showed that those who made more mistakes in the first block showed a lower level of accuracy, rather than a higher one, in opposition to this hypothesis.

Alternatively, it could be the case that participants in the interactive condition learnt better because we explicitly asked them to generate responses, regardless of the feedback they were later provided with. Research in the field of memory shows that when participants are asked to generate an answer, they are more likely to retain it than when they read it or it is provided to them by the experimenter, a phenomenon known as "generation effect" (Slamecka & Graf, 1978). In line with our study, the size of the "generation effect" has been shown to increase when participants are asked to recall the answers one day or more after learning them (see McCurdy et al., 2020 for a meta-analysis). One of the proposed explanations to the effect is the Lexical Activation Theory, according to which generating a response requires an activation of previous knowledge and an integration of the answer within it, improving its semantic encoding. Consequently, the effect is lower for the encoding of pseudowords with no meaning than for real words (Gardiner & Hampton, 1985, McCurdy et al., 2020). The effect has not been tested in the context of learning pseudowords that can be semantically encoded, like it is the case in our study. However, we could argue that our results converge with this hypothesis, as we found an effect of interactivity over the semantic aspect of the task (linking pseudowords to meaning), but not the phonological or grammatical one (encoding of the wordforms and their linguistic structure).

Another closely linked domain-general phenomenon is the "guessing effect" (Potts & Shanks, 2014; Potts et al., 2019), which shows that asking participants to guess the meaning of a word before providing them with the right definition can lead to better learning. According to this account, the mechanism behind it is the curiosity generated by asking participants to provide a guess, which is exacerbated when they receive negative feedback (Potts et al., 2019). Contrary to the "generation effect" described before, this has been tested in the context of second language learning. Our results seem to match this hypothesis. However, when testing whether those participants in the interactive condition who made more matching mistakes in Block 1 obtain better matching scores in Block 4, we find the opposite effect: the matching accuracy in blocks 1 and 4 show a positive correlation for participants in the interactive condition.

Unfortunately, we did not collect any indicator of prediction generation from participants in the non-interactive condition which could help us test whether any of these accounts can explain our results. It could be the case that some participants (those who performed best) were generating guesses. A follow-up study could look at this further by including an interactive condition that did not provide feedback .It is also important to note that Experiments 1 and 3, in which the effect of interactivity was clearer, were conducted online, whereas Experiment 2 was conducted in person. Aside from the described methodological differences between these studies, it could have been the case that participants in the non-interactive condition in Experiment 2 were paying more attention and were more engaged than those in Experiments 1 and 3, in line with the findings in Finley and Penningroth (2015). This, together with the additional cognitive load that the interactive condition in Experiment 2 had, could explain the lack of differences between conditions in the initial blocks of the task. Even if we do not know if this was the case, with other studies suggesting very little to no difference in attention between lab-based and online studies

(Clifford & Jerit, 2014), it is important to explore this aspect further, particularly given the rise of virtual language learning platforms.

In any case, this dissociation between the total accuracy, phonological accuracy and grammatical accuracy raises interesting questions for psycholinguistics, as the effect of interactivity was only present in the aspect of language that most closely resembles that studied in other domain-general fields, such as memory (i.e., studies used by McCurdy et al., 2020). Could domain-general mechanisms be behind it, or could it be explained by the fact that the feedback participants received related to the mapping between semantics and wordform? Krishnan et al. (2018) ran a study in which participants learnt both new words for unfamiliar fish and semantic information about them, and manipulated the aspect of learning they received corrective feedback for (either accuracy of the label for the fish, accuracy of the semantic information, or none). Feedback focused on accuracy of the label showed a positive effect on the retention of these new wordforms, whereas feedback on semantic information did not. This suggests that, aside from the effect of generating responses, including feedback could be beneficial for some aspects of learning. However, their paradigm did not focus on the interactive aspect of learning but on explicit exposition of material, nor did it include grammar or explore the learning of wordforms. Future research could expand on this, testing how the effect of feedback focused on different aspects of language interacts with other potential factors (such as the generation effect).

The scope of the results of our study is limited by the homogeneity of the sample (university students) and the lack of potentially relevant information such as participants' linguistic background, as well as the specific conditions under which they completed the experiment, which was conducted online. Equally, although we showed that a small change in interactivity in our paradigm can have a lasting impact on learning, further research is needed to identify the concrete mechanisms involved in complex human interaction and how the type of feedback we identified interacts with other social and cognitive processes in real-life communication. The collection of additional measures, such as the time spent in each screen, eye-tracking, mouse-tracking, or even measures of brain activity through techniques such as EEG, with which prediction errors can be relatively easily shown, could help us understand what participants in different conditions were actually doing when completing the task, and shed light on the specific mechanism behind the effect.

We believe this new paradigm can combine the benefits of artificial language learning research with studies involving online games, opening avenues for new research. This study showed that rapid learning of linguistic input is possible through a referential communication task, and introduced measures sensitive enough to capture the effect of small methodological manipulations. Future research using this paradigm could address outstanding questions in sociolinguistics, for instance by exploring learning from different sources (virtual, human, virtual introduced as human or vice versa, different levels of authority, etc.). Furthermore, more complex artificial languages could be used to see how different linguistic components are encoded during interaction. Most studies looking at the cognitive factors that affect language acquisition and the evolution of the language itself rely on explicit instruction (Feher et al. 2016, 2019; Philip et al., 2013; Roseberry et al., 2014; Saldaña et al., 2019a; Smith et al., 2014, 2017), which is a departure from the conditions in which it often occurs (i.e. implicitly). Alternatively, they explored language acquisition in natural languages (Weisleder & Fernald, 2013; Kartushina et al., 2022), therefore missing the opportunity to look at language change beyond the confounds and barriers of existing languages. This paradigm bridges the gap between those two lines of research, and it can help to start detangling the complex interaction between the multiple factors involved in the effect of interaction over language learning. Although, of course, social interaction has many more elements to it than the ones we explore in this chapter, we believe that these paradigm and results can contribute to start understanding the multiple mechanisms involved in the effect of interaction in a granular manner. In summary, it allows us to explore language learning how it happens in most contexts: implicitly, through interaction and with the goal of communicative success, with the advantage of the control of variables and measurements that experimental studies bring.

Finally, these results and further research in this area have clear implications for language learning in online environments. In a highly digitalised world, it is vital to pinpoint and promote those factors that can make learning without real interaction more effective.

# Chapter 3: A novel experimental paradigm to study the effect of observation vs. interaction in language acquisition, use, and evolution

In Chapter 2, we explored the effect of simple interactivity over the acquisition of a novel language with a simple grammar. However, as stated in the discussion, the proposed paradigm does not allow for social interaction between two or more people to take place and is not sensitive to sociolinguistic variables. While the process of language acquisition in the studies in Chapter 2 is closer to natural conditions than the one we observe in classic paradigms using artificial language learning, and even closer to virtual language learning environments, it does not allow us to discern how some basic cognitive mechanisms, such as attention and memory, interact with social variables. Research in the fields of sociolinguistics, psycholinguistics, and evolutionary linguistics show how factors such as social status (Labov, 2006; Fedzechkina et al., 2022; Roberts & Fedzechkina, 2018), ingroup-outgroup biases (Fedzechkina & Roberts, 2020; Iacozza et al., 2019), size and characteristics of the community of learners (Garrod et al., 2010; Fay et al., 2010; Kirby et al., 2015; Raviv et al., 2021), or even beliefs about whether the interlocutor is a human or a machine (Feher et al., 2016) affect the use and evolution of languages.

The effect of interaction in language acquisition and change has been studied using a variety of methods using both natural languages, through observation (Newport, 1999; Singleton & Newport, 2004) or corpus analysis (see Barron & Sneider, 2009 for a review), modelling (Kirby et al., 2008; Perfors, 2012, 2016; Smith et al., 2017), languages or communication systems generated by participants (Galantucci, 2009; Fay et al., 2010; Roberts & Clark, 2020; Roberts & Fedzechkina, 2018), and artificial language learning paradigms (see Newport 2020 for a review).

As we raised in Chapter 2, a limitation common to all of these paradigms is that the process of language acquisition is either explicitly instructed (Fedzechkina et al., 2012, 2017, 2022; Feher et al., 2015; Kirby et al., 2015; Smith et al., 2017), entirely generated by participants (Galantucci, 2009; Fay et al., 2010; Roberts & Clark, 2020; Roberts & Fedzechkina, 2018), or simply observed (Newport, 1999; Singleton & Newport, 2004), which does not allow the systematic exploration of how interaction affects the process of acquisition.

Studies in Chapter 2 showed how it is possible for participants to acquire an artificial language without explicit instruction, and we looked at the effect of interactivity on acquisition. Building up on that, here we present a paradigm that takes this further by allowing participants to learn implicitly through real interaction with other participants. We hope this could further explore not only how interaction affects language acquisition, but directly observe how variability is treated in acquisition, and how it may lead to changes in the structure of the language. In other words, this extends the work presented by adding the possibility of human interaction at the stage of acquisition, instead of after having experience with the language, hence, allowing for the exploration of how different ways of language acquisition could affect language use.

As discussed in Chapter 2, literature shows that, even if infants as young as 18 months-old have shown their ability to acquire language through observation (Akhtar et al., 2001; Akhtar, 2005; Floor & Akhtar, 2006; Gampe et al., 2012), interaction boosts language learning (see Hiver et al., 2021 for a review). In line with that, in Chapter 2 we observed that even an added element of interactivity could boost people’s ability to learn the association between a word and its meaning. We hypothesised that the explanation behind this effect could be a difference in the perception of the structural characteristics of the language, or the mere act of generating a response. However, our design did not include interaction with another real participant and did not allow us to discern between these two alternative explanations or explore additional contributing factors that are only present in real social situations, such as joint attention (Tomasello & Todd, 1983; Pruden et al., 2006; Ataman-Devrim et al., 2023). The paradigm we developed and describe in these chapter addresses these limitations while allowing us to simultaneously observe and manipulate interactional structure in language acquisition and observe its effect on language use and language change.

This paradigm is a further adaptation of the director-matcher task (Clark & Wilkers-Gibbs, 1986), which addresses those limitations. By developing different modalities of three-way interactions in which one of the interaction partners is familiar with the language (that is, interactions formed by two participants and a confederate), we can explore how implicit learning happens, and we get even closer to a natural context (in which at least one of the interlocutors knows the language) while retaining experimental control. This study contains three conditions based on two different settings: in the first setting participants will learn by alternating between interacting with a confederate introduced as a participant who has been trained in the artificial language and by observing the other participant’s interactions with the confederate. Both participants in this condition will be in the "interact-and-observe" condition, and they will learn half of the language through interaction and the other half through observation. In the second setting, one of the participants will interact with the confederate ("interact-only" condition) while the other participant learns by observing their interaction ("observe-only" condition).

That way, we will be able to compare those participants who learn purely through observation, participants who learn purely through interaction, and participants that learn part of the language through observation and part of it through interaction. This would allow us to conduct both within participant and between participant comparisons and detangle the effect of learning through interaction vs. observation and the effect of participating in the interaction.

In addition, this experiment aims to explore how variability is treated in language acquisition. Language seldom contains unpredictable variation (Givón, 1985), that is, linguistic elements whose probability is not predicted by the context (grammatical, social, syntactic, etc.). Several studies in artificial language and natural languages such as creoles have shown that this tends to be eliminated either by reducing the variants to one (i.e., regularising) or by conditioning the variants to some other aspect of the language (Hudson Kam & Newport, 2005, 2009). Some of the explanations given for this phenomenon have been the additional cognitive load of encoding unpredictable variation (Aslin & Newport, 2012; Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005, 2009), a bias toward towards more efficient forms of communication (Fedzechkina, 2012, 2017, 2018, 2022), or participants' prior beliefs on the nature of variation and the goal of the tasks (Perfors, 2012, 2016). Interaction itself has shown to reduce unpredictable variation through the priming and alignment processes that occur when using a language in interaction (Feher et al., 2016, 2019) or the pressure imposed by communicating with various interlocutors (Raviv et al., 2020). Moreover, first language acquisition by children is considered to be one of the drivers of language change and the reduction of variation, according to observations in children who learn their first language from inconsistent input from their parents, such as deaf children whose parents use home sign (Goldin-Meadow, 2020). Following up on this, several artificial language studies have compared the treatment of unpredictable variation by children and adults (see Newport, 2020 for a review), but in all cases, all participants first receive direct instruction in the language and then are asked to use it. That makes it hard to observe how the language is treated during acquisition, in the learning stage, and disregards the impact that different forms of acquisition, aside from direct instruction, could have over the later use of the language. In this study, we include an element of unpredictable variation: there will be two different variants for one of the verbs, that is, the same concept will be expressed with two different words. These will be used with a different frequency (the majority variant 75% of the time and the minority variant the remaining 25% of the time), but they will be unpredictable: no contextual cue will affect which of the words is most likely to be used in a given utterance. Here, the confederate would use the artificial language as predetermined, modelling the unpredictable variation, and would not change their linguistic behaviour in line with that of participants. In conversations in natural contexts, if both interlocutors are equal in terms of social and linguistic status, they usually adapt different aspects of the language, such as prosody, lexicon, or syntactic structure, to match each other, a process known as alignment (Branigan et al., 2005). However, the degree of alignment is dependent on different factors affecting the linguistic status or social status of the interlocutors, with interlocutors of higher status triggering higher levels of alignment from those with lower status (Weatherholtz et al., 2014). Based on the accommodation hypothesis, given that the confederate would be presented as the one most familiar with the language, we would expect participants who are interacting to mimic the confederate's behaviour. However, based Schoot et al. (2014), which shows that not being involved in a communicative interaction reduces the strength of priming we predict that, when learning through observation, the priming effects should not be as strong as for the interaction participant, and thus, we would expect the observer participant to be more likely to regularise to the majority variant instead of matching the behaviour of the confederate. Paradigms until date did not allow to test these predictions, as they did not include two learners who are simultaneously learning from a confederate and from each other in a highly controlled setting.

Furthermore, by including the "interacting-and-observing" condition, in which participants acquire a different set of nouns and adjectives either through interaction or through observation, we can not only test the predictions described above between participants, but also within participants, by testing the use of the variable particle with different lexicon. This condition has arguably a higher cognitive load than the observer only condition. Hence, we would expect it to be more likely that participants regularised (Hudson Kam & Chang, 2009). However, as explained earlier, the interactive pressure should lead participants to prime each other, and accommodate their productions to match those of the confederate priming (Branigan et al., 2005; Feher et al., 2016; Weatherholtz et al., 2014).

This design allows us to test whether the use of variants is consistent across those linguistic elements learnt through interaction and those learnt by observation. Finally, both in the interaction and observation, and observation only condition, participants will be exposed both to the use of the language of the confederate and of their peer. Even if there is no interaction between peers, this paradigm will let us explore whether participants match the overall probability averaged across the confederate and their peer, the confederate's, the peer's or any weighted approximation  From a sociolinguistic perspective, we could expect participants to be more likely to align their productions with those of the confederate, based on the higher authority in terms of knowledge of the language (Weatherholtz et al., 2014), or with their peer, based on ingroup-outgroup bias (Fedzechkina & Roberts, 2020). For example, if there are two verbs to express the same idea, "addun" and "puttun", and the confederate uses "addun" in 75% of their interactions with Participant 1, and "puttun" the in the remaining 25%, but when interacting with Participant 2, they use "addun" 25% of the times and "puttun" 75% of the times, then Participant 1 could: a) use "addun" 75% of the times, imitating the confederate's behaviour towards them, b) use each of the options 50% of the times, imitating the confederate's behaviour across all interactions with them and with Participant 2, or c) use "addun" with whichever frequency Participant 2 uses it, imitating their behaviour rather than the confederate's, or d) any combination of the former. Any of these probabilities, across all participants in different conditions, could give us valuable information on how statistical information is processed in social contexts, and how priming effects, communicative pressures, and sociolinguistic variables interact with each other.

Finally, this paradigm gives as a valuable opportunity to explore how these processes evolve and change over learning. Studies manipulating the complexity of an artificial language, or comparing children to adults, have observed that the mastery of the language impacts how unpredictable variation is treated (Newport, 2020), and observations from second language learning show that language proficiency affects priming processes (Sinclair et al., 2019). By using the measures described in Chapter 2 (total accuracy, grammatical accuracy and phonological accuracy) in addition to traditional measures in artificial language learning that explore language use, such as regularisation and conditioning, we will be able to observe how the described processes change over time as learning progresses, allowing us to obtain data on each of these measures at different time points and compare participants both within and between-conditions.

In summary, the current study, aims to: a) compare learning by observation vs. by interaction, between- and within-participants, in terms of total accuracy, phonological accuracy, grammatical accuracy and speed, b) compare the treatment of unpredictable variation in the process of language acquisition through interaction vs. through observation, both between- and within-participants, c) explore alignment/ accommodation processes in the treatment of unpredictable variation in different learning settings, and d) explore the relationship between the treatment of unpredictable variation and learning outcomes.

Some studies have focused on pairwise interactions (Feher et al., 2016), or group dynamics (Fay et al.,2010), have directly compared interaction vs. observation (Anderson & Pembek, 2005), or compared learning from a teacher vs. from peers (Ibsen-Jensen et al., 2018), but to our knowledge, no other experimental paradigm in the psycholinguistic literature allows to systematically and simultaneously manipulate and explore language interaction and observation, and learning from different agents. Here we will describe the design for a study that compares learning and language use through interaction, observation, or both, between and within participants, as well as how this impacts language use. In the discussion, we will cover further manipulation that this paradigm allows, and how they could potentially be used to fill further theoretical gaps.

## Methods

### Description of the paradigm

This paradigm aims to test learning in different group structures. Two participants at a time are invited to each experimental session, in which they are introduced to each other and to the experimenter. They are told that they will be playing a videogame through which they will be learning an artificial language. In every case, there will be three players taking part in a cooperative game: one experimenter and two participants. To impede verbal and non-verbal communication, they will be separated by panels, and each of them will play the game on a tablet or computer. In order to communicate, players will be using exclusively an artificial language designed ad hoc for this game, in written format, through their device. The experimenter will be familiar will the artificial language and will produce perfect linguistic input. They can be introduced to the participant as a player who is familiar with the artificial language. Participants will need to learn the language through interaction.

In this game, the players will be building a tower together. They will take turns to add a series of bricks to the tower. There are 16 types of bricks, which differ in two parameters:

shape and filling pattern. As shown in Figure 3.1, bricks can be filled with four different patterns and have four different shapes. We designed the bricks to be visually easily distinctive.

Figure 3.1.

*The sixteen bricks forming the stimuli.*

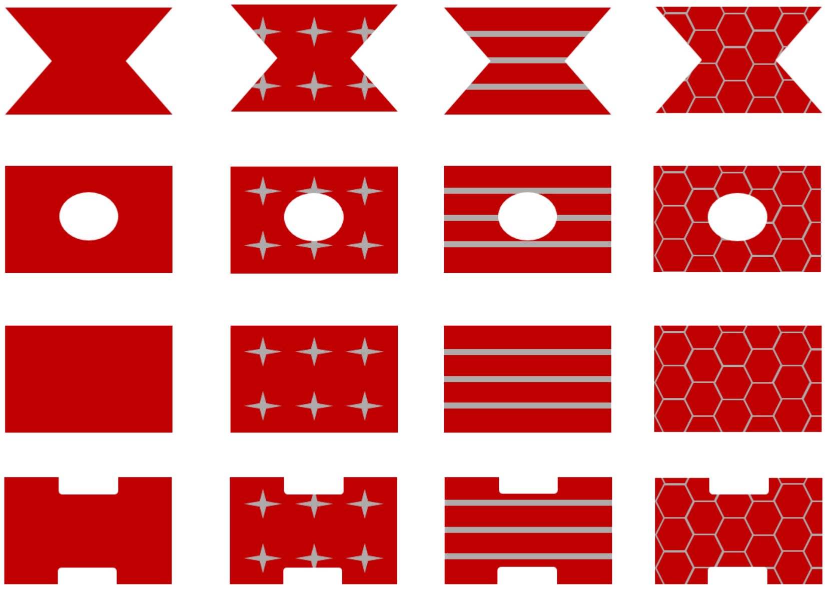
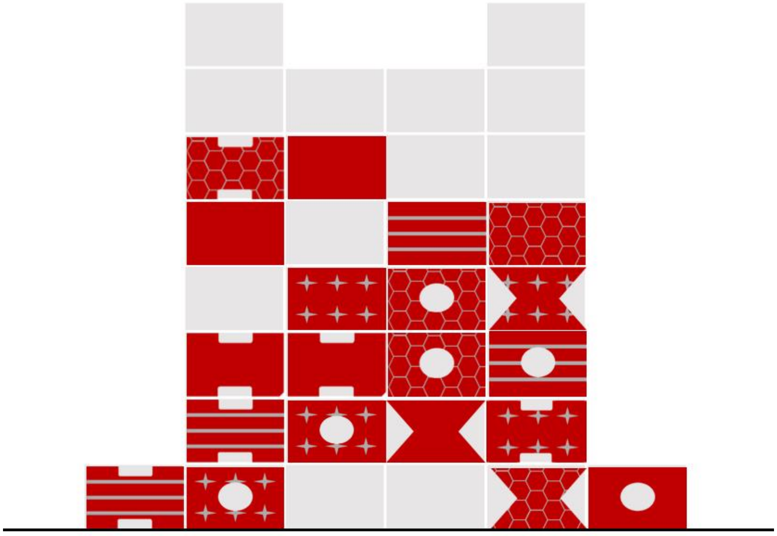


Figure 3.2

*An example of a tower in construction*

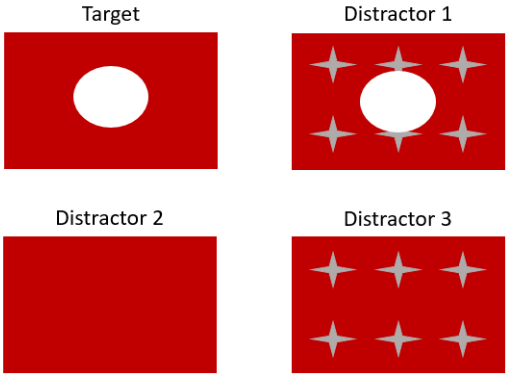


*Note.* The grey squares represent the bricks that were not added to the tower because communication was not successful.

Every tower will contain a maximum of 32 bricks (two of each type), displayed in a matrix of 6*8 (see Figure 3.2). To be able to build the tower, the players need to put the bricks in a specific sequence. This sequence is generated for each tower fully randomising the existing sixteen bricks twice. As stated before, this paradigm is an adaptation of the director-matcher paradigm (Clark & Wilkers-Gibbs, 1986), so in each trial one of the players will act as the director, instructing which brick to add next, one as a matcher, selecting the brick to include according to the director´s instructions, and one as an observer, who will not take part in the interaction but will be able to see what the instructions from the director were, which brick the matcher selected, and whether it was the correct one or not.

In a given trial, hence, the player acting as the director will be told by the program which brick comes next. Then, they will communicate this to the matcher using the artificial language, which contains different nouns and adjectives to describe the shapes and the fillings, respectively (see section Language Structure for the description of the language). The player acting as the matcher for a given trial will receive the utterance produced by the director and be asked to select the right brick from an array of four options: the target brick and three distractors. As for the Study in Chapter 2, one of the distractors will have the same filling as the target brick but a different shape, the second distractor will have the same shape as the target brick and a different filling, and the third distractor will have the same shape as the second distractor and the same shape as the second distractor (but none of the features in common with the target brick) (see Figure 3.3 for an example). This way, the player acting as matcher will be required to be familiar with the words describing both the shape and the fillings of the bricks to be able to give the correct answer. The position of the target brick and the three distractors will be randomised.

Figure 3.3.
*Target and distractors.*

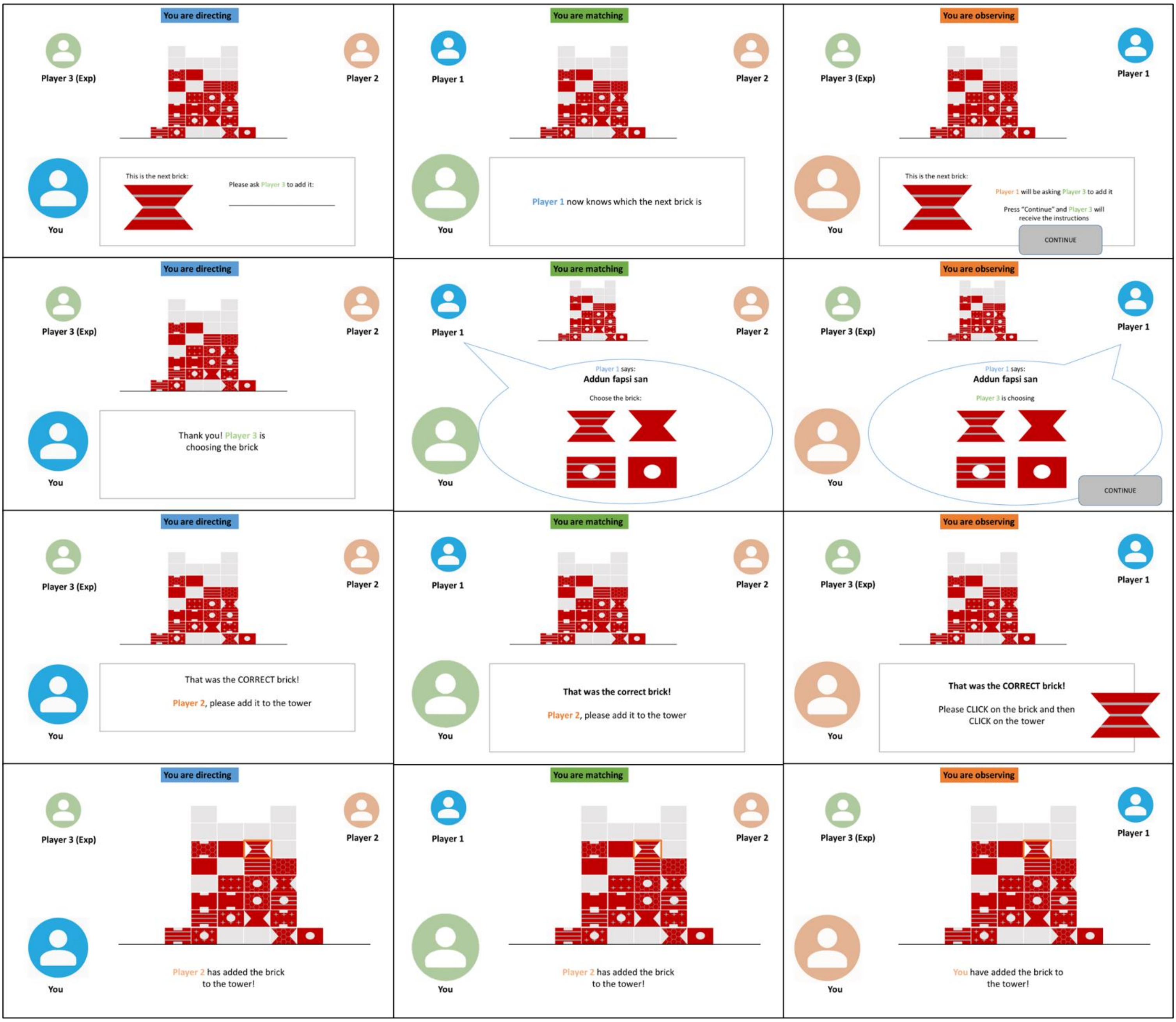


*Note.* Distractor 1 has the same shape as the Target brick, but a different filling, Distractor 2 has the same filling as the Target brick but a different shape, and Distractor 3 has the same shape as Distractor 2 and the same filling as Distractor 1, with no parameters in common with the Target brick.

If the communication is not successful (either because the description that the director provided was not correct or because the other participant did not pick the right brick), participants will receive feedback indicating that they did not manage to include the block in the tower and the space for the given block will remain empty, whereas if the communication is successful the brick will be added to the tower (see Figure 3.2). The towers will be built from left to right and from bottom to top. When the experimenter is acting as the matcher in a trial, and receives an utterance from the director, they will select the brick that most closely matches the director´s production, whether that is the correct option or not, as a native speaker of the artificial language would presumably do. The program will aid in this process by comparing the production that would perfectly describe each of the bricks in the selection array (the target and the three distractors) with the director´s production and indicating which one has the lower Levenshtein distance to the director´s production. If two or more bricks are equidistant to the production, the confederate will choose randomly. This will be calculated automatically, indicating to the experimenter which block to select, in order to avoid any individual biases or errors.

The participant with the role of observer, in order to keep them engaged and avoid social exclusion, will be the one pressing the button “Continue” to progress across the different stages of the trial, and will be asked to click the brick in the trial to add it (or not, if the trial is unsuccessful) to the tower. There will be a 500ms blank screen intertrial interval between trials. Figure 3.4 shows what an example trial for each of the roles: director, matcher, and observer.

Figure 3.4.
*A sample trial as shown in each participants' screen.*



This is the next brick:

Please ask Player 3 to add it:

Player 1 now knows which the next brick is

This is the next brick:

Player 1 will be asking Player 3 to add it

Press “Continue” and Player 3 will receive the instructions

Thank you! Player 3 is choosing the brick

Player 1 says:
Addun fapsi san

Choose the brick:

Player 1 says:
Addun fapsi san

Player 3 is choosing

That was the CORRECT brick!

Player 2, please add it to the tower

That was the correct brick!

Player 2, please add it to the tower

That was the CORRECT brick!

Please CLICK on the brick and then CLICK on the tower

Player 2 has added the brick to the tower!

Player 2 has added the brick to the tower!

You have added the brick to the tower!

*Note.* The left column represents the trial from the viewpoint of the director, the middle column from the viewpoint of the matcher and the right column from the viewpoint of the observer.

## Design

Participants will be randomly divided in pairs. Each pair will be assigned to either of two settings, which differ in two ways: with whom they interact and from whom they receive input. Participants can be placed in three different conditions within these settings depending on their roles: interact-and-observe, interact-only and observe-only.

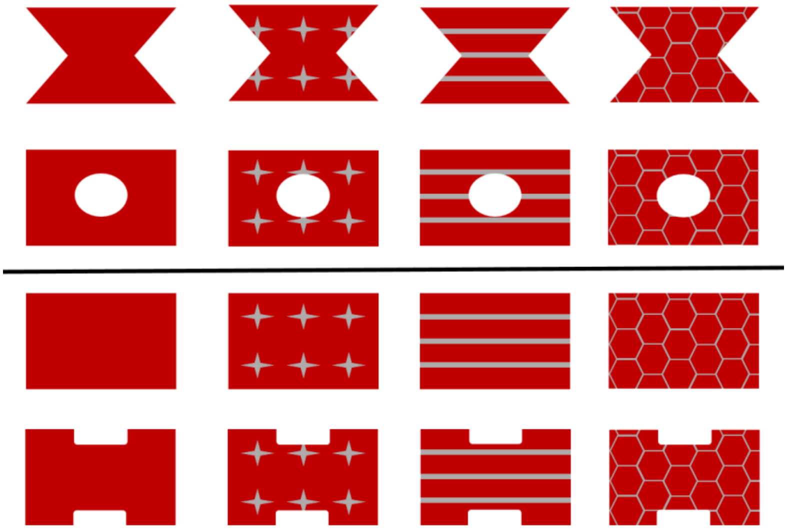
***Setting 1. Experimenter-learners***

In this setting, both participants will receive input from and interact with the experimenter. Participants will be able to see the interaction between the other participant and the experimenter. These two participants will both interact with the experimenter and observe interactions, so both will be in the interact-and-observe condition.

However, each of the participants will interact using half of the set of bricks. The set will be divided by shapes. Thus, each of the participants will be interacting with the experimenter using only eight bricks in two of the shapes. This division will apply both to the bricks that they describe for their interaction partner (director trials), as well as those presented as both the target and distractors when their interaction partner asks them to select a brick (matching trials). as well as the bricks that will be shown in the matching trials both target and distractors. Figure 3.5 shows an example of how the set could be divided.

Figure 3.5.

*Example of the division of bricks by shape.*



*Note.* In this example, within the “experimenter-learners” setting, Participant 1 would learn the description corresponding to top half of the bricks by interacting with the experimenter, whereas Participant 2 would learn them by observing these interactions. The opposite would be true for the bottom half of the bricks: Participant 2 would learn their description by interacting with the experimenter and Participant 1 by observing these interactions.

As described in the previous section, each tower contains 32 bricks, two of each type, and therefore, each block contains 32 trials. Each participant in this setting will be learning the nouns and adjectives corresponding to eight of the sixteen bricks through interaction and eight through observation. Within a block, they will be director for one of the trials in which each of the bricks appears and matcher for the other time it appears, and in consequence, they will observe each of the eight remaining blocks once being described by the experimenter and once by the other participant.

There are four types of trials participants in this setting will be taking part in: director-trials, matcher-trials, observer trials with the experimenter as the director, and observer trials with the experimenter as the matcher. We will divide the 32 trials in 8 sections, each of them containing one of the four types of trials, and randomise the order of these trials within the section.

To illustrate this with an example (Table 3.1), imagine that Participant 1 was learning bricks 1 to 8 by interacting and bricks 9 to 16 by observing, whereas this was the opposite case for participant 2. The order in which each of the bricks if presented would be randomised by section for each of the participants and roles. Then, the order of each of the trials would be randomised within each block. For example, the four trials of section 1 in the example below would go as follows:

1. Participant 1 would describe Brick 1 to the Experimenter, who would select it; Participant 2 would observe the interaction
2. Participant 2 would describe Brick 9 to the Experimenter, who would select it; Participant 1 would observe the interaction
3. The Experimenter would describe Brick 3 to Participant 1, who would select it; Participant 2 would observe the interaction
4. The Experimenter would describe Brick 14 to Participant 2, who would select it; Participant 1 would observe the interaction

Table 3.1.

*Illustration of trial organisation across a block.*

| **Sub-block** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
|---|---|---|---|---|---|---|---|---|
| **Participant 1- Director** | B1 | B3 | B4 | B2 | B5 | B6 | B8 | B7 |

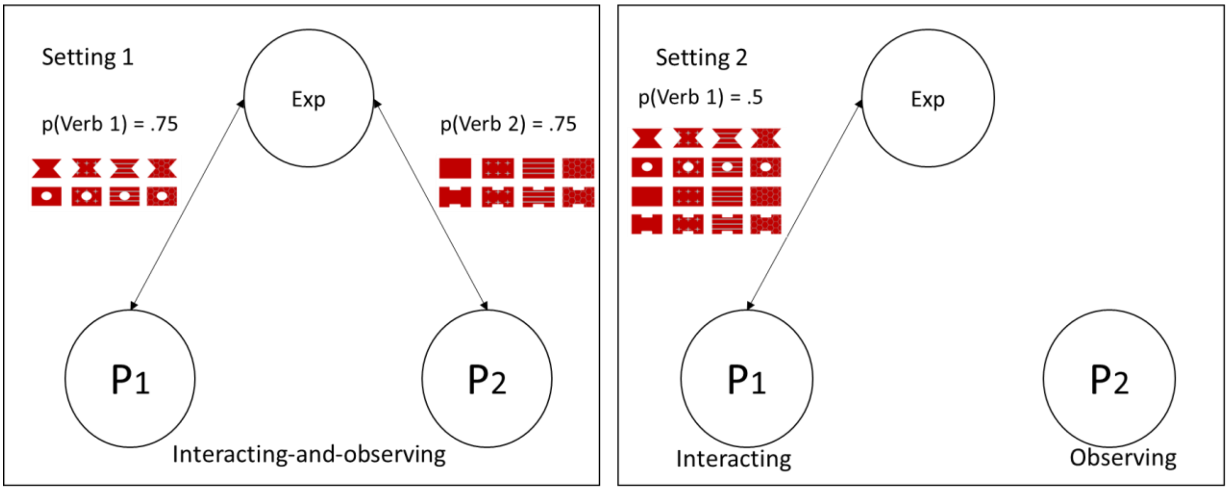
| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Participant 1- Matcher** | B3 | B4 | B1 | B5 | B8 | B7 | B2 | B6 |
| **Participant 2- Director** | B9 | B16 | B15 | B14 | B11 | B10 | B12 | B13 |
| **Participant 2- Matcher** | B14 | B11 | B13 | B9 | B10 | B12 | B15 | B16 |

*Note*. “B” is short for “Brick”. This table illustrates potential randomisation for the trials within each of the blocks. Each block would be divided into eight sub-blocks, each containing one trial of each type (Participant 1- Director, Participant 1 – Matcher, Participant 2 – Director, Participant 2 – Matcher). The order of presentation of these trials would be randomised within each of the sub-blocks. For example, the block in this example could start with Participant 1 acting as director with Brick 1, followed by Participant 2 acting as director for Brick 9, then Participant 2 acting as matcher for Brick 14, and Participant 1 acting as director for Brick 3, before progressing to sub-block 2.

### *Setting 2. Experimenter-learner-observer*

In this condition, Participant 1 will interact with the experimenter (interact-only condition). Participant 2 will be observing the interaction between Participant 1 and the experimenter without taking part on it (observe-only condition). In order to maintain the attention of Participant 2, we will ask them to type what Participant 1, or the confederate have said, or what option they have selected when matching, before their peers are provided with feedback. In half of the trials they will be asked to repeat the director’s production (half of the times when Participant 1 is directing and half of the time when the confederate is directing), whereas in the other half of the trials, they will be asked to repeat what the matcher has selected (half of the time when the matcher is Participant 1, and the other half when it’s the confederate who is the matcher). This will allow us to keep those participants who are observers in a given trial engaged in the task, and to have a proxy measure for attention in the observer trials. As in Setting 1, the block will be divided in eight sections of four trials each: two with the Participant 1 as the director and two as the matcher. The order of the four trials will be randomised within each section.

Figure 3.6.
*Experimental design*



**Sample considerations**

The sample will consist of English-speaking adults. Importantly, to obtain an equal number of participants for each condition, we would need to run twice as many experiments with Setting 2 as with Setting 1.

**Structure of the artificial language**

The artificial language, named "Babelian" (see Table 3.2), will be composed of ten words: two verbs, four nouns and four adjectives. The verbs are "puttun" and "addun" and they both mean "put" or "add". They are based on the English words so that participants find it easy to map them to their meaning. The nouns will refer to the shape of the bricks. The adjectives will refer to the filling. Table 1 shows the ten words, with their category and mapping. Both adjectives and nouns are pseudowords, which were obtained from the ARC non-word database (Rastle et al., 2002). They contain the same number of phonemes and letters and the same word structure within category so that they are easily identifiable as nouns or adjectives. In addition, nouns were added an "i" as the last letter for distinctiveness. Also, they do not have any letter or phoneme in common within category except for the ending vowel, so that they are easily distinguishable. All the pseudowords contain only legal bigrams in English and have a low number of phonological and orthographical neighbours, as well as a low summed frequency of orthographic (<100), body (<100) and phonological neighbours in English (<400). When forming a sentence, this is the permitted word order: Verb + Noun + Adjective. The mapping of shapes and fillings to nouns and adjectives will be randomised per participant pair.

Table 3.2.

*Vocabulary in the artificial language*

| Word in AL | Syntactic category | Mapping |
|---|---|---|
| puttun | Verb | Put/Add |
| addun | Verb | Put/Add |
| fapsi | Noun | Shape S1 |
| zecti | Noun | Shape S2 |
| gulbi | Noun | Shape S3 |
| jondi | Noun | Shape S4 |
| nal | Adjective | Filling F1 |
| zoc | Adjective | Filling F2 |
| dep | Adjective | Filling F3 |
| jud | Adjective | Filling F4 |

In Setting 1, the experimenter will use one of the verbs with a probability of .75 with Participant 1 (e.g. “addun” with a probability of .75 and “puttun” with a probability of .25), whereas the proportion of use will be inverted for Participant 2 (e.g. “addun” with a probability of .75, and “puttun” with a probability of .25). Within the setting, hence, participants will be exposed to each verb with a probability of .5. In Setting 2, the experimenter will use both verbs with a probability of .5 with Participant 1, while Participant 2 observes the interaction (see Figure 3.6). This way, all participants will be exposed to Verb 1 with a probability of .5, with the difference that in Setting 1 the proportion with which each of the verbs is used will differ by interaction partner. This allows to observe, within Setting 1, whether participants reproduce the overall frequency of use of Verb 1 in their productions, averaging that frequency observed through interaction and through observation, or whether their production is biased towards the frequency of use observed during the interaction trials or during observation trials.

In Setting 2, the probability of Verb 1 cannot be manipulated in the same way. However, this condition can serve as a baseline of whether participants, in general reproduce the frequency of verb use they observe at all when learning a language through interaction vs. through observation.

**Procedure**

Two participants and the experimenter will gather in a room and told that they will be playing an online interactive game on a tablet/computer. They will be given the information sheet and the experimenter will explain that the three of them will be playing an interactive game together, in which they cooperate to build a tower, but that they can only communicate by using “Babelian”, a language that they will be learning as they play. After clarifying any questions that they may have, they will be placed at individual tables separated by dividers and given a device each. Participants will then sign the consent form, and individually read the instruction that correspond with their condition. They will press “Proceed” when they are ready to start and sent to a waiting screen, until all players are ready to start. When both participants in a session have read the instructions and pressed “Proceed”, the experiment will commence. It will be divided in four blocks. Each block will consist of an interaction phase, in which they build a tower, and a testing phase, in which they are tested on their knowledge of the artificial language so far.

### *Interaction phase*

In the interaction phase, participants will build one tower per block, acting as directors and matchers once per each brick. However, the order of presentation of the bricks will be randomised. The experimenter will always be the first one to direct to one of the participants, who will direct back.

### *Testing phase*

The testing phase will consist of a production and a recognition test. In the production test, participants will be asked to produce a sentence asking the experimenter to put the next brick. They will need to produce the sentence for four of the bricks with no overlapping in shapes or fillings within them (thus covering the whole range of possible shapes and fillings). In the recognition test, participants will be given a sentence and asked to match which of the bricks it refers to from the set of 16 bricks. As in the production test, they will be asked to recognise only four of the bricks, different from those presented in the production test and with no overlap within them in terms of shape and filling. Participants will not receive any feedback on their performance during the testing phase. After participants have completed four rounds of the game, with their corresponding “Interaction” and “Testing” phases, they will be thanked for their participation and debriefed.

## Planned measures and comparisons

In order to measure learning outcomes, we propose using the measures described in Chapter 2: total accuracy, phonological accuracy, and grammatical accuracy. For the measure of the treatment of variability, we propose using the classical measures used in studies in unpredictable variation: entropy, which captures the preference of one variant over the alternative, and lexical mutual information, which captures the predictive value that each of the individual items have for the variant. Here, the variants are the two forms of the verb "addun" and "puttun", and the nouns (corresponding to the shapes) and the adjectives (corresponding to the fillings). We will calculate lexical mutual information independently for nouns and adjectives. Therefore, noun mutual information will represent the predictive value of nouns for a given verb variant, and the adjective mutual information will represent the predictive value of adjectives for a given verb variant. Both measures, entropy, and mutual information, are described in Samara et al. (2017).

In order to answer the first research question, which focuses on the impact of interaction vs. observation on learning, we plan to compare:

- Within Setting 1, learning outcomes in the part of the language learnt through interaction and the part of the language learnt through observation, within participants.
- Within Setting 2, learning outcomes in "interaction-only" condition against in "observer-only" condition, between participants.
- Between Settings, learning outcomes in "interaction-only" in Setting 2, against the part of the language learnt through interaction in Setting 1, between participants.
- Between Settings, learning outcomes in "observation-only" in Setting 2, against the part of the language learnt through interaction in Setting 1, between participants.

In order to answer the second research question on the impact of interaction and observation on the treatment of unpredictable variation, we will conduct the same comparisons with entropy and noun and adjective mutual information, instead of learning outcomes, as the outcome variables.

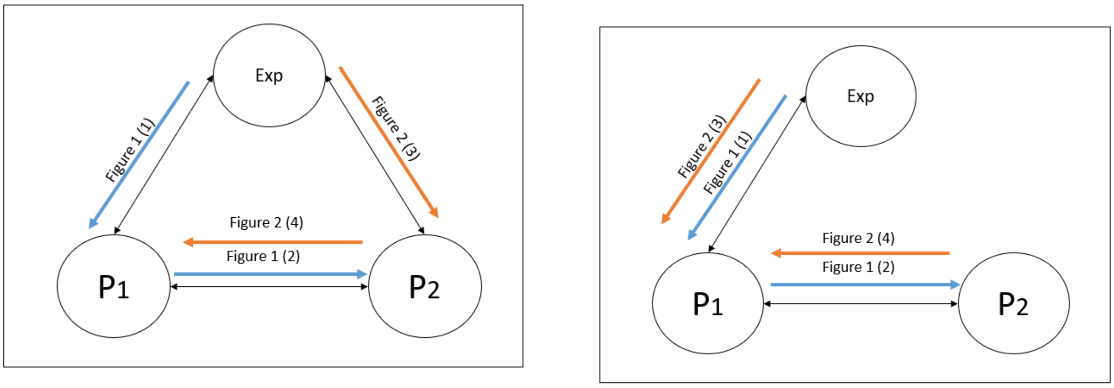
Third, in order to answer the next research question, which explores the impact of sociolinguistic variants (i.e. the status of the different agents in the interaction on unpredictable variation), we will compare, across settings, whether status (experimenter vs. peer) moderates the strength of the correlation between the entropy in verb variation observed in the language of the other players (the experimenter and the other participant) and the entropy in the verb variation in the language produced by the participant themselves. In other words, we will explore whether the verb variability in a given participants' production correlates with that of the other players, and to what extent.

Relatedly, we will explore local priming processes, by observing whether players productions can be predicted from those of their interlocutors within the "interact-and-observe" setting. In this setting, each participants productions will have been preceded by a production by either the experimenter or the other player, as part of an interaction they have observed, or they have taken part in. There will also be a small number of trials in which the preceding trial was their own production. We will compare these five prime types[2]. Finally, we will explore the correlations between entropy and mutual information and learning outcomes, within participants and within each of the learning blocks, and the moderation of the condition they were on.

## Additional settings

This paradigm allows manipulations at multiple levels. One of the most interesting elements is the further manipulation of the settings. We propose an extension of the previously described design with the two settings represented in Figure 3.7.

Figure 3.7. Additional experimental settings



*Note.* The panel in the left represent the structure of the "three-way-interaction" setting and the panel in the right represents the structure of the "hierarchical learning interaction" setting.

[2] This planned comparison will be dependent on power analysis and sample size.

***Setting 3. Three-way interaction condition*** (see Figure 3.7, left panel)

In this setting, participants interact with and receive input from both the other Participant and the experimenter. Participant 1 and Participant 2 are in the same condition in this setting. Participants will be able to see what the rest of the players have written and the blocks that they have selected. Adding this condition would allow comparison between those trials in which participants interact with each other and when they interact with the experimenter, and observation of those interactions too. As in Setting 1, this setting allows manipulation. of the part of the language participants learn from their peer and from the experimenter, and from observing vs. from interacting.

***Setting 4. Hierarchical learning interaction*** (see Figure 3.7, right panel).

In this condition, Participant 1 interacts with the experimenter and with Participant 2. Participant 2, in contrast, will only interact with Participant 1. Participant 2 will be able to observe the interactions between Participant 1 and the experimenter. This condition also permits the comparison of learning through interaction and through observation with both experimenters and peers, but in contrast to Setting 3, Participant 2 only interacts with Participant 1, and never directly with the experimenter.

As shown in Table 3.3, the combination of these four settings would cover all possible combinations of possible source of input through interaction (with the experimenter, the other participant, both, or none) or through observation (between the participant and the confederate or none)[3]. The combination of these would lead to six conditions: interact with experimenter only (Setting 1, Participant 1), observe-only (Setting 1, Participant 2), interact with experimenter and observe (both participants in Setting 2), interact with experimenter and participant, and observe (both participants in Setting 3), interact with experimenter and participant but not observing (Setting 4, Participant 1) and interact with participant and observe (Setting 4, Participant 2).

[3] The remaining two conditions that would arise from all combinations of the possible modalities of input through interaction and input through observation, namely interaction only with participant and no observation, and no interaction, and no observation, would not be feasible while maintaining the amount of input equal for both participants in the condition. If participants interacted with the other participant but did not observe any of the interactions, they would receive half as much input as their peer in that condition.

Table 3.3.

*Possible combinations of the manipulations on the origin of the input and their relationship with Settings.*

| **Input through interaction from** | **Input from observation from** | **Participants** |
|---|---|---|
| Experimenter | Experimenter – Participant interactions | Setting 2: Participants 1 and 2 |
| Experimenter | None | Setting 1: Participant 1 |
| Experimenter and participant | Experimenter – Participant interactions | Setting 3: Participants 1 and 2 |
| Experimenter and participant | None | Setting 4: Participant 1 |
| Participant | Experimenter – Participant interactions | Setting 4: Participant 2 |
| None | Experimenter – Participant interactions | Setting 1: Participant 2 |

## Additional manipulations

### Variability and complexity in the Artificial Language

Multiple studies have looked at the treatment of unpredictable variation in relation to Universal Grammar laws (Bickerton, 1984; Chomsky, 1984; Goldin-Meadow et al., 2008), and economic and communicative pressure (Fedzechkina et al., 2012, 2017; Gibson et al., 2019; Kirby et al., 2008, 2015). However, this has not been looked in relationship and in interaction with those pressures emerging from interaction, such as the processes of priming or sociolinguistic factors. This paradigm can be further expanded to answer those questions by manipulating the structure of the artificial language to learn adding variability (for example, in terms of word order), or manipulating its statistical properties, to explore the effect of competing forces, beyond the ones that existing paradigms allowed to explore (see Fedzechkina & Roberts, 2020).

### Social structure

There are known sociolinguistic biases that shape how the information that is received from an interlocutor is received and processed by the receptor, depending on social characteristics, such as social status, authority, perceived knowledge, or perceived ingroup-outgroup belonging (Iacozza et al., 2019; Weatherholtz et al., 2014). Studies with artificial languages have looked at these factors and how they affect reproduction of an input (Fedzechkina & Roberts, 2020), but never in an interactive context. This paradigm can easily be modified to tackle this question by manipulating the framing of the introduction of the confederate (as an experimenter or a fellow peer, the creator of the language or a more advanced learner, etc.).

### Communicative pressure

As mentioned earlier, one of the mechanisms proposed for the reduction of unpredictable variation in language learning is communicative pressure. This paradigm allows manipulating communicative pressure by introducing (or not) a reward for successful communication, that could be dependent either on individual behaviour, or in cooperation. This would allow brining the principles of game theory to the field of language acquisition and language change.

### Size of the community

Bringing together the field of experimental semiotics and evolutionary linguistics, we could manipulate the size of the community, generating more complex settings, as these have shown to influence the structure of the language (Fay et al., 2010; Raviv et al., 2021).

### Iterated learning

This design allows to generate chains of learners. In further generations, instead of using a confederate, a randomly selected participant from the previous generation can act as the confederate, or alternatively, a confederate whose input replicates that recorded for a participant in a previous generation, or a combination of several participants' output. These allows testing of a closer approximation of the conditions modelled in Smith et al. (2017), and directly observe the evolution of language with the introduction of new learners through interaction, rather than explicit training, making it closer to natural learning conditions.

### Meaning space

Meaning space refers to the number of unique items/parameters that a language can describe. In the case of this experiment, this is composed by the four shapes and the four fillings that the bricks could adopt, and the action of adding the bricks to the tower. No other meanings need to be communicated in the context of this experiment, such as the size of the bricks, or the action of removing them from the tower. The size of the meaning space can be easily manipulated by reducing or incrementing the number of parameters (adding bricks of different sizes or reducing the possible fillings to one), the number of variants within a parameter (adding additional brick shapes), allowing to test the effect of interaction and observation not only on learnt stimuli, but in novel stimuli. Raviv et al. (2019) showed that an increase of the meaning space could lead to the development of a linguistic structure, that is, when participants were presented with meanings they had not been trained on, they produced descriptions recombining the existing linguistic input that they had received, leading to more structured languages. However, this research did not include interaction, missing the effect that the pressured derived from it led to, as Smith et al. (2017) observed. For example, in the existing design, the correct description of shapes and filling is essential for communicative success. If the director does not refer to them in their description, the matcher cannot add the correct brick. Using this paradigm, we can increase the meaning space as in Raviv et al. (2019), adding for example bricks of different colours, and manipulate whether they are relevant or not for the success of the communication by accepting as correct (or not) bricks of any colour. According to Kirby et al. (2008), given the lack of pressure for expressivity, the increase in compositionality should not appear or should be reduced in this instance.

### Feedback

In their study, Krishnan et al. (2018) showed that evaluative feedback (whether the answer is right or wrong) can help the learning of new vocabulary, but not that of semantic facts. The interpretation of their results was based on cognitive and metacognitive aspects of learning. This paradigm allows as to manipulate the aspect in which the feedback focuses on, such as mapping, grammar, phonological accuracy, etc., in the context of real interaction, allowing us to understand how communicative pressure affect the effect of feedback in learning.

### Individual differences

Of course, all of this can be looked at and explored in relation to individual characteristics such as short-term memory skills (Husdon Kam & Newport, 2005, 2009; Ferdinand et al., 2019; Perfors, 2012) or learning abilities (Johnson et al., 2020).

## Implications and future research

We believe that this paradigm has the potential to bridge the gap between multiple disciplines with cognitive psychology, sociology, and evolutionary linguistics through its flexibility and its ability to simultaneously manipulate aspects of language and context that had not been looked at so far. All in all, it would expand the boundaries of the field in artificial language learning bringing it closer to the context in which first and second language acquisition occur. This would allow the testing of competing predictions from different research fields, which traditionally use different methods, and the exploration of opposing forces, such as those arising from the structural characteristics of the language (Hudson Kam & Newport, 2005, 2009; Hudson Kam & Chang, 2009; Perfors, 2012) and those coming from contextual demands (Fay et al., 2010; Garrod et al., 2010; Kirby et al., 2008, 2015;  Perfors, 2016;), in conjunction. In summary, it provides a framework for exploration and close observation of language during its learning, use, and over generations of learners.

Of course, as any other paradigm, ours is not free from limitations, such as the artificiality of the context, the limit in the complexity of the language that participants can learn in a limited amount of time in the laboratory, or the potential difficulties in adapting it to diverse samples, such as children or people with learning or sensory disabilities. In addition, we have limited the scope of the paradigm to written language and focused on English native speakers. Language is multimodal, including an oral aspect, as well as gesture. Expanding it to other modalities of language would require significant adjustments to the design and add higher level of complexity. However, most of these limitations are common to the field of experimental psycholinguistics, and we believe that, despite the mentioned limitations, and in combination with the results and predictions coming from other types of research, from natural languages to computer simulations, it can lead to an advancement of the knowledge on the field.

In addition, it can generate interesting insights for education, shedding light on the mechanisms involved in the acquisition of different aspects of language, and the best contexts for an easier learning, improving student outcomes. Aside from being able to closely manipulate and explore the effect of learning through interaction and observation, this paradigm and our proposed study can inform on the role of priming over the effect on interaction in language acquisition. According to the error-based language-learning model by Chang et al. (2006) priming can support long-term learning of syntactic structures. Using this paradigm, we can directly compare the effect of priming on language-learning and manipulating the effect of being part of the interaction. Previous studies (Feher et al., 2016, 2019; Muylle et al., 2021; Smith et al., 2017; Weber et al., 2018) have explored priming using artificial languages, but after directly instructing participants on the existing structures. This paradigm provides the opportunity to explore the effect of priming in second language acquisition from the start of the learning process, and directly contrast its effect depending on the social set-up (observing and interaction vs. being part of it) and the source of the prime sentence (the experimenter vs. a peer). Finally, it allows a granular analysis of different levels of language learning, such as phonological, grammatical, syntactic, and semantic ones.

Unfortunately, as stated in the COVID statement, even if the design of the study and the ethics application were finalised and the programming of the study was ongoing, data collection for this study was not possible during the duration of the doctoral project. However, we hope that the method that the careful design and its methodological advancements can help inspire future research.

# Chapter 4: Can illusion of causality lead to linguistic conditioning?

Natural languages exhibit variation at multiple levels, including phonetics, lexicon, syntax, and grammar, providing linguistic richness, and allowing for diverse ways of expressing ideas. However, this variation is not random (Givón, 1985): different aspects of the linguistic structure such as the grammatical rules, or the extra-linguistic context such as the level of formality (Labov, 2006), the drive for effective communication (Fedzechkina et al., 2012, 2017), and characteristics of the interlocutor (Weatherlotz et al., 2014), predict which of the possible variants the speaker is likely to use. A simple example of this is the use of contractions in English (e.g. "I'm" for "I am"). The contracted and the full form have the same meaning, yet their use depends on sociolinguistic factors (e.g., the contracted form would not be acceptable in formal texts), and syntactic factors (e.g., the contracted form would not be considered grammatical at the end of a clause, such as in "This is who I'm").

There are several factors that help explain the absence of random variation in natural languages. Usually, language acquisition happens in the context of a community where learners are acquiring the language from multiple sources simultaneously, which are difficult to track for researchers (Clark, 2009). Also, the input children are exposed to is usually grammatically correct (Singleton & Newport, 2004; Newport, 2020). There are contexts, however, in which the divergence between the received input and the output is bigger and the evolution of the language across different generations of speakers is easier to observe, that is, people produce a language that differs from that they were exposed to. This is the case of creolisation. Creole languages are born in contexts where two communities that speak different languages are in contact (Bickerton, 1981). Usually, a group of speakers needs to accommodate to the language of the dominant group, from which they adopt lexicon and functional words. This has often been the case in colonised territories, in which the enslaved people brought to the area generated a language based on the lexicon of the language of the enslavers, such as French creole in Haiti, Portuguese creole in Cape Verde, or English creoles in Jamaica or Papua New Guinea (DeGraff, 1999). The first result of the contact is often what is called "pidgin," a simplified version of the dominant language that includes features of the other languages. The grammar of pidgin languages has variables rules and often lacks structures to convey complex meanings. When this pidgin language is transmitted over generations, it often transforms into a creole language, which has clearly defined grammatical rules, lexicon, and native speakers (i.e., speakers whose first language is the creole language). It has been observed that children may be driving the creolisation process, by creating new consistent rules that give structure to the language, thus transforming the unstructured variable input from their parents (Bickerton, 1981, 1984; Kocab et al., 2016; Mufwene, 2007; Sankoff & Laberge, 1978; Sankoff, 1979; Senghas & Coppola, 2001).

A similar thing happens in the case of deaf children who are born in families that do not speak any sign language. Children use structures in a more consistent way than their hearing parents, creating consistent home sign languages (Singleton & Newport, 2004). This has also been observed in the development of Nicaraguan sign language (Senghas et al., 2004). This language started to develop in Nicaragua when the deaf community started to be more connected through the creation of a school for deaf children. Over new generations of speakers the language started acquiring structural elements such as compositionality of the signs. For example, when describing a motion, the older generations represented the shape of the motion (e.g., bouncing) and direction of the motion (e.g., down) in the same sign, and hence used an individual sign for each possible combination of direction and shape, whereas the younger generation produced separate signs for motion and shape, and combined them. Interestingly, the change seems to be driven by the younger speakers of the language, those who have started the acquisition when they were younger and transmitted up to the older speakers.

Similar results have been obtained from artificial language learning studies. These studies use very reduced experimenter-designed languages, ad hoc for every study. This permits one to observe language acquisition, use and transmission processes in a controlled context and with relative independence from the speakers' linguistic knowledge. When presenting children and adults with an artificial language that contained unpredictable variation, that is, the alternation of two or more possible forms independent from any lexical, grammatical, or social cues, children were more likely to regularise than adults (Hudson Kam & Newport, 2005). Regularisation in these studies is defined as using one of the alternative forms with a higher probability than that present in the input.

The reason for this difference between children and adults is not clear. According to the Newport's Less is More hypothesis (Hudson Kam & Newport, 2005; Newport, 1988), children simplify the language during acquisition to facilitate the learning process. This would explain why, when presented with unpredictable variation they choose a favourite form and use it more frequently than the alternative(s). If the difference between children and adults was due to the higher processing limitations in children, adults would show the same behaviour when presented with a more complex input (Ferdinand et al., 2019; Hudson Kam & Newport, 2009). This is what happened in a study by Hudson Kam and Newport (2009). Adults were presented with a language that contained multiple markers that were used unpredictably, Instead of alternating the different markers matching the probabilities they found in the input for each of the markers, they overproduced the most common marker, that is, they regularised. However, if they were presented with the same number of markers but there were clear rules about when to use which, varying predictably according to the noun, adult participants learnt these rules and replicated them in their output, instead of regularising, as when the variation was unpredictable.

Further research into regularisation in the face of unpredictable variation in adults has shown that the effect might be led by the processing load of language production (Hudson Kam, 2019; Hudson Kam & Chang, 2009), by a previous bias for regularisation (Perfors, 2012) or by pragmatic assumptions about the aim of the task and the reason behind variation (Perfors, 2016).

Other studies on unpredictable variation have shown that regularisation happens quicker when participants use the artificial language in interactions (Feher, Ritt, & Smith, 2019). Participants prime each other to use one or the other alternatives, a process known as alignment (Branigan et al., 2005). In those cases, in which one of the participants in the interaction pair uses exclusively one of the forms, the partner accommodates to the more regular speaker, leading to an even quicker regularisation.

Regularisation is not the only way in which participants reduce variation. In some cases, participants condition the variability (Hudson Kam & Newport, 2005, 2009; Hudson Kam, 2019; Smith et al., 2017; Feher et al., 2016; Perfors, 2016; Wonnacott, 2011). They use the alternative forms depending on other linguistic elements such as sentence structure, a co-occurring lexical item, etc., by imposing rules that are not present in the input. However, the conditions under which this conditioning happens are not well known yet. In language evolution research, iterated learning studies show how conditioning can appear as a result of the accumulation of individual learning biases (Smith & Wonnacott, 2010), but the process slows down when the task involves learning from multiple sources (Smith et al., 2017).

Communicative pressure seems to play a role in linguistic conditioning. For example, Fedzechkina, Jaeger and Newport (2012) and Fedzechkina, Newport, and Jaeger (2017) found that, when presented with unpredictable particles, participants conditioned them on the word order to favour communicative efficiency. Nevertheless, some participants in studies in which there is no communication, also show this behaviour (e.g., Hudson Kam & Newport, 2008). Researchers tried to explain conditioning in this case through the Less is More Hypothesis (Newport, 1988): participants would have conditioned the variation to simplify the rules and remember them better. However, it is not possible to know whether participants used conditioning as a strategy, or they did perceive that the variation was predictable and conditioned. Also, there is a significant amount of individual variability in these behaviours, and it is hard to predict which behaviour participants will show, with some of the participants reproducing the probabilities in their linguistic input, while other either get rid of one of the alternatives or condition it (e.g., Fehér et al., 2016). Research from domain-general cognitive processes could help us answer this question.

For instance, statistical learning has been shown to be one of the mechanisms in language acquisition. Humans are very sensitive to statistical regularities that have been shown to aid processes like word and sentence parsing (Saffran et al., 1996; Aslin et al., 1998) or the detection of grammatical rules (Gerken, 2005; Gomez & Gerken, 1999; Reeder et al., 2013; Wonnacott et al., 2008; see Romberg & Saffran, 2010, for a review). In addition, previous research has shown that some of the principles of general associative learning can be applied to the linguistic domain (see Ferdinand et al., 2019; Ramscar et al., 2013; Saldaña et al., 2019b). However, oftentimes in the linguistic field, statistical learning is interpreted as an accurate and unbiased process, in contrast with the findings in domain-general research.

Literature on statistical learning shows that the human brain is not only very proficient at perceiving statistical regularities, but also biased to perceive them even in their absence. An example of this is *illusion of causality* (Matute et al., 2015 for a review), defined as perceiving a relationship of causality between a cue (i.e., potential cause) and an outcome that are not contingent on each other, that is, when the presence or the absence of a cue does not predict the presence or the absence of the outcome. The bias to perceive illusory statistical patterns has been said to serve an evolutionary function, as the risk for survival of missing an existent pattern (a real danger or opportunity) is higher than the risk of perceiving a pattern that is absent (Haselton & Buss, 2000; Haselton & Nettle, 2006; Blanco, 2017).

Illusion of causality has been widely studied in the context of causal learning. The classic paradigm presents participants with a series of a trials in which they can see whether a cue/set of cues is present or absent and whether the outcome/outcomes occur (Allan & Jenkins, 1983; Jenkins & Ward, 1965). After being presented with some trials, participants are usually asked to judge the predictive power of the cues with respect to the outcomes. A typical example of this paradigm is the allergy prediction task (Van Hamme & Wasserman, 1994), in which participants are shown whether a person had contact with a potential allergen (the cue) and whether they developed an allergy (the outcome). These studies have helped uncover the conditions in which illusion of causality develops.

The statistical distribution of the cue and the outcome has been shown to affect participants' causal perceptions. When the number of times a cue and an outcome are paired together increases, so does the illusion of causality, even if they are not contingent on each other. This can happen when the base probability of the outcome is high (e.g., the number of people who develop allergies is high, regardless of the presence of an allergen), which is called *outcome density effect* (Allan & Jenkins, 1980; Alloy & Abrahamson, 1979; Msetfi et al., 2005). Equally, when the base probability of the cue is high (e.g., the allergen is present very often, regardless of the latter development of an allergy), the pairings between cue and outcome increase, which is called *cue density effect* (Allan & Jenkins, 1983). Finally, those situations in which both the base probability of the cue and the outcome are high are the ones which the causal perception estimates are the highest (Blanco et al., 2013). This effect further increases when participants have a prior belief that a causal relationship exists (Blanco et al., 2018).

In parallel, research in social psychology has long explored a phenomenon called *illusory correlation*. Illusory correlation can be defined as the perception of a statistical relationship between two unrelated events, and it is more likely to occur when both are high in frequency. It has mostly been studied in the context of stereotype formation, as this is believed to be one of its drivers (Fiedler, 1991; Hamilton & Gifford, 1976; Smith & Alpert, 2007; Van Dessel et al., 2021). In the classical paradigms, participants are presented with descriptions of individual behaviours/traits of people belonging to two groups of different sizes. The prevalence of one of the behaviours/traits is higher than for the other, but both are equally prevalent for both groups (Hamilton & Gifford, 1976). After the presentation, they are normally asked to estimate the frequency of the traits/behaviours for each of the groups and/or the likeability of the groups. In their original study, Hamilton and Gifford (1976) found that participants tended to report the majority trait to be more common in the majority group, and the minority trait to be more common in the minority group, even if the majority trait was equally frequent in both groups. They argued that, given that positive behaviours are overall more frequent than negative behaviours, people tend to associate infrequent negative behaviours with minority groups. However, this phenomenon has also been explored beyond social psychology, using for example, letters and shapes (Fiedler & Armbuster, 1994).

Although the two lines of research have remained relatively independent, the predictions and findings in the field of Illusory Correlation align with those in the field of Illusion of Causality: they both predict that a high frequency cue is likely to be paired with a high frequency outcome, and they both show, from different paradigms and theoretical accounts, that biases in statistical learning can lead to illusory associations. The conditioning behaviour that we observe when there is unpredictable variation, in which participants associated the variation to a cue, could be read as an example of *illusion of causality.* Participants, in trying to predict an unpredictable variant, generate associative rules, conditioning the variation to a linguistic cue. If these two processes (linguistic conditioning and causal illusion) are guided by the same underlying statistical learning principles, we can predict linguistic conditioning behaviour to be more prevalent under the same circumstances in which *illusion of causality* appear. Saldaña et al. (2019b) tested the predictions from the illusory correlation account alone in the linguistic field, using both semantic (animacy category) and social cues (gender of the speaker). Contrary to their predictions, they did not find skewness in any of the cues to predict a higher level of conditioning. However, the measures that they used for conditioning (Mutual Information, see Samara et al., 2017) contained an important flaw that would not have allowed these differences to be perceived: its range varied depending on the frequency of the cue, making their conditions not comparable.

This study built up on an unpublished study by Saldaña et al.'s (2019b), correcting the flaw identified in the measure, and testing in addition to the predictions from the *illusory correlation* literature, those from the *illusion of causality* literature. Whilst *illusory correlation* literature exclusively focuses on those situations in which both cue and outcome are high in frequency (e.g., when there is a majority group and a high frequency trait), *illusion of causality* literature finds the phenomenon also to be present when only either the cue or the outcome are high in frequency (cue- and outcome- density effects; e.g., when there is a majority group and a 50%-50% distribution of good and bad traits). Precisely, we aim to test whether the outcome-density and cue-density effects (or illusory correlation effect) replicate in the linguistic domain. We used a variation of the classic illusory correlation paradigm, adapting it to the linguistic domain. With that aim, we designed a language that contained nouns from two semantic categories (animate and inanimate), and two different particles to mark plurality. In this experiment, the cue was the category of the noun (an animal or an object) and the outcome was which plural marker it was associated with. As in illusory correlation paradigms (and in the classic allergy task; Van Hamme & Wasserman, 1994), participants were presented with different sentences one by one, and then they were asked to judge the relationship between the semantic categories and the markers. As it was a linguistic task, we also asked them to produce some sentences in the language in order to observe whether their productions reflected the input they received or they had regularised or conditioned. Finally, as in the classical Allergy Task, we asked participants to estimate the frequency with which each of the markers appeared with each of the categories.

To explore the effects of outcome-density and cue-density in the linguistic domain, we used a 2*2 design, manipulating the language on which we trained participants on. The input language varied in the skewness of a linguistic cue, in this case semantic category, and the skewness of the outcome, in this case, the frequency of two alternative plural markers.

In the uniform marker distribution conditions, each of the markers was used with the same frequency, whereas in the skewed marker distribution condition one of the markers was used more frequently than the alternative. Similarly in the uniform category distribution conditions, there was an equal number of items belonging to each animacy category, whereas in the skewed category distribution condition, the majority of the nouns were of one of the categories, and the remaining to the other (see Table 4.2 in the *Design* section).

Following from the literature on illusion of causality and illusory correlation, we predicted that:

1) Participants who were exposed to a language in which the frequency distribution of the marker was skewed would be more likely to perceive a relationship between category and marker, and hence, condition them on each other more than in conditions in which the frequency distribution of the marker was uniform.
2) Participants who were exposed to a language in which the frequency distribution of the animacy category was skewed would be more likely to perceive a relationship between category and marker, and condition them on each other more than in those conditions in which the frequency distribution of the animacy category was uniform.
3) Skewness in the distribution of the marker and animacy category would interact, leading to a level of conditioning higher than the sum of the individual effects.

4) When both the marker distribution and the animacy distribution are skewed, participants will condition the more frequent (majority) marker on the majority animacy category.

Our study allowed us to explore how domain-general mechanisms affect the perception of statistical regularities in linguistic input. If participants' perception of the probability of variable linguistic elements was biased, then that could lead to a biased production, which, when transmitted, due to an amplification of existing biases, could lead to the generation of novel grammatical rules (as in Reali & Griffiths, 2009, Smith & Wonnacott, 2010 or Smith et al., 2017). This could shed light on the process of implicit grammar acquisition and the conditions under which unpredictable variation is reduced, and ultimately, on the processes of language change and evolution.

## Methods

This study was approved by the ethics committee of the Psychology Department of the University of Warwick on the 30th of July2021 (reference PGR_20-21/15). The methods and analysis were preregistered and can be found at https://osf.io/5bjkd.

### Participants

We collected data until we reached a sample size of at least 40 participants per condition. That rendered a final sample of 320 participants after excluding 64 participants (see exclusion criteria below). We collected our sample through Prolific Academic, between the 2nd of August 2021 and 16th June 2022, and compensated participants with £3.63, equivalent to the prorated UK minimum wage for 25 minutes of participation at the time of ethics application in 2020. The inclusion criteria we established within the Prolific platform were: 1) being over 18, 2) being a native English speaker, 3) residing in the UK at the time of participation, 4) not having declared any language-related disorders or hearing difficulties. We also established certain exclusion criteria based on the data participants produced, excluding those who had not learnt the nouns in the artificial language and had produced a high number of responses which were not usable. Following pre-determined learning and performance thresholds, we excluded 1) 34 participants for having to discard more than 25% of their trials in the plural production phase (either because they failed to produce the right lexical item, a valid marker, or both), 2) 31 participants for failing to produce the correct noun in more than 25% of the filler trials in the plural production phase, 3) 28 participants for failing to produce the correct noun for more than 25% of the trials in the production task of the noun testing phase, 4) four participants for failing to select the correct answer in more than 25% of the filler trials in the noun comprehension phase, and 5) six participant for clicking the same side of the screen for more than 87.5% of the trials in the comprehension task (in 30 out of 32 trials)[4]. Some of these participants failed several of the exclusion criteria. The sample distribution per condition before and after exclusion can be seen in Table 4.1. The exclusion rate was relatively high (16.67%) presumably because of the difficulty in learning the stimuli and the online nature of the study. It varied between conditions (ranging from 9% to 23%) but it was not predicted by marker condition, animacy condition or task order.

Table 4.1.

*Number of participants per condition before and after exclusions*

| Marker condition | Animacy condition | Order condition | N before exclusion | N after exclusion | Exclusion rate (p) |
|---|---|---|---|---|---|
| skewed | skewed | Comprehension first | 45 | 40 | 0.11 |
| skewed | skewed | Production first | 49 | 40 | 0.18 |
| skewed | uniform | Comprehension first | 50 | 40 | 0.20 |
| skewed | uniform | Production first | 45 | 40 | 0.11 |
| uniform | skewed | Comprehension first | 48 | 40 | 0.17 |
| uniform | skewed | Production first | 51 | 40 | 0.22 |
| uniform | uniform | Comprehension first | 44 | 40 | 0.09 |
| uniform | uniform | Production first | 52 | 40 | 0.23 |

Regarding the sample demographics, 180 of our participants identified as female, 136 as male, 3 as non-binary, and one participant chose not to share their gender. The average age was 36.68 (SD = 13.19, range 18-75). In our sample, 60% only spoke English (192 participants), 25.93% declared to speak two languages (83 participants), and 14.06% of our participants (45 participants) spoke three or more languages.

**Development of materials**

To test the materials and measures, we conducted two pilot studies. Here we briefly describe the pilots and the changes in methodology they led to.

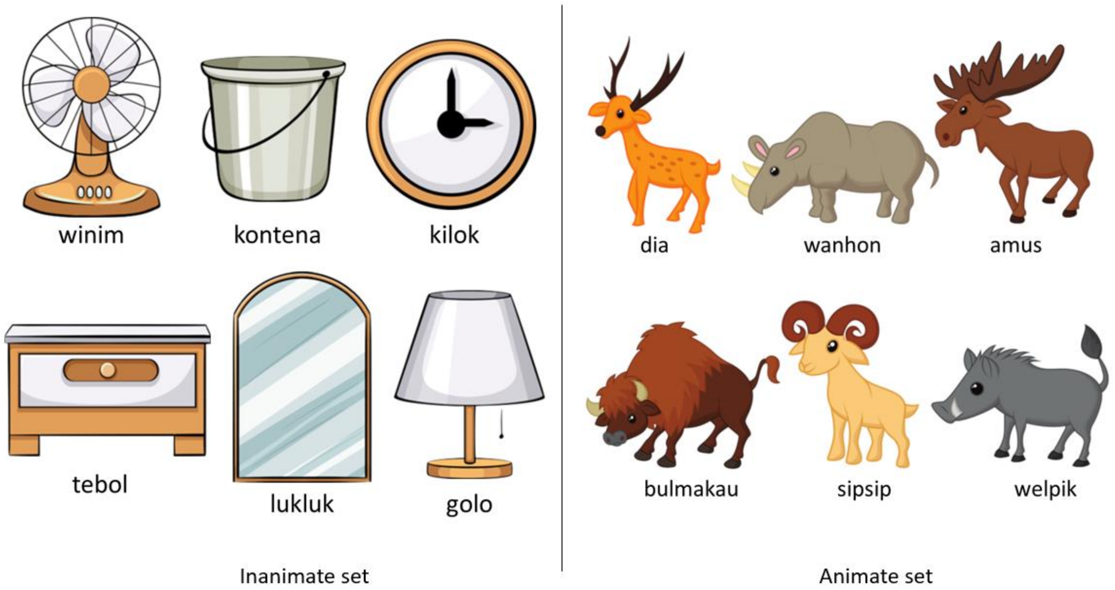
[4] This cut-off point was decided upon inspection of the screen clicking patterns, with most participants alternating between both side and a small subset clicking the same side consistently across the task, except for in up to one or two trials. This behaviour indicated poor attention to the task, and hence we discarded those.

In Pilot 1, with 120 first-year psychology students, we used an artificial language, called Panitok, from Saldaña et al. (2019b) which was comprised of 12 nouns and three plural markers. The nouns were based on the language Tok Pisin (an English-based creole from Papua New Guinea), so that it was easy for participants to learn them. Six of the nouns referred to animals ("sipsip", "bulmakau", "welpik", "dia", "wanhon", and "amus", meaning goat, bull, wild pig, deer, rhinoceros, and moose, respectively) and six to objects ("tebol", "golo", "kilok", "winim", "lukluk", and "kontena", meaning table, lamp, clock, fan, mirror, and bucket respectively). The three plural markers were "hap", "nim", and "tog".

Participants were presented with two types of sentences: singular sentences, only containing the noun (i.e., "table" in Panitok would be represented as "tebol"), and plural sentences, with a MARKER + NOUN structure (i.e., "tables" in Panitok would be represented as "tog tebol"). Each of the referents was represented with an image picture obtained from the following image database: http://123rf.com (see Figure 4.1).

Figure 4.1.

*Set of images for the animate (left panel) and inanimate (right panel) referents with their names in Panitok.*



The results of Pilot 1 revealed some issues with the materials:

1) When asked about whether they had perceived the items to belong to different categories, only 37 out of 80 participants (46.25%) identified animacy as a category.
2) Some of the artificial words were not easily learnt by participants, and were accurately produced under 80% of the times, whereas others were produced correctly over 99% of the time. The words that were correctly produced in less than 80% of the trials were “bulmakau”, “wanhon”, and “welpik”.
3) “Welpik” and “Wanhon” were often interchanged, leading to errors.
4) Due to a phonetic similarity, participants were highly biased to choose “nim” with the word “winim”.

To address these issues, we made the following changes:

1) In line with the results by Culbertson et al. (2019) which showed that category salience was an important factor on the acquisition of conditioning, we made the categories more salient by increasing the within category homogeneity. Pilot 1 included household items with different uses, sizes, and frequency. We swapped all household items for vehicles (taxi, moped, ambulance, bus, tractor, and digger, as “taksi”, “moto”, “karsik”, “bas”, “trakta” and “diga”, respectively), to make this difference between animate and inanimate objects more salient. The vehicles were also visually different as they had been designed by a different artist to the one producing the images for animals.
2) We substituted “bulmakau” (bull), for the easier “bulkau”, and “welpik” (wild boar), for “pumba”. Also, we removed “wanhon” (rhinoceros) and substituted for an elephant which was assigned the pseudoword “tronki”.
3) As the overall transparency of the pseudowords had increased for all items except for “sipsip” (goat), we changed its name to “bili”.
4) To make sure that the difference between categories was perceived and given the importance of early phonological cues (Culbertson et al., 2019). we added a prefix that differed by category (“te-“for vehicles, and “da-“ for animals).
5) The original markers were “tog”, “nim” and “hap”. Given that with the changes that we had made, all inanimate objects would start with the phoneme /t/ and that Pilot 1 showed that phonetic similarity between the marker and an item could lead to a bias, we changed the marker “tog” to “bok”.

Pilot 2 (60 participants) trialled the new stimuli and showed that the accuracy for all pseudowords was of over 90%, and that the proportion of participants who declared having perceived an animate and inanimate category had increased to 45 out of 60 participants (75%). Equally, we did not find any obvious biases in the pairings between individual markers and pseudowords.

**Language structure and stimuli**

The final artificial language, which we continued calling Panitok, was comprised of 12 nouns and three plural markers. Six of the nouns referred to animals, “dabili” (/dabɪlɪ/), “dabulkau” (/dabʊlkaʊ/), “datronki” (/datrɒnkɪ/), “dadia” (/dadɪa/), “dapumba”(/dapʊmba/), “damus” (/damʊs/), meaning goat, bull, elephant, deer, wild boar, and moose, respectively, and six to vehicles “tetaksi” (/tetaksɪ/), “tediga” (/tedɪga/), “tekarsik” (/tekarsɪk/), “tebas”(/tebas/), “tetrakta” (/tetrakta/) and “temoto” (/temɒtɒ/), meaning taxi, digger, ambulance, bus, tractor, and motorcycle respectively). The animals started with the prefix “da-”, and the vehicles with the prefix “te- “, to facilitate the recognition of the categories. The three plural markers were “hap” (/hap/), “nim” (/nɪm/), and “bok” (/bɒk/).

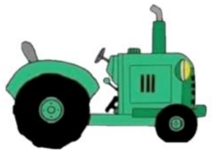
Each participant was presented with a subset of eight nouns and two markers. The proportion of animals (animate referents) and vehicles (inanimate referents) depended on the condition they had been assigned to (see the Design subsection). Two markers were randomly selected from the set for each participant.

Participants were presented with two types of sentences: singular sentences, only containing the noun (i.e., “taxi” in Panitok would be represented as “tetaksi”), and plural sentences, with a MARKER + NOUN structure (i.e., “taxis” in Panitok would be represented as “bok tetaksi”).

Also, to facilitate learning, as the phonetic representations of the nouns is Panitok were closer to their English counterparts than their written representations, each sentence was presented together with its audio recording. We used the grapheme-to-phoneme translation rules of Tok Pisin to record the stimuli, based on Smith (2008). The full set of images and their associated nouns is shown in Figure 4.2.

Figure 4.2.

*Set of images for the animate (upper panel) and inanimate (lower panel) referents with their names in the new version of Panitok.*



## Design

We used a between participants 2*2 design, with skewness in the frequency linguistic marker (marker condition: uniform or skewed) and skewness of the frequency of animacy category (animacy condition: uniform or skewed) as independent variables. Participants in the uniform marker conditions were presented with a language in which both markers were used with the same frequency, both across the language and with each of the individual nouns. Participants in the skewed marker condition were presented with a language in which one of the markers was used 75% of the time and the other 25% of the time, both across the language and with each of the individual nouns.

In relation to animacy condition, participants in the uniform animacy condition learnt a language that contained four nouns of each category, whereas participants in the skewed animacy condition learnt a language that contained six nouns of one animacy category and two nouns of the remaining animacy category (e.g., six animals and two vehicles). Table 4.2

presents the specific number of trials of each type that participants in each condition would have been presented with within a block of 32 trials. We selected the specific number of trials based on the skewness that we wanted to achieve in each of the parameters. As it can be inferred, there was no relationship between plural marker and category in any of the conditions.

However, based on previous results in the non-linguistic domain, we expected those participants in the skewed conditions to develop illusion of causality and show conditioning behaviour. We used two different tasks (comprehension and production) to obtain our outcome measures, and their order was counterbalanced within each of the conditions (task order: comprehension task first vs. production task first). Task order was not a variable of interest, and we did not hold any predictions in its regard but acknowledging that it could interact with our fixed effects in unexpected ways, we included it as a fixed effect in our models (see Statistical analysis section).

Table 4.2.

*Summary of the design*

| Animacy condition | Marker condition | Category distribution | Marker distribution | Out of one block of 32 trials…… |
|---|---|---|---|---|
| Skewed animacy | Skewed marker | 6 nouns in Category 1 | 75% times Marker 1<br>25% times Marker 2 | 24 trials Category 1<br>- 18 with Marker 1<br>- 6 with Marker 2 |
| | | 2 nouns in Category 2 | 75% times Marker 1<br>25% times Marker 2 | 8 trials Category 2<br>- 6 with Marker 1<br>- 2 with Marker 2 |
| Uniform animacy | Skewed marker | 4 nouns in Category 1 | 75% times Marker 1<br>25% times Marker 2 | 16 trials Category 1<br>- 12 with Marker 1<br>- 4 with Marker 2 |
| | | 4 nouns in Category 2 | 75% times Marker 1<br>25% times Marker 2 | 16 trials Category 2<br>- 12 with Marker 1<br>- 4 with Marker 2 |
| Skewed animacy | Uniform marker | 6 nouns in Category 1 | 50% times Marker 1<br>50% times Marker 2 | 24 trials Category 1<br>- 12 with Marker 1<br>- 12 with Marker 2 |

| | | | | |
|---|---|---|---|---|
| | | 2 nouns in Category 2 | 50% times Marker 1<br>50% times Marker 2 | 8 trials Category 2<br>- 4 with Marker 1<br>- 4 with Marker 2 |
| Uniform animacy | Uniform marker | 4 nouns in Category 1 | 50% times Marker 1<br>50% times Marker 2 | 16 trials Category 1<br>- 8 with Marker 1<br>- 8 with Marker 2 |
| | | 4 nouns in Category 2 | 50% times Marker 1<br>50% times Marker 2 | 16 trials Category 2<br>- 8 with Marker 1<br>- 7 with Marker 2 |

The lack of contingency between markers and categories described earlier in this section, can also be expressed formula introduced in Chapter 1 (Allan, 1980, Equation 2). Here, P(O|C) represents the probability of the outcome in the presence of the cue (here, the presence of one of the two markers with one of the categories), whereas P(O|¬C) represents the probability of that same outcome in the absence of the cue (here, the probability of the same marker in the presence of the alternative cue). In all of our conditions, these two values are the same, and hence, the contingency (ΔP) is 0. Let's see for example, the skewed marker – skewed animacy condition. As shown in Table 4.2, the probability of marker 1 with category 1 is of .75 (18 out of 24), while the probability of marker 1 with category 2, is also .75 (6 out of 8), even if category 2 is, overall, less frequent than category 1.

$$\Delta P= P(O|C)-P(O|\neg C). \qquad (3)$$

$$\Delta P = .75 - .75 = 0$$

## Tasks

### *Phase 1, Noun training*

Participants were presented with each of the eight nouns in a random order, both in written and auditory format together with their corresponding image. After each of the presentation, they were asked to repeat the noun aloud. Their voice was recorded, and their answers were used as an attention check. Each noun was presented three times, once within each block of eight trials.

### *Phase 2, Noun testing*

Participants completed a two-alternative forced-choice (2AFC) task, and subsequently a free recall task, to test their knowledge of the nouns. In the 2AFC task, they were either presented with a noun and two images to choose between (the one corresponding to the noun and a randomly chosen foil image) or with an image and two nouns (the one corresponding to the image and a randomly chosen foil noun). These two types of trials were interleaved. Each of the nouns was the target twice (once in each of the types of trials). The order of the nouns and the position of the target in the screen (right or left) were randomised.

In the free recall task, participants were presented with the image of each of the objects, one by one, together with some dotted lines representing the number of characters in each of the words and they were asked to orally produce the noun that corresponded to it while their voice was recorded.

***Phase 3, Plural training***

This was an 2AFC task containing two types of trials: plural training trials and filler trials. In both types of trials, participants could see an avatar in the upper left corner, who symbolised the speaker who produced the utterance in the trial. In the plural training trials participants were presented, both in written and auditory format, with a sentence containing a plural marker and a noun and were asked to choose between two images, the target containing two of the referents corresponding to the noun, and the foil containing a single referent (e.g., they were shown "bok tebas" and presented with an image with two buses and an image with a single bus to choose between). The filler trials were identical to the 2AFC trials in the noun testing phase: participants were presented with a noun (in singular) and asked to choose between an image with the referent corresponding with the noun and a randomly chosen foil image. This way, participants implicitly learnt that the presence of the marker encoded plurality, and had the opportunity to explore the meaning of the two markers. Participants received feedback after each trial (after they clicked on an image, they could either see "Correct!" or "Wrong." on the screen for 1000s, together with a distinctive sound for each option). Participants were presented with four blocks of 40 trials each: 32 plural training trials (four with each noun as the target) and 8 filler trials (one with each noun as a target). The frequency distribution of the marker depended on the experimental condition, as well as the number of animate and inanimate nouns in the language they learnt. Responses in this task were used as an attention check and to ensure that participants had understood that the markers marked plurality.

***Phase 4, Plural testing***

This task was divided into two parts: a 2AFC comprehension task and a free recall production task. In the comprehension task, participants could see an avatar and a marker (both in written and auditory format) and were presented with two images, one containing an animate referent and the other containing an inanimate referent, together with their corresponding nouns. The avatar was the same as in the “Plural Training phase”, and was the one producing the marker, and whose next production they were asked to predict. Participants were asked to choose the one that they thought followed the marker. They did not receive any feedback on their selections. The task also contained singular filler trials, which followed the same structure as the ones in the plural training phase, and for which they received feedback. The task contained 40 trials (32 testing trials, and 8 filler trials, one per noun). In the plural trials, each of the referents was the target four times, and the associated marker represented the distribution participants were presented with in the plural training phase. For instance, a participant in the skewed marker condition for which “nim” was the majority marker and “bok” the minority marker would have been presented with three trials in which “nim” was the marker and one in which “bok” was the marker for each of the target nouns. The order of presentation of the trials was fully randomised.

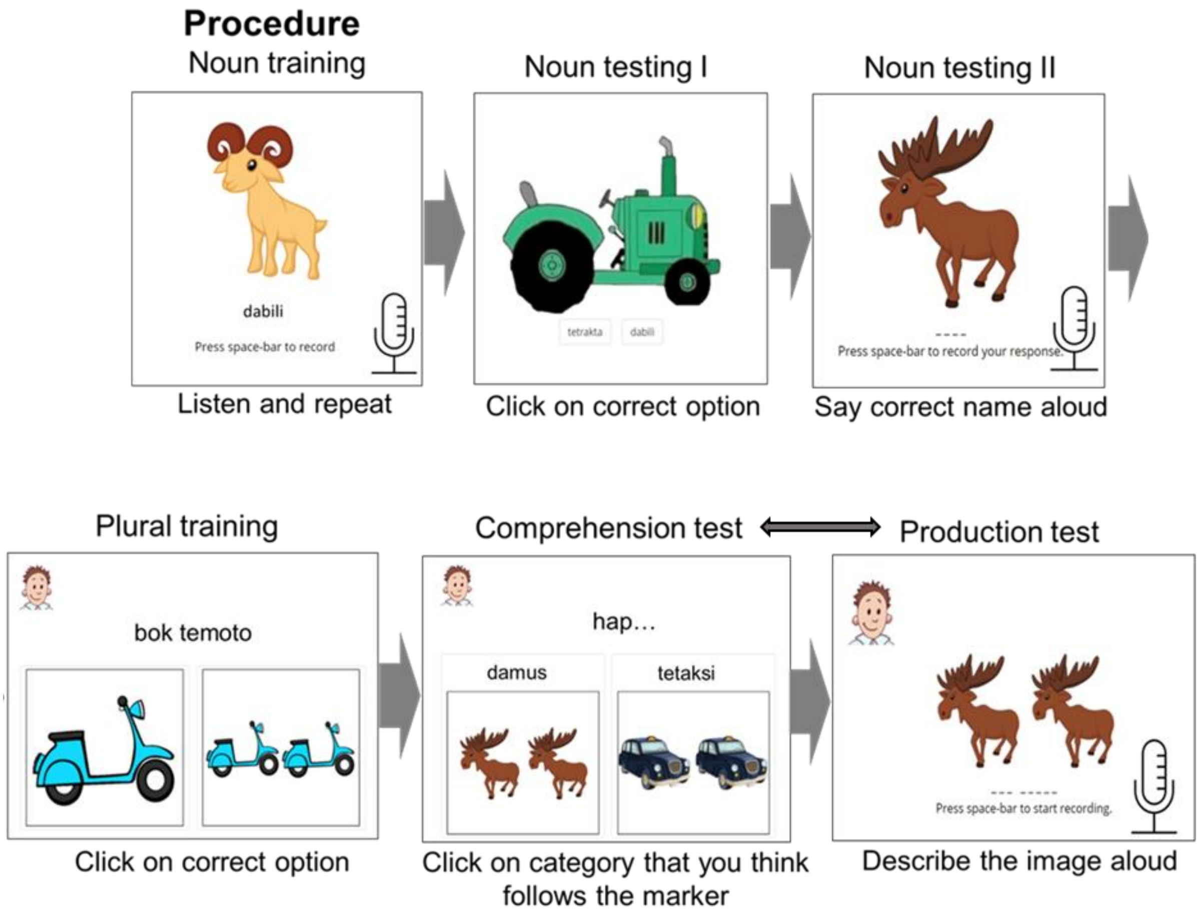
The production task was a free recall test in which participants were shown an image and asked to describe it aloud, while we recorded their answer. They could see dotted lines under the image, representing the number of characters the target production contained. The image could either contain two referents (test trials) or a single referent (filler trials). Participants only received feedback on the filler trials, where they were presented with the image and its corresponding noun both in written and auditory format after they had produced their response. The task contained two blocks of 40 trials (32 test trials and 8 fillers). The order of presentation was randomised within each block. Figure 4.3 shows a visual representation of the different tasks.

***Phase 5, Post-test questionnaire***

Participants completed a short demographic questionnaire (gender, age, number of languages they speak). Then, they were asked to estimate the frequency of different elements of the artificial language through four different sliders. Finally, they answered open-ended questions about their perception of categories, the skewness of this categories, and the conditioning between categories and markers. We will come back to the results of the open questions and slider questions, and how they related to participants’ behaviours in Chapter 6.

Figure 4.3.

*Visual representation of the different experimental tasks*



*Note.* The order of the comprehension and production tasks was counterbalanced. The microphone icon represents those tasks in which participants were asked to verbally produce and record their answer.

## Procedure

Participants completed the study online using their own devices. They were required to use a computer or a laptop to complete the experiment, using Chrome or Firefox as their browser, for which the experiment was tested and optimised. We asked to complete the study in a quiet environment free from interruptions and we also asked participants not to take breaks or to take any notes. Participants were first presented with an information sheet presenting the study and its aims, and after 10 seconds a button appeared to enable participants to proceed to the next screen. Then they were presented with an informed consent form in which the conditions of participation were described. Participants were asked to complete the experiment in a single session. If they were inactive for more than 5 minutes after the noun training phase has started, their session timed out and they were not able to take part in the experiment. Before starting the experiment, participants were shown a sample audio at the same volume as the audio recordings asking them to adjust the volume to a comfortable level. They were also asked to test their microphone, by recording a short audio and testing whether their production was audible.

The experiment software was developed in JavaScript using the jsPsych libraries (de Leeuw et al., 2023). Participants were taught and tested on the artificial language through a series of tasks presented in the order described in the task section. At the start of each task, participants were presented with text and audio instructions in English by a single male speaker (the purported "native speaker" of the language) on how the task would proceed. This same speaker also produced the audio stimuli for the first two task phases. After this, a different male "native speaker" was introduced by the first speaker, who taught participants more complex aspects of the language. This speaker produced the audio stimuli in the remaining tasks. For those tasks involving oral production, participants were asked to press the spacebar to start and stop the recordings at their will. At the end of the experiment, the post-test questionnaire was administered, and participants were debriefed. We also asked them whether they had taken any notes, making it clear that this would not affect their payment, to encourage honesty. If they had declared having taken notes, we would have excluded their data from the analysis, but this was not the case for any of the participants. Participants took an average of 26.41 minutes to complete the task (SD = 4.78, range 15-46).

## Measures & Indexes

### *Comprehension task*

Our main analyses were based on participants' behaviours in the comprehension task and in the production task. In the comprehension task, we coded the category to which the images clicked in each of the trials belonged as animate or inanimate. We then calculated the proportion of animate and inanimate selections with each of the markers. From there, we extracted the following indexes, which we then used as the outcome or fixed factor of our models:

**Most Conditioned Category.** This referred to the category that a participant chose with the highest frequency for a given marker (whether or not this category was the one used the most across both markers). For example, as Participant 1 in Table 4.3, if when presented with "hap", a participant clicked 50% of the times in an "animate" category image and 50% in an "inanimate" category one, whilst when presented with "bok" the distribution of their selections was 25% “animate” and 75% “inanimate”, the “Most Conditioned Category” would be “inanimate”. If the proportion of category selection was equal for both markers, the mapping for most/least conditioned category was selected randomly.

**Most Skewed Marker.** This was the marker for which the category distribution was most skewed (further from .5). In the example of Participant 1 in Table 4.3, this would be the marker “bok”, as the difference between the frequency with which participants chose each of the categories was higher for this marker. Once again, if the proportion of category selection was equal for both markers, the mapping for most/least conditioned marker was selected at random.

**Output Majority Category.** Across both markers, we calculated which was the category that was selected most frequently by each participant. For Participant 1 in Table 4.3, this would be “inanimate”. Most Conditioned Category and majority category did not necessarily match, as the proportion of trials with one marker varied by condition (as it was the case for Participant 3 in Table 4.3).

Table 4.3

*Examples of comprehension task measures for different participant behaviour*

| Participant | Marker condition | Marker 1 | | Marker 2 | | Most Conditioned Category | Most Skewed Marker | Output Majority Category |
|---|---|---|---|---|---|---|---|---|
| | | Inanimate | Animate | Inanimate | Animate | | | |
| 1 | Uniform | .5 (8/16) | .5 (8/16) | **.75 (12/16)** | .25 (4/16) | Inanimate | Marker 2 | Inanimate |
| 2 | Uniform | .125 (2/16) | **.875 (14/16)** | .75 (12/16) | .25 (4/16) | Animate | Marker 1 | Animate |
| 3 | Skewed | .125 (1/8) | **.875 (7/8)** | .75 (18/24) | .25 (6/24) | Animate | Marker 1 | Inanimate |

*Note.* The table represents what the measures “Most Conditioned Category”, “Most Skewed Marker” and “Output Majority Category” would be for three participants, based on the number of trials they had selected each category in. The second column indicates the marker condition a participant was assigned to. The next four columns indicate the proportion of trials in which a participant chose each of the categories for each of the markers. The values indicate proportion of trials in which each of the categories was chosen with a marker, with the number of trials and the total number of trials with that marker between parentheses. The highest proportion of all four columns is bolded for each participant.

### *Production task*

In the production task, we recorded the marker that the participant produced in each of the trials. It was transcribed following the phoneme-to-grapheme rules of Tok Pisin for it to be comparable with Panitok (see Appendix B for a description of how each of the phonemes was transcribed) and then coded. The transcribers were blind to the condition participants had been assigned to.

To code the markers, we started by measuring the Levenshtein distance[5] (Levenshtein, 1966) between the transcription of the marker the participants' produced and each of the markers they were presented with. If the distance was 0 or 1 to any of the markers, and over 2 to the other marker they were presented with, we coded the response as the former marker. For instance, if a participant who was presented with "hap" and "bok" as markers produced "han" in a given trial, this would be coded as "hap", since the Levenshtein distance between "hap" and "han" is of 1, and between "bok" and "han" is of 3. We excluded those trials in which the distance was more than 1 to both markers (e.g., if that same participant produced "hun"), and those in which the distance was 1 to one of the markers and 2 to the other (e.g., a participant that was been presented with "hap" and "bok" says "hop"). We coded the noun production as correct or incorrect by calculating the distance between a production and a trial.

We also excluded those trials in which the noun participants had to produce had a distance of more than 2 to the target noun (e.g., if the target noun was "tetaksi" and the participant produced "temoto"). As with the category choice in the comprehension task, we calculated the proportion of trials in which participants produced each of the markers with each of the categories, and based on that, we computed the following indices:

**Most Conditioned Marker.** This index referred to the marker that was produced with the highest frequency for a given category. We followed the same strategy as for the "Most Conditioned Category" in the comprehension task, with the difference that "Marker" was the outcome measure in this task.

**Most Skewed Category.** This was the category for which the marker distribution was most skewed (further from .5). Once again, for this measure, we followed the same strategy as for "Most Skewed Marker" in the comprehension task, with the main difference laying in the fact that "Category" was our predictor here.

[5] Recall that Levenshtein distance represents through an integer value the minimum number of operations (addition, deletion, and replacement) required to transform a text string into another. A value of 0 represents that the two strings are identical. The higher the value, the more dissimilar the strings are.

**Output Majority Marker.** Across both categories, we calculated which was the marker that was produced most frequently for each participant. This was equivalent to the "Output Majority Category" in the comprehension task. Table 4.4. shows these measures in action in the production task, for three fictitious participants.

We selected these measures instead of the more traditional Mutual Information (i.e., Samara et al., 2017) because the skewness in our stimuli biased this measure, which led to artificial differences between conditions, and as suggested by Kirby (2008), the use of logistic measures was more advisable than the use of proportions.

Table 4.4.

*Examples of production task measures for different participant behaviour*

| | | Animate | | Inanimate | | Most | Most | Output |
|---|---|---|---|---|---|---|---|---|
| | Animacy | | | | | Conditioned | Skewed | Majority |
| Participant | condition | Marker 1 | Marker 2 | Marker 1 | Marker 2 | Marker | Category | Marker |
| 1 | Uniform | .5 (16/32) | .5 (16/32) | **.75 (24/32)** | .25 (8/32) | Marker 1 | Inanimate | Marker 1 |
| 2 | Uniform | .125 (4/32) | **.875 (28/32)** | .75 (24/32) | .25 (8/32) | Marker 2 | Animate | Marker 2 |
| 3 | Skewed | .125 (2/16) | **.875 (14/16)** | .75 (36/48) | .25 (12/48) | Marker 2 | Animate | Marker 1 |

*Note.* this table represents what the measures "Most Conditioned Marker", "Most Skewed Category" and "Output Majority Marker" would be for three participants, based on the number of trials they had used each marker in. The second column indicates the animacy condition a participant was assigned to. The next four columns indicate the proportion of trials in which a participant used each of the markers with each of the categories. The values indicate proportion of trials in which each of the markers was used with each of the categories, with the number of trials and the total number of trials with that category between parentheses. The highest proportion of all four columns is bolded for each participant.

### Statistical analyses

We used R 4.2.2 (R Core Team, 2022) to analyse our data. Given that all our outcome variables (described in measures) were binomial, we ran two logistic mixed effects models for our hypotheses on linguistic conditioning, one for the comprehension task and a second one for the production task. Each of the models had the specific behaviour in the trial as the outcome (the category selected in the comprehension task and the marker produced in the production task), and the type of trial as a predictor (which of the markers participants were presented with in the comprehension task, and which of categories the noun belonged to in the production task). Hence, a significant effect of the type of trial would suggest conditioning behaviour. For example, in the comprehension task, if we found a difference on that the probability of selecting a given category, the outcome, differed by the marker participants were presented with, the type of trial, it would mean that participants were showing a different pattern of category choice by marker, or in other words, that they were conditioning their category choice to the marker. We then included the marker and category conditions as fixed factors in the models, and interpreted any interaction of these with the trial type as evidence that the experimental conditions had an effect on the degree of conditioning behaviour.

Specifically, our comprehension task model included “Most Skewed Marker”, animacy condition, marker condition, task order, and their interactions as fixed effects, random intercepts for participant, marker subset (which were the two markers participants learnt out of the possible three), noun subset (which eight nouns participants learnt), item (the target item shown on the screen in that particular trial), and position on the screen, as well a by-participant slope for animacy condition, marker condition and/or task order. The output variable was whether the category that a participant had selected in a given trial was the “Most Conditioned Category” (see measures) or the alternative. A significant effect of “Most Skewed Marker” would indicate that participants had conditioned the categories on the markers, and an interaction with any of the of the other fixed effects indicated that the level of conditioning differed by experimental condition.

Our production task model included "Most Skewed Category”, animacy condition, marker condition, task order, and their interactions as fixed effects, and “Most Conditioned Marker” as the output variable. The random intercept and slope structure was identical to that for the model in the comprehension task, excluding the random intercept for position on the screen, which was not relevant to this task. We established the threshold for significance in .05 and used Laplace’s method of approximation to obtain the degrees of freedom, using lmerTest package (Kuznetsova et al., 2017). We used glmmTMB to execute the models (Brooks et al., 2017).

In the case of non-convergence, we followed the Barr et al. (2013) method and simplified the structure of the models by removing random effects that were highly correlated to each other or that accounted for little variance until convergence was achieved. We used Tukey HSD method to perform nested pairwise comparisons to further explore the significant interactions, and carried out nested analyses of the effects, applying Type III Sum of Squares formulas to estimate the effects within the model ("joint tests" function in the emmeans package; Lenth et al., 2022)

## Results

### Learning outcomes

Before testing our hypothesis, we checked that our participants were proficient in the artificial language after the noun training and plural training phases, and crucially, that there were no significant differences in learning across conditions that could contribute to differences between conditions in our variables of interest. As Table 4.5 shows, the proportion of correct trials was near ceiling for participants in all the training tasks and conditions. We ran 2*2 ANOVA tests with animacy condition and marker condition as predictors and the proportion of correct trials as the outcome and did not find any significant effects.

Table 4.5

Proportion of correct trials in the training tasks by condition.[6]

| | Skewed marker | | Uniform marker | | Differences | |
|---|---|---|---|---|---|---|
| | Skewed animacy | Uniform animacy | Skewed animacy | Uniform animacy | Marker condition | Animacy condition |
| **Task** | Mean (SD) | Mean (SD) | Mean (SD) | Mean (SD) | P - value | P - value |
| Noun testing – 2AFC | .990 (.027) | .991 (.025) | .994 (.021) | .993 (.022) | .243 | 1.000 |
| Noun testing- free recall | .933 (.091) | .956 (.075) | .956 (.075) | .944 (.086) | .553 | .553 |
| Plural training | .988 (.011) | .989 (.010) | .989 (.009) | .989 (.010) | .677 | .930 |

### Conditioning behaviour

#### *Comprehension task*

In the comprehension task, participants were presented with a marker and asked to predict which word would come next. They were presented with two options, one of each category, and asked to click on the one they thought would come next. The outcome variable here was the animacy category their choice in each trial belonged to (whether this was the "Most Conditioned Category" as described in measures, to be precise), and the predictors

[6] In the free recall task of the noun testing phase, we transcribed the participants' oral productions following the indications in Appendix B and considered as correct any production with a Levenshtein distance of 2 or less to the target production.

were the marker participants were presented with in a trial (whether this was the "Most Skewed Marker" as defined in measures), marker condition, animacy condition, and task order. Here, with the aim of increasing the readability of the results, we will refer to each of the effects as it follows:

- The effect of "Most Skewed Marker" shows that participants conditioned the category choice to the marker presented in a given trial. As such, we will refer to it as "evidence of conditioning".
- The interaction between "Most Skewed Marker" and marker condition shows that the conditioning behaviour differed by marker condition, so we will refer to it as "effect of marker condition on conditioning".
- The interaction between "Most Skewed Marker" and animacy condition shows that the conditioning behaviour differed by animacy condition, so we will refer to it as "effect of animacy condition on conditioning". The interaction between "Most Skewed Marker" and task order shows that conditioning behaviour differed by task order. As such, we will refer to it "effect of task order on conditioning".
- Any three- and four-way interactions between these effects will be referred to as two- and three-way interactions between the already-described interactions, and other effects. For example, a triple interaction between "Most Skewed Marker", marker condition and animacy condition, will be described as an interaction between the effects of marker condition and animacy condition on conditioning.

As Figure 4.4 shows, we found significant *evidence of conditioning* [7](β = -1.258, z = -48.13, p < .001), indicating that participants conditioned category choice on marker. We also found a significant *effect of marker condition on conditioning* (β = -.149, z = -5.83, p < .001). Participants in the skewed marker conditions showed a higher degree of conditioning than those in the uniform marker conditions. We did not find an *effect of animacy condition on conditioning* (β = -.030, z = -1.17, p = .240), suggesting that animacy condition did not have an effect on conditioning behaviour. Finally, interaction between the *effect of marker condition and animacy condition on conditioning* was not significant (β = -.009, z =-.36, p = .721).

[7] All interpretative terms has been italised to prevent literal understanding of these effects.

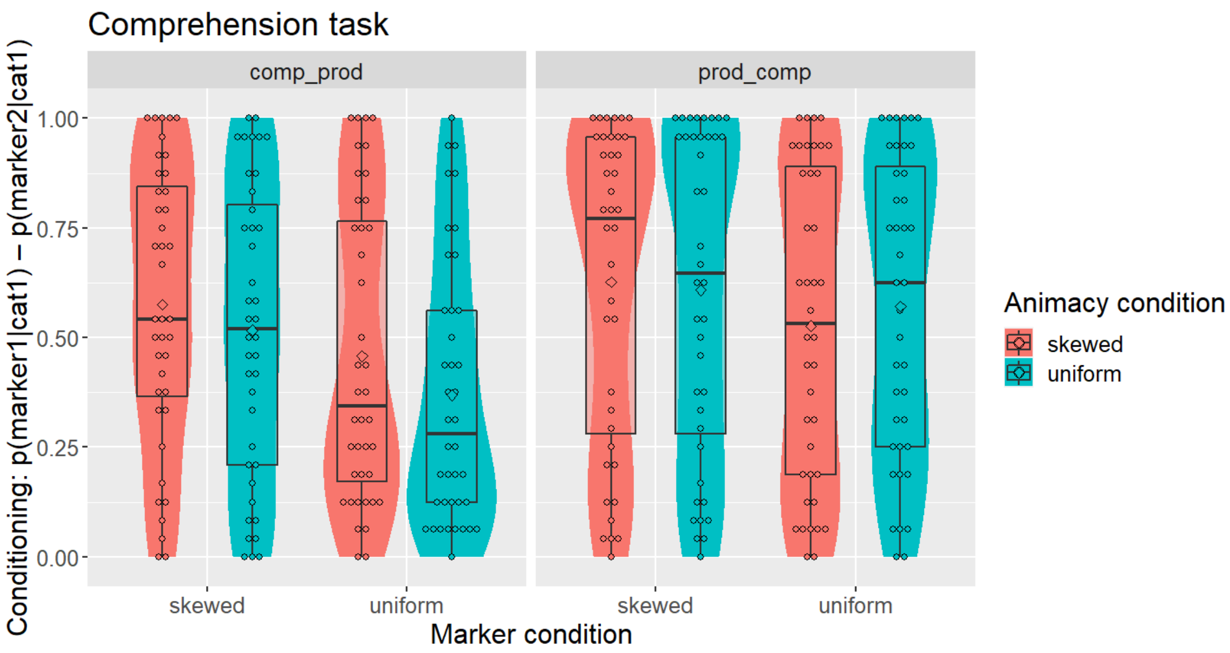
Unexpectedly, we found a significant *effect of task order on conditioning* (β =.143, z =.5.64, p <.001). Participants who completed the comprehension task after the production task conditioned to a higher degree than those who completed the comprehension task first. In addition, we found an interaction between the *effect of task order and of animacy condition on conditioning* (β = -.076, z =-2.97, p = .003), suggesting that the effect of animacy was dependent on task order. To understand the significant three-way interaction, we ran a post-hoc nested models within each of the task-orders.

As expected from interaction, when comprehension task was completed first, there was a significant *effect of animacy condition on conditioning*, F(1, 10219) = 9.181, p = .0025 in the predicted direction (conditioning was higher in the skewed animacy condition than in the uniform animacy condition), but the *effect* was not present when the comprehension task was completed second, F(1, 10219) = 1.505, p = .220.

In summary, and as Figure 4.4 shows, in the comprehension task, a) participants conditioned their category choice on the marker they were presented with, b) this was more likely to happen when the distribution of the marker frequency was skewed, c) this was more likely to happen when the comprehension task was the second one to be completed, and d) this was more likely to happen when the distribution of the animacy frequency was skewed, but only if the comprehension task was the first one to be completed. Animacy condition and marker condition did not interact with each other in any case.

Figure 4.4

*Degree of Conditioning by Task Order, Animacy Condition, and Marker Condition*



*Note.* Y-axis represents the absolute difference in the proportion of trials in which one of the categories with one of the markers vs. the alternative. A value of 0 indicates no difference and no conditioning: the proportion of selection of one category or the alternative was equal for both markers. A value of 1 indicates full conditioning: with one of the markers, one of the categories was selected a 100% whilst with the alternative marker, the alternative category was selected 100% of the times. Left panel shows the degree of conditioning by animacy and marker condition when the comprehension task came first, and the right panel shows conditioning when the comprehension task came second. The two columns on the left within each of the panels belong to the skewed marker conditions, and the two columns on the right to uniform marker conditions. Colours represent animacy condition: orange for skewed and green for uniform.

### *Production task*

In the production task, participants were asked to describe an image containing two items (either two animals or two vehicles). The outcome variable was which of the markers they produced in their description (more precisely, whether it was the "Most Conditioned Marker" or the alternative), and the predictors were the category the item belonged to (whether it was the "Most Skewed Category" as per the description in Measures or the alternative), marker condition, animacy condition, and task order.  Due to a convergence issue with the original model (see Statistical analyses section), we removed "set of nouns" as a random intercept, given that it had the lowest variance. Following the same strategy as for the

Comprehension task and to increase readability, we refer to the effects and interactions as it follows:

- The effect of “Most Skewed Category” shows that participants conditioned the marker production to the category the noun in a given trial belongs to. As such, we will refer to it as “evidence of conditioning”.
- The interaction between “Most Skewed Category” and marker condition shows that the conditioning behaviour differed by marker condition, so we will refer to it as “effect of marker condition on conditioning”.
- The interaction between “Most Skewed Category” and animacy condition shows that the conditioning behaviour differed by animacy condition, so we will refer to it as “effect of animacy condition on conditioning”.
- The interaction between “Most Skewed Category” and task order shows that conditioning behaviour differed by task order. As such, we will refer to it “effect of task order on conditioning”.
- Any three- and four-way interactions between these effects will be referred to as two- and three-way interactions between the already-described interactions, and other effects.

We found significant *evidence of conditioning* (β = -1.115, z = -50.13, p < .001), indicating that participants conditioned marker use on animacy category. We also found a significant *effect of marker condition on conditioning* (β = -.118, z = -6.46, p < .001). Participants in the skewed marker conditions conditioned to a higher degree (showed a higher difference of marker use between categories) than those in the uniform marker conditions. The *effect of animacy condition on conditioning* was marginally significant and in the opposite direction as the predicted one (β = -.038, z =-1.68, p = .093). Participants in the uniform animacy conditions had a slight tendency to condition to a higher degree than those in the skewed animacy conditions. Finally, the interaction between the *effect of marker condition and of animacy condition on conditioning* was significant (β = .045, z =1.96, p = .049). We explored this further in the post-hoc analyses described below.

Once again, we found a significant *effect of task order on conditioning* (β =-.588, z =-25.73, p <.001). Participants who completed the production task second conditioned more than those who completed it first. In addition, we found an interaction between the *effect of* *animacy condition and of task order on conditioning* (β = -.053, z=-2.30, p=.021), as well as between the *effect of marker condition and of task order on conditioning*" (β = -.106, z =-4.63, p<.001), and a four-way interaction between all four factors (β =-.047, z = -2.04, p =.041).

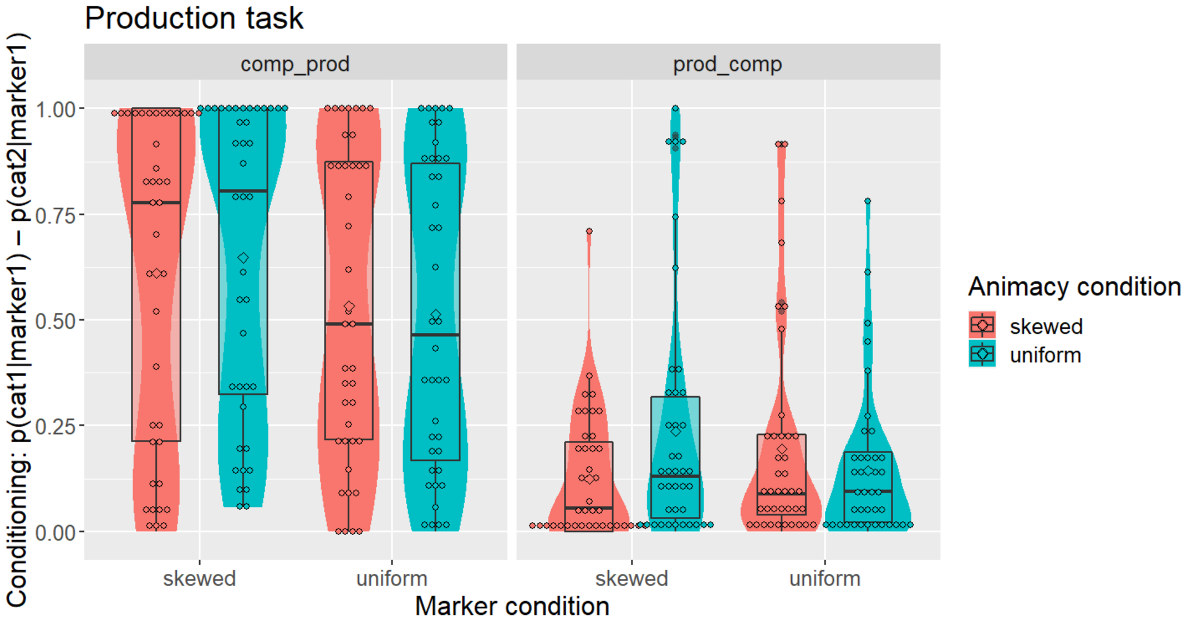
In order to understand the three- and four-way interactions that we found; we ran post-hoc nested analyses of the effects within each of the task orders. When the production task came second, the results were similar to those found in the model that did not account for order: a significant *effect of animacy condition on conditioning*, showing that participants in the uniform animacy condition categories conditioned to a higher degree than those participants in the skewed animacy categories, F(1, 19584) = 6.714, p =.0096; a significant *effect of marker condition on conditioning* showing that participants in the skewed marker conditions conditioned to a higher degree than those in the uniform marker condition, F(1, 19584) = 52.047, p<.001, and no interaction between the *effects of marker and animacy condition on conditioning*, F(1, 19584) = .003 p =.958.

When the production task came first, however, the results were different to those found in the general model. We did not find an *effect of animacy condition on conditioning*, F (1, 19584) = .235, p =.682, nor an *effect of marker condition on conditioning*, F(1, 19584) = 2.066, p =.151. However, as suggested by the four-way interaction in the general model, we found an interaction between the *effects of marker condition and animacy condition on conditioning*, F (1, 19584) = 9.823, p =.0017. Further nested tests on the general model showed that, within the production first task order, the *effect of animacy condition on conditioning* was significant both when marker condition was uniform, F(1, 19584) = 3.961, p= .047, and when it was skewed, F(1, 19584) = 5.888, p = .0153, but in opposite directions (see Figure 4.5).

In summary, in the production task, a) participants conditioned their marker production to the animacy category of the items, b) conditioning behaviour was higher when participants completed the production task after the comprehension task, c) conditioning behaviour was overall higher when the distribution of the frequency of the marker was skewed, d) the effect of a skewed frequency distribution of animacy was dependent on task order and marker condition. To be precise, we found that, when the production task came second, contrary to what we had predicted, participants conditioned to a higher degree when the distribution of the frequency of animacy was uniform than when it was skewed. This was also the case when production task came first, and the frequency distribution of marker was skewed. However, within the production first task order, we found the predicted effect of animacy condition when the frequency distribution of marker was uniform. This can be seen in Figure 4.5.

Figure 4.5.

*Degree of conditioning in the production task by task order, animacy condition, and marker condition*



*Note.* The Y-axis represents the absolute difference in the proportion of trials in which one of the markers was used with one of the categories vs. the alternative. A value of 0 indicates no difference and no conditioning: the proportion of use of one marker or the alternative was equal for both categories. A value of 1 indicates full conditioning: with one of the categories, one of the markers was selected a 100% whilst with the alternative category, the alternative marker was selected 100% of the times. Left panel shows degree of conditioning by animacy and marker condition when the comprehension task came first, and the right panel shows conditioning when the comprehension task came second. The two columns on the left within each of the panels represent skewed marker conditions, whereas the two columns in the right represent uniform marker conditions. Orange violin plots represent skewed animacy conditions, and green plots represent uniform animacy conditions.

## Direction of the conditioning

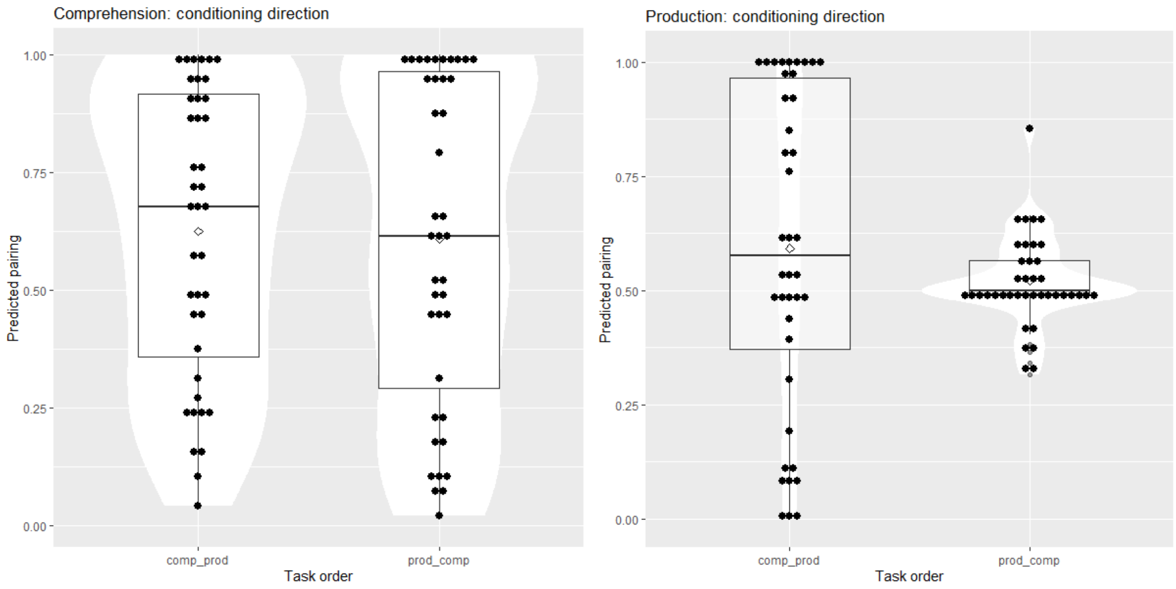
We aimed to test the predictions from illusory correlation theory (Hamilton & Gifford, 1976), which states that participants are more likely to relate majority groups to majority behaviours, and minority groups to minority behaviours. In the context of this study, we expected participants in the skewed animacy-skewed marker condition to associate the use of the majority marker more often with the majority category, and the minority marker with the minority category. To test this, we ran two models, one for each of the tasks.

For the comprehension task, we ran a logistic mixed effects model with "Majority category" as the outcome and "Majority marker", task order, and their interaction as predictors (see Measures for a description), as well as random intercepts for participant, set of nouns, and set of markers, as well as a by-participant slope for task order. This was to test whether the presence of the majority marker predicted the presence of the majority category. The effect of "Majority marker" was significant, indicating that the presence of the majority marker predicted the selection of the majority category, that is, that participants conditioned in the predicted direction (β =.524, z = 10.625, p <.001). Task order did not influence this effect (β = .026, z = .736, p = .462) (see left panel of Figure 4.6).

For the production task, following the same logic, we ran an identical model but including "Majority marker" as the outcome measure, and "Majority category" as one of the predictors. In this task again, we found that participants conditioned the majority marker to the majority category (β = .324, z = 7.498, p <.001), but the effect was bigger when the production task came second (β = .162, z =3.782, p <.001). Figure 4.6 shows the proportion of participants that conditioned in the predicted direction, by task and task order (see right panel of Figure 4.6).

Figure 4.6.

*Proportion of trials in which participants paired the majority marker with the majority category, by task (left panel comprehension, right panel production) and task order.*



*Note.* It is easy to observe that the variability in the proportion of trials in which participants showed the predicted pairing in the Production task, when this came in the first place, was much lower than in the other tasks and task orders, centring around .5. This is due to fact that, as shown in Figure 4.5, participants showed a very low level of conditioning, with many of them regularising to a single marker. If they produced the same marker for all trials, they would naturally produce the predicted pairing in around 50% of the trials.

Before diving into the discussion, Table 4.6 presents a summary of the results in relation to our predictions.

Table 4.6. Summary of results for Chapter 4

| | | | By Task order | |
|---|---|---|---|---|
| Effect | Task | Across task orders | Comp-prod | Prod-comp |
| Conditioning behaviour | Comprehension | Conditioning present | Yes, lower in this task order | Yes, higher in this task order |
| | Production | Conditioning present | Yes, higher in this task order | Yes, lower in this task order |
| Effect of marker condition | Comprehension | Predicted direction –conditioning in skewed > uniform marker condition | No effect of task order | |
| | Production | Predicted direction –conditioning in skewed > uniform marker condition | Predicted direction –conditioning in skewed > uniform marker condition | No effect of marker condition within this task order |
| Effect of animacy condition | Comprehension | Dependent on task order | Predicted direction –conditioning in | No effect of animacy |

| | | | | |
|---|---|---|---|---|
| | | | skewed > uniform animacy condition | condition within this task order. |
| | Production | Dependent on task order | Opposite direction –conditioning in uniform > skewed animacy condition | No effect of animacy condition within this task order. |
| Interaction between animacy condition and marker condition effects | Comprehension | No interaction | No effect of task order | |
| | Production | Four-way interaction with task order | No interaction between marker and animacy conditions | When marker condition is uniform: predicted effect of animacy condition.<br>When animacy condition is skewed: opposite effect of animacy condition |
| Conditioning in the predicted direction | Comprehension | Yes | No effect of task order | |
| | Production | Yes | Higher in this task order | Lower in this task order |

## Discussion

In this experiment, we trained participants on an artificial language containing 8 nouns belonging to two categories, animate and inanimate, and two plural markers. The language they were trained on differed in its statistical properties in two parameters: the skewness of the frequency distribution of the marker (skewed, 75-25% or uniform, 50%-50%), and the skewness of the frequency distribution of the animacy category (skewed, 75-25%, or uniform 50%-50%). We tested them with a comprehension task, where they were provided with a plural marker and asked to choose what they predicted should come next between two options: an animate noun and an inanimate noun, and a production task, where we showed them an image of two animate or inanimate nouns and asked them to describe it. We also manipulated the order in which they completed these tasks.

We found that, overall, regardless of the statistical properties of the artificial language, all participants tended to condition one of the markers to one of the categories beyond what we would predict by chance, despite the fact that these were not contingent on each other in the language they were trained in. We also found that conditioning increased across both tasks, being higher in the second task than in the first, regardless of whether this was the comprehension or the production task. This goes in line with the results by Perfors (2016), who found participants to be biased to find explanations for linguistic variability, and hence, to identify patterns, whether these exist or not. Equally, this does not contradict the results found in research in causal illusion. In our study, as Perfors (2016) predicted, participants seemed to show a bias towards conditioning, regardless of the condition they were placed in. Research in causal illusion shows that participants' prior beliefs mediate the extent of outcome-density and cue-density effects (Blanco et al., 2018; Vicente et al., 2023). For example, Vicente et al. (2023) found that prior beliefs about treatment effectiveness made participants overall more likely to develop an illusion of causality between the use of the treatment and the healing. Hence, we could argue that in our study, participants might have shown a general bias towards conditioning, which has made the illusion of causality more likely to occur when the properties of the language were skewed.

Importantly, regarding our hypotheses, we found that skewness in the distribution of the marker led to a higher degree of conditioning, and this effect was consistent across tasks and task orders. The effect of marker skewness is consistent with the outcome-density effect, commonly cited in the causal illusion literature (Allan & Jenkins, 1980; Alloy & Abrahamson, 1979; Blanco, 2017; Haselton & Buss, 2000; Haselton & Nettle, 2006; Msetfi et al., 2005): when an outcome (here the marker) is high in frequency, participants are more likely to believe that it is linked to a cause that it is not contingent with.

The effect of skewness in the distribution of animacy did not consistently lead to a higher degree of conditioning. We only found this effect in the comprehension task, when it came first, and in the production task, when it came first and only within the conditions that had uniform frequency distribution of the marker. Indeed, we found the opposite effect within some of the conditions in the production task. Furthermore, the size of the effect of animacy condition, regardless of the direction, was notably smaller than that of marker condition. Finally, we did not find the predicted interaction between marker condition and animacy condition.

There are many potential explanations for this. It could be the case that, as we were using categories that existed in the natural world (animals and vehicles), participants were biased by their prior frequency distributions of these elements and were hence not sensitive to our manipulation, particularly given that prior beliefs are known to affect causal illusion (Blanco et al., 2018; Diaz-Lago et al., 2023; Vicente et al., 2023). For example, Diaz-Lago et al. (2023) ran a study looking at the effect of a medicine price over its perceived effectiveness. They found that, even if the medicine was not effective, it was perceived as more effective when it was more expensive. Most importantly, they found that this effect was moderated by the number of doses of the medicine that they estimated they had been given. Participants who were presented with the expensive medicine believed that the number of doses given out was higher than participants who were presented with the cheap medicine, even though the number was the same. The estimated number of doses correlated with the perceived effectiveness of the medicine. In summary, it was participants perceptions of the frequency of the cue (the medicine) that altered their perception of causality, and not the frequency of the cues themselves. In this case, participants prior beliefs on the distribution of animals and vehicles could have altered the perception of their skewness in the miniature language, and therefore the results.

In addition, some of the participants showed no evidence of having consciously perceived the categories, precisely 121 of them (further details in Chapter 6). These participants were less likely to condition category on marker. Indeed, we explored conditioning when excluding those participants that had not perceived the categories and found that some of the effects that were previously not significant, became significant. Initially, we had only found the predicted effect of animacy condition on the comprehension task when this came first. After excluding those participants who did not perceive categories, we also found the predicted effect of animacy condition in the production task when this came second ($\beta = -.070962$, $z =-2.51$, $p =.012$ for the triple interaction between category, animacy condition and task order, and $F(1,14356) = 6.658$, $p =.009$, for the post-hoc interaction between category and animacy condition within the comprehension first task order). Furthermore, when excluding these participants, we also find the predicted effect of animacy condition across both task orders in the comprehension task: participants in the skewed animacy conditions show a higher degree of conditioning than participants in the uniform animacy conditions ($\beta = .069$, $z = -2.17$, $p = .029$).

Also, when looking at the direction of the conditioning within the skewed marker and skewed animacy conditions, we find that, as predicted, participants conditioned the majority marker to the majority category and the minority marker to the minority category. Hence, their pairing was not arbitrary, even if conditioning was not always higher than when animacy category was uniform. This follows the findings and predictions in the field of illusory correlation (Fiedler, 1991; Hamilton & Gifford, 1976; Smith & Alpert, 2007; Van

Dessel et al., 2021), who found that people tend to associate majority behaviours to majority groups and minority behaviours to minority groups.

Though unexpected, the effect of task order can be seen as revealing. The task in which participants conditioned the most was the second, whether that was the comprehension task or the production task. That shows that participants continued learning during testing, after training. The conditioning level was the lowest when the production task came first. We observed that many of the participants were fully regularising the language, using exclusively one of the markers for all nouns. This could be due to participants forgetting the other marker, as the responses to the open-ended questions suggest, as well as participants' failed attempts to produce an alternative marker, often starting to produce one that was similar to the original alternative. If participants could only produce a single marker, they would use that marker 100% of the times for each category, and hence, the level of conditioning could not be above 0.

We also explored how conditioning changed block-by-block within each of the tasks. We observed that, within the comprehension task, conditioning increased from the first 16 trials to the last 16 trials, regardless of the condition and task order. However, we did not find that pattern within four 16 trial blocks of the production task, with the level of conditioning remaining relatively stable across blocks, regardless of the condition and task order. When production task was first, this could be due to the full regularisation and lack of experience with the artificial language required to complete a free recall task. However, we also find this pattern when production was second. Whichever level of conditioning participants in a given condition had reached in the comprehension task was then maintained in the production task. Hence, conditioning seems to be developing during the comprehension task. This could be due to the lower cognitive demands of the task (Hudson Kam & Chang, 2009; Hudson Kam, 2019), or to the salience of the categories: in the comprehension task, participants have to choose between two categories in each trial. This task, closer to the paradigms in illusory correlation, in which participants are often presented with a cue and asked to predict whether the outcome will follow or not, but further away from language production, could have given participants' cues on the structure of the language and the importance of the categories. As stated before, perceiving the categories, whether implicitly or explicitly, was a prerequisite for the development of conditioning, and explicitly identifying which are the cue and the outcome whose contingency participants are asked to judge is common in the field of causal judgment and illusory correlation (Allan & Jenkins, 1980; Alloy & Abrahamson, 1979;

Blanco, 2017; Blanco et al., 2013; Haselton & Buss, 2000; Haselton & Nettle, 2006; Matute et al., 2015; Msetfi et al., 2005; Vicente et al., 2023).

It is also important to note the conceptual difference between these two tasks: in the comprehension task, participants were presented with an avatar of the person that had been "teaching" them the language and ask to predict their productions. In the production task, they were asked to imagine themselves using the language. It could be the case that the results in the comprehension task reflect participants perceptions of the language they leant, whereas additional processes could be involved in the production task, which could include active choices by participants not to align with the probabilities they had perceived. Studies in the field of psycholinguistics have shown how other factors aside from the perception of linguistic patterns affect production, such as pragmatic assumptions (Perfors, 2016), production effort (Hudson Kam & Newport, 2009; Hudson Kam, 2019), biases for communicative efficiency (Fedzechkina et al., 2012, 2017; Kirby et al., 2008, 2015), transmission (Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Smith, 2022) or interaction (Feher et al., 2016, 2019; Smith et al., 2017).

Future research could look at these questions using materials for which participants do not have any priors and making categories more salient, for example, using categories that do not exist in the natural world, such as types of alien food, similar to the stimuli used by Lai et al. (2020) and making them more salient, by explicitly presenting them as noun categories. It is also important to extend the training in the language, to ensure that participants are familiar enough with it to produce it.

Here we brought together the predictions from domain-general associative learning and statistical processing and language acquisition, and evolution. These fields, however, have some important differences. Most research in illusion of causality focuses on the link between causes and effects that are relevant for survival, such as the identification of allergens (Van Hamme & Wasserman, 1994; Lee & Lovibond, 2021) or the effectiveness of medicines (Blanco et al., 2018; Diaz-Lago et al., 2023; Vicente et al., 2023), in which a bias towards identifying patterns can have lifesaving consequences, with its evolutionary implications (Haselton & Buss, 2000; Haselton & Nettle, 2006; Blanco, 2017). However, the theoretical underpinnings of this phenomenon are based on associative learning, which has been found to predict behaviour in variable contexts with a lower evolutionary relevance, such as the development of conspiracy and pseudoscientific beliefs (Blanco et al., 2011; Matute et al., 2015). Equally, previous research in the linguistic field have found some of the principles of associative learning to apply in language (Nixon, 2020; Ramscar et al., 2013). It is therefore reasonable to assume that, given that our results matched some of the predictions stemming from associative learning, there would be linked to the same underlying processes.

In Chapter 6, we will return to these results and explore them in relation to the reported perception of the statistical properties of the language participants were presented with and the language they produced, as well as discussing how they relate to the existing theories and research in statistical learning and psycholinguistics, and the wider implications of our findings.

# Chapter 5: Category accentuation and linguistic conditioning

In Chapter 4, we showed how the tendency of individuals to reduce unpredictable variation and linguistic uncertainty through conditioning is sensitive to the statistical properties of the language, following some of the patterns that we find in domain-general learning of cue-outcome relationships, where people perceive statistical relationships that do not exist (Allan & Jenkins, 1980; Alloy & Abrahamson, 1979; Blanco, 2017; Haselton & Buss, 2000; Haselton & Nettle, 2006; Msetfi et al., 2005). However, the increase in structure we observe in natural languages (Bickerton, 1984; Singleton & Newport, 2004) is often the outcome of a gradual process, in which some level of statistical relationship (probabilistic conditioning) gradually transforms into deterministic conditioning, that is, in which the degree of conditioning as defined in the previous chapters goes from more than zero (probabilistic conditioning) to 1 (deterministic conditioning). This phenomenon is the focus of Chapter 5.

Interactive processes have been shown to affect this shift from probabilistic to deterministic conditioning. Feher et al. (2019) showed in their study how, when a person that uses a linguistic element variably interacts with someone that uses it deterministically (either always or never), the former tends to mimic the latter, and they both converge in a deterministic use of the variant. However, other studies have also found the change in conditioning to happen without the effect of interaction. As covered in earlier sections, studies in language evolution have tried to replicate this phenomenon in the laboratory, using artificial languages (see Smith, 2022 for a review). For instance, Smith and Wonnacott (2010) found that when using iterated learning, that is, when creating chains of participants in which the next generation learns from the output that the previous one produced, small statistical biases in individual participants gradually amplify and evolve into deterministic rules. Further studies have replicated this gradual increase in structure using different mechanisms such as an increase in the size of the community of speakers (Fay et al., 2010; Raviv et al., 2021) or the meaning space (Raviv et al., 2019), starting from unstructured languages containing unpredictable variation and leading to compositional linguistic systems that follow deterministic grammatical rules rather than probabilistic rules.

One of the accounts from linguistics for this phenomenon is the bottleneck theory, which explains that, given the limited cognitive ability to process language while it is received, participants reconstruct the structure of the language from the elements they have retained from the limited exposure they have had, which leads to an imperfect reproduction of the language (see Christiansen & Chater, 2016 for a review). This effect biases the languages towards learnability, which stems from simplicity and structure, as learnable languages are more likely to be reproduced accurately. However, learnability is not the only pressure that shapes languages. Successful communication requires language to communicate meanings unambiguously. The combination of these two forces leads languages to be highly efficient, and nearly perfect in the balance between learnability and expressivity.

Artificial language learning studies have gathered support for this hypothesis. Kirby et al. (2008) showed that eliminating the pressure for communication, and hence for expressivity, led a randomly generated language to evolve to one that was high in learnability (as it was composed by a small number of short words) but low in expressivity (the words did not allow meanings to be unambiguously distinguished, with a high number of homonyms), whereas when including a pressure for communication, like in Kirby et al. (2015), this randomly generated languages evolved into efficient languages, with a higher learnability than the input language whilst retaining the expressivity. Numerous studies since have found results supporting the theory of communicative efficiency (Gibson et al., 2019 for a review). As we found in Chapter 4, with the reduction of unpredictable variation, however, other domain-general theories in the perception of probabilities can provide us with information about the specific conditions under which this reduction in variation occurs and how the statistical distribution of the elements in the language affect it.

In parallel to research in psycholinguistics, cognitive and social psychology have long explored how the perception of all sorts of stimuli can be biased when they are divided in discrete categories. This phenomenon is referred to as category accentuation, by which the difference between elements in a continuum is amplified, and the differences between elements of the same category minimised, when these are categorised. In their original studies, Tajfel and Wilkes (1963) asked their participants to estimate the length of several lines which varied in length across a continuum. When the lines were presented in categories that were consistent with their length (i.e., category A for the shortest half and category B for the longest half), participants tended to overestimate the difference between the lines in the boundary, by underestimating the length of the longest line of category A, and overestimating the length of the shortest line in category B. This phenomenon was found to apply not only to lines, but to other sorts of visual stimuli (Fried & Holyoak, 1984), and most importantly, to the social categorisation of facial features (Corneille et al., 2004) and associated traits

(Krueger & Rothbart, 1990). This led to the "social self-categorisation theory" which has been used to study and understand many different social phenomena such as stereotyping and political polarisation (McGarty & Penny, 1988; Fernbach & Van Boven, 2021).

A well-documented similar phenomenon in psycholinguistics is that of the impact of literacy and reading skills on phonemic categorisation (Morais et al., 1979, 1986; O'Brien et al., 2018). Phonemic categorisation is the process by which we are able to distinguish the phonemes in our language, and appropriately classify them. For example, the phoneme /t/ in English is pronounced in many ways depending on its position in the world or the accent of the speaker, among many other factors. From infancy, we learn to identify all those sounds as a /t/ based on, for example, their frequency (Maye et al., 2002, 2008). However, even in adulthood, the ability to correctly classify these phonemes is variable, and more importantly strongly related to reading skills (Kolinsky et al., 2021; Hoonhorst et al, 2011; Morais et al., 1979, 1986; O'Brien et al., 2018). Several studies have shown that, learning an alphabet, increases the accuracy of these classifications, and makes the boundaries between phonemes more defined (Kolinsky et al. 2021; Hoonhorst et al., 2011; Morais et al., 1986), such as the boundary between voiced and voiceless phonemes like /d/ and /t/. In summary, learning the graphemes that correspond to different phonemes makes people perceive a bigger difference between phonemes in a continuum, depending on which grapheme they correspond to. Equally, this process is language-dependent, and learning a different spelling and phonetic system leads participants to identify the boundaries in different places on the continuum (Reinisch et al., 2013). This process is arguably similar to that of category accentuation. In both cases, the presence of categories (or graphemes) leads to the perception of a bigger difference between items in a continuum (whether these are lines or sounds).

Aside from the effect of categorisation on the perception of continuous variables, further studies in category accentuation have used categorical stimuli to explore this phenomenon and how, similarly to the results that we find in illusory correlation, the frequency distribution of members of each group shapes the results (Krueger & Rothbart, 1990; Krueger et al., 1989). They proposed that participants paid more attention to those members who confirmed the stereotype, leading to a biased perception, which is similar to the explanation provided for the illusory correlation effect.

Indeed, category accentuation and illusory correlation have been linked in the past. As we covered in Chapter 4, McGarty et al. (1993) and later Costello & Watts (2019) and Bott et al. (2021) argue that illusory correlation is no more than an amplification of existing differences from a sample that is assumed to be biased. Concretely, McGarty et al. (1993)'s account stated that both illusory correlation and category accentuation could be explained by the same phenomenon. They stated that both effects were based on a general numerical difference in the number of majority and minority traits in each of the groups, with the difference being bigger in absolute numbers for the majority group. Hence, their account states that the association between the majority group and the majority behaviour to be stronger than that between the minority behaviour and the minority group.

Sherman et al. (2009) also tried to use an integrative account of these two phenomena. They propose that Attention Theory (AT) could account for both phenomena. According to this theory, participants first learn the association between the majority group and their trait(s), as they are exposed to more exemplars of this. Then, they shift their attention to the minority group and focus on those traits that distinguish the group from the majority group. This process has been explored through eye-tracking, which showed that, as the learning process progresses, participants shift their focus from the traits that characterise the majority group to those instances that distinguish the minority group from the majority one (Kruschke et al., 2005)

Finally, the original account, based on distinctiveness (Hamilton & Gifford, 1976), stated that participants find examples of the minority group, as well as the least frequent behaviours, distinctive due to the low frequency, and that hence, those examples in which people from the minority group are described the low frequency traits will be particularly salient. Hence, the association between the minority group and the minority trait will be made first and be stronger, whereas the association between the majority group and the majority trait will only be made by contrast with the existing perceived relationship.

The main difference between these accounts focuses on two aspects. The Attention Theory (AT) predicts that the effect of category accentuation will be amplified if there is a skewness in the number of elements in each category, that is, if there is a majority and minority category, whereas Hamilton and Guilford (1976)'s accounts and McGarty et al. (1993)'s account would not predict such difference. Crucially, Attention Theory predicts that the association between the minority category and its associated trait will be higher than that between the majority category and its associated trait, in contrast with Hamilton and Guilford (1976)'s accounts and McGarty et al. (1993)'s accounts, which predict this asymmetry to be the opposite, with a stronger association forming between majority groups and their associated traits. In a series of experiments, Sherman et al. (2009) put these accounts to test and found that Attention Theory best explained the observed results.

But how can this be translated into the linguistic field, and how can we test the specific predictions of these theories? As we saw in Chapter 4, participants not only add structure to the language by reducing uncertainty, but they do it following the patterns that we find in illusory correlation and illusion of causality research. In addition, iterated learning research has observed how existing differences tend to be amplified (Reali & Griffiths, 2009; Smith & Wonnacott, 2010). In this instance, we aim to test whether the increase in conditioning that we observe in language is also dependent on the statistical distribution of the frequencies of the category.

To answer these questions, we designed a study in which the marker distribution was dependent on category: it was conditioned, and we manipulated the skewness in the frequency of the categories. Each of the markers was prevalent with one of the categories (e.g., Marker 1 was used for category 1 75% of the time, whereas Marker 2 was used for category 2 75% of the times). As in the experiment in Chapter 4, we manipulated the frequency distribution of the animacy across conditions. We expected both groups to condition marker use on the category, as in the input language. In the uniform animacy condition, both linguistic and cognitive approaches to language structure predict participants to condition the language more than in their input. However, we had a few competing hypotheses regarding the skewed animacy condition.

According to the Attention Theory account (Sherman et al., 2009), participants in this condition would show a higher level of conditioning than participants in the uniform category condition. However, as we covered in Chapter 2, when faced with a skewed distribution of a variant, participants sometimes reduce uncertainty by regularising to the most frequent variant, in this case, the most frequent marker. This is particularly likely in the production task, due to its highest cognitive load on retrieval (Ferdinand et al., 2019; Hudson Kam & Chang, 2009; Hudson Kam, 2019): participants in this task have to remember and retrieve the correct nouns and markers, in contrast with the comprehension task, where they are provided with them. Therefore, participants could reduce uncertainty regularising to the most frequent marker and discarding the alternative, given its higher frequencies across categories (precisely, .625). Regularisation and conditioning often have an inverse relationship. If a participant fully regularised their production, producing exclusively the most common variants, the level of conditioning would be 0, as this majority variant would be used equally frequently across both categories. Equally, if a participant fully conditioned the variant use to a linguistic cue, the regularisation can never be full, as the frequency of production of its variant will fully mimic the frequency of the categories. Hence, here we find two competing hypotheses: participants may regularise to the most frequent variable, or they may accentuate the existing level of conditioning in their linguistic input.

In addition, we planned to test the predictions of Attention Theory (AT) regarding the symmetry of the conditioning. As stated before, AT predicts that participants will form a stronger association between the minority category and its associated variant than between the majority category and its associated variant. In practice, this would mean that participants are more likely to use the minority variant when presented with the minority cue than they are to use the majority variant when they are presented with the majority cue, whereas McGarty et al. (1993) account would predict the opposite asymmetry. Finally, linguistic approaches to conditioning would not in principle predict any asymmetry. This study allowed to test the symmetry in the conditioning, to explore which of the accounts best described the process of conditioning in the linguistic domain.

In summary, we predicted that:

1. Participants would condition to a higher degree than that in their input, whether the frequency distribution for the category is skewed or not.
2. Participants would condition to a higher degree in the skewed animacy condition than in the uniform animacy condition, or alternatively,
3. Participants would regularise their marker use more in the skewed animacy condition than in the uniform animacy condition.

   Additionally, within the skewed animacy condition, we predicted that:
4. The association between the majority marker and majority category would be stronger than that between minority category and minority marker, or alternatively,
5. The association between the minority marker and minority category would be stronger than that between majority category and majority marker.

## Methods

This study was approved by the ethics committee of the Psychology Department of the University of Warwick on the 30th of July2021 (reference PGR_20-21/15). The methods and analysis for this study were preregistered in https://osf.io/5bjkd.

**Participants**

196 participants took part in our study, from which we excluded 34, yielding a final sample of 162 participants. We recruited via Prolific Academic, together with the sample for the experiment in Chapter 4 and given the similarities on the structure and demand between both experiments, we used the same inclusion/exclusion criteria and compensation. These were the reasons for the exclusion of the 34 participants: 19 of them failed to produce an analysable answer in 25% or more of the plural trials in the production task, 17 failed to produce the right noun in more than 25% of the filler trials in the production task, 16 failed to produce more than 25% of the nouns in the free recall task of the noun testing phase, four failed more than 25% of the filler trials in the comprehension task, and two clicked the same side of the screen for more than 87.5% of the items in the comprehension task. Some of the participants failed several of the exclusion criteria. The exclusion rate was relatively high (17.34%) and did not differ by condition. The sample distribution by condition can be observed in Table 5.1, with a minimum of 40 participants per condition and counterbalancing, as stated in the pre-registration.

Table 5.1.

*Number of participants per condition, before and after exclusions*

| Animacy condition | Order condition | N before exclusions | N after exclusion | Exclusion rate (p) |
|---|---|---|---|---|
| skewed | Comprehension first | 49 | 42 | 0.14 |
| skewed | Production first | 49 | 40 | 0.18 |
| uniform | Comprehension first | 49 | 40 | 0.18 |
| uniform | Production first | 49 | 40 | 0.18 |

Regarding the demographics, 98 participants identified as female, 61 as male, two as non-binary, and one preferred not to disclose their gender. Their mean age was 34.7 (SD = 12.96, range 18 – 77). 101 of them were English monolinguals (62.35%), 31 bilinguals (19.14%), and 30 spoke three or more languages (18.52%).

**Language structure and stimuli**

We used the same language and stimuli as in Chapter 4. This was an artificial language that contained 12 nouns that resembled their English counterparts and three plural markers (“bok”, “hap”, and “nim”). Six of the nouns belonged to vehicles (“temoto” for moped, “tetaksi” for taxi, “tediga” for digger, “tekarsik” for ambulance, “tebas” for bus, and “tetrakta” for tractor) and six to animals (“dabili” for goat, “datronki” for elephant, “dapumba” for boar, “dabulkau” for bull, “dadia” for deer, and “damus” for moose). Each participant saw a subset of eight nouns (with a different proportion of vehicles and animals depending on their animacy condition) and two markers (randomly selected out of the three).

**Design**

In order to test our hypotheses in the linguistic domain, and in line with Chapter 4, we used animacy as the category (and the linguistic cue) and plural marker as the associated trait (the outcome). We used a between-participants design with skewness of the frequency of animacy category (skewed vs. uniform) as the independent variable. In the skewed animacy condition, participants learnt a language that contained six nouns from an animacy category, and two items from the other, whereas participants in the uniform animacy category learnt a language in which there were four nouns of each category. Majority/minority categories were assigned randomly for each participant, as well as the specific nouns they would learn, from the array of six nouns per category. The marker use was conditioned to the category in both conditions to the same degree: one of the markers was 75% of the times with one category, and 25% with the other, and vice versa, regardless of the skewness of the category frequency distribution. We counterbalanced the order of the two tasks we used for our outcome measures: comprehension task and production tasks. Table 5.2 shows a representation of an example of the frequency of each kind of trials per condition.

Table 5.2.

*Experimental design*

| Condition | Category distribution | Marker distribution | Out of 32 trials… |
|---|---|---|---|
| Skewed animacy | 6 nouns in Category 1 | 75% times Marker 1<br>25% times Marker 2 | 24 trials Category 1<br>- 18 with Marker 1<br>- 6 with Marker 2 |
| | 2 nouns in Category 2 | 75% times Marker 1<br>25% times Marker 2 | 8 trials Category 2<br>- 6 with Marker 1<br>- 2 with Marker 2 |
| Uniform animacy | 4 nouns in Category 1 | 75% times Marker 1<br>25% times Marker 2 | 16 trials Category 1<br>- 12 with Marker 1<br>- 4 with Marker 2 |

|  |  |  |  |
|---|---|---|---|
|  | 4 nouns in Category 2 | 75% times Marker 1 | 16 trials Category 2 |
|  |  | 25% times Marker 2 | - 12 with Marker 1 |
|  |  |  | - 4 with Marker 2 |

*Note.* This table shows the frequency distribution of the nouns of each category by condition, as well as the frequency with which each of the markers appeared with each noun of each category. Column 1 indicates the animacy condition; Column 2 indicates the number of nouns of each category that participants learnt in each condition; Column 3 indicates the frequency of each of the markers with the nouns of each category and conditions. Column 4 represents how those frequencies would translate in a block of 32 trials.

Once again, using the formula for contingency (Allan, 1980; Equation 3) that calculates the difference in probability of the one of the outcomes (here, one of the markers) for each of the cues (each of the categories), we see that the level of contingency between the category and the marker is .5 for both conditions (the probability of marker 1 of category 1, P(O|C), is .75, and .25 with category 2, P(O|¬C), with a difference of .5 between values).

$$\Delta P= P(O|C)-P(O|\neg C). \qquad (3)$$

**Tasks**

We used the same task as in the Experiment in Chapter 4. The only difference was the probability distribution of the different kinds of trials, which we covered in the Design subsection. As a brief reminder, participants learnt an artificial language called Panitok that consisted of 8 nouns and 2 possible plural markers. They were presented with the language both auditorily and in written mode. They completed five tasks. Participants were first trained in the nouns, by being presented with them together with the corresponding image and asked to repeat them. They were subsequently tested on the nouns through an AFC task, followed by a production task in which they were presented with an image and asked to say the name aloud. Their responses were recorded and transcribed.

In the third task, the plural training task, participants were presented with a sentence, both written and auditorily, containing a plural marker and a noun and were asked to choose between two images, the target one containing two items of the referent corresponding to the noun, and the foil one containing a single item (e.g. they were shown “bok tebas” and presented with an image with two buses and an image with a single table to choose between). This task contained 128 trials, and 32 filler trials that looked like the noun testing 2AFC trials. The fourth task was the plural testing task, which contained the comprehension task and the production task in a counterbalanced order. In the comprehension task, participants were shown a marker and asked to choose what would follow between two options, each showing an image and a noun from each category. This task contained 32 trials. In the production task, participants were presented with an image with two of the objects and were asked to describe them. They were expected to produce a verbal response that contained a marker and the noun corresponding to the image. There were 64 of these trials. Finally, in the last task, participants completed a few questions on their perception of the language and provided their demographic information. Please refer to Tasks subsection in Chapter 4 for a full description (p. 114).

### Procedure

We used the same procedure as in the Experiment in Chapter 4. Participants took an average of 25.49 minutes (SD = 4.6, range 16 – 41 minutes) to complete the experiment.

### Measures

We used the same measures as in the experiment in Chapter 4 (p. 118).

## Results

### Learning outcomes

Before testing our hypothesis, and as in Chapter 4, we checked that our participants were proficient in the artificial language after the noun training and plural training tasks, and crucially, that there were no significant differences between conditions that could contribute to differences between conditions in our variables of interest. As Table 5.3 shows, the proportion of correct trials was near ceiling for participants in all the training tasks and conditions. We ran one-way ANOVA test with animacy condition as predictor and the proportion of correct trials as the outcome and did not find any significant effects for any of the outcome variables, aside from a marginally significant difference in the free recall task, suggesting that participants in the skewed animacy condition might have had higher scores in this task than participants in the uniform animacy condition.

Table 5.3.

Proportion of correct trials in the training tasks by condition.[8]

[8] In the free recall task of the noun testing phase, we transcribed the participants' oral productions following the indications in Appendix B and considered as correct any production with a Levenshtein distance of 2 or less to the target production.

|  | Skewed animacy | Uniform animacy | Difference |
|---|---|---|---|
| **Task** | Mean (SD) | Mean (SD) | P - value |
| Noun testing – 2AFC | .993 (.022) | .984 (.025) | .376 |
| Noun testing- free recall | .960 (.071) | .936 (.093) | .062 |
| Plural training | .989 (.011) | .988 (.010) | .756 |

**Conditioning behaviour**

We used logistic mixed effects models for the exploration of conditioning, one per task. These aimed to look at the degree of conditioning between marker and category against the degree of conditioning in the input, regardless of whether the conditioning followed the same direction as the input or not. This allowed to test if participants were overall reducing variation by conditioning marker use to animacy category, regardless of the existing patterns in their input language, as it happened in Chapter 4, where there was no conditioning. Moreover, this allowed to compare the overall degree of conditioning in each condition with the degree of conditioning in the predicted direction, analysed later. The model for production task had “Most skewed category” (see Chapter 4 for its description, p. 121), animacy condition, task order, and their interaction as main effects, and random intercepts for participant, marker subset, noun subset, as well as by-participant random slopes for all fixed factors. The outcome variable was the marker that participants had selected, 1 if it was the “Most conditioned marker” and 0 if it was the “Least conditioned marker”. We sum coded animacy condition and task order, and we used custom coding for category. This allowed us to reflect in the model the difference in the probability of each of the markers with each of the categories. Therefore, it entered the contrast to the input conditioning rate in the hypothesis matrix and allowed us to compare participants’ degree of conditioning with the input one. As the hypothesis matrix was not centred, we used its generalised inverse as our contrast matrix. The original hypothesis matrix and the resulting contrast matrix can be seen in Table 5.4. The model did not initially converge, so following the procedure by Barr et al. (2013), we excluded the random intercept for marker subset, as it explained the least of the variance.

Table 5.4.

*Hypothesis and custom contrast matrices for the fixed factor of category in the conditioning models.*

| Hypothesis matrix | Intercept | Contrast 1 |
|---|---|---|
| Minority category | ¼ | 5/8 |
| Majority category | ¾ | -5/8 |

| Contrast matrix | Intercept | Contrast 1 |
|---|---|---|
| Minority category | 1 | 1.2 |
| Majority category | 1 | -.4 |

We followed the same strategy for the comprehension task, this time using “Most skewed marker” as a factor and the chosen category in a given trial as the outcome, coding it as 1 if it was the “Most conditioned category” and 0 if it was the “Least conditioned category”. Once again, we sum coded animacy condition and task order, and used custom contrast representing the input level of conditioning for marker.

***Comprehension task***

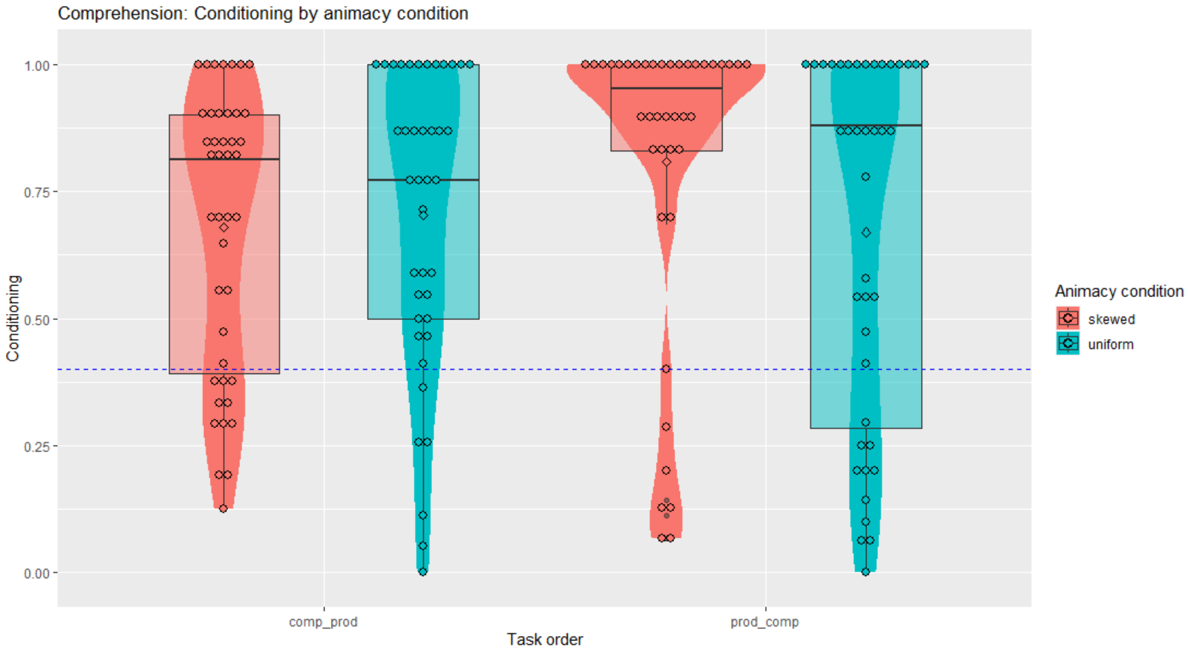
We aimed to explore whether participants conditioned their category selection to the marker more or less than in their input, and whether the skewness of the animacy condition had an impact on it. We also explored the effect of the order of the tasks, given that it had had a big effect in the experiment in Chapter 4. In line with the strategy followed in Chapter 4, and given the parallelism between the analyses in the two chapters, we also renamed our effects to improve the readability of the section.

- The effect of “Most Skewed Marker”, having used custom contrast, showed that the conditioning has higher than in the input. As such, in this section we refer to it as *evidence of change in conditioning.*
- We also refer to the interaction between “Most Skewed Marker” and animacy condition as *effect of animacy condition on conditioning.*
- We refer to the interaction between “Most Skewed Marker” and task order as *effect of task order on conditioning.*
- Finally, we refer to the interaction between “Most Skewed Marker”, animacy condition and task order as the interaction between the *effect of task order and of animacy condition on conditioning.*

As Figure 5.1 shows, we found *evidence of change in conditioning* (β = -.913, z =-51.63, p <.001), showing that participants were conditioning their category selection to the marker to a higher degree than in their input (a degree of conditioning of .4) across both conditions and task orders. We also found a significant *effect of animacy condition on conditioning*, (β =.101, z =5.69, p <.001), showing that, as predicted, participants in the skewed animacy condition showed a higher degree of conditioning. In addition, we found a *effect of task order on conditioning* (β =-.125, z =-7.05, p <.001), showing that participants who completed the comprehension task second showed a higher degree of conditioning than those who completed it first. Finally, the *interaction between the effect of task order and of animacy condition on conditioning* was not significant (β =.003, z =.19, p = .849).

Figure 5.1.

*Conditioning in the comprehension task by animacy condition and task order*



*Note.* The blue dashed line represented the degree of conditioning in the input.

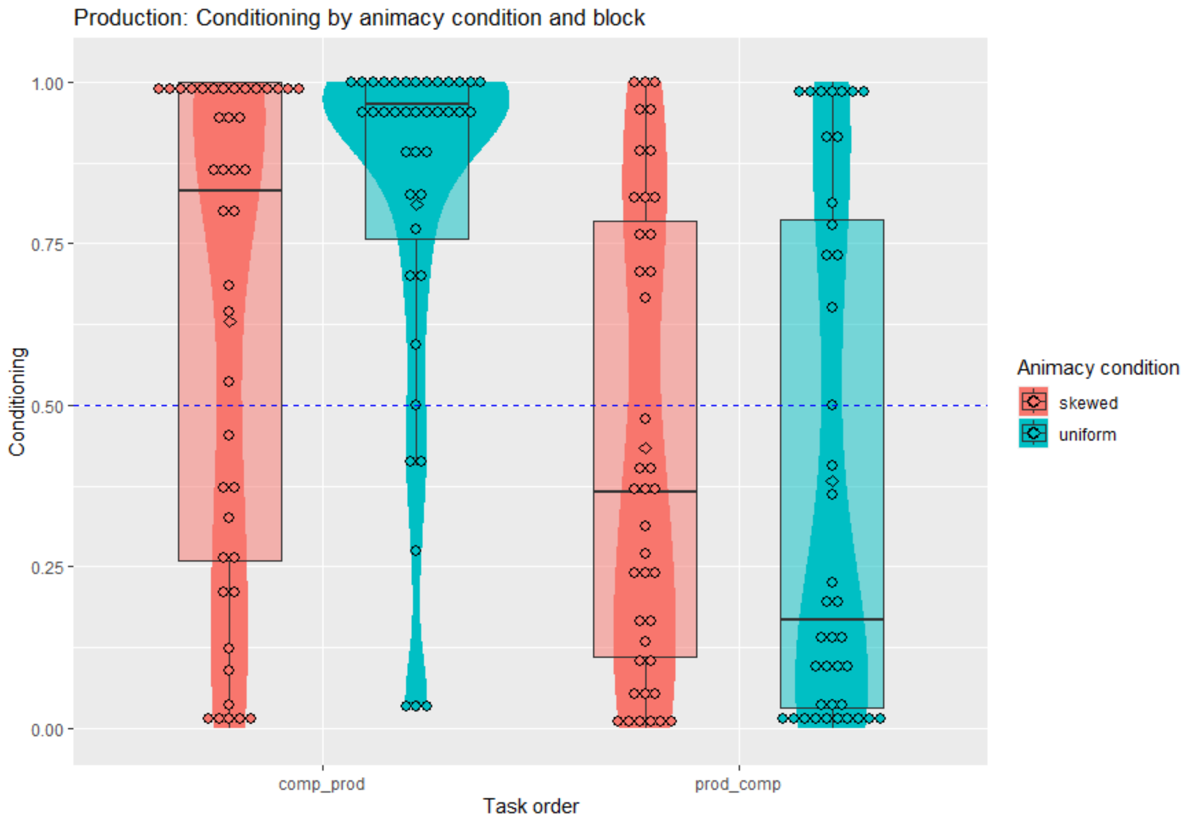
### *Production task*

As in the comprehension task, here we refer to the effect of “Most Skewed Category” as *evidence of change in conditioning,* to the interaction between “Most Skewed Category” and animacy condition as *effect of animacy condition on conditioning*, and to the interaction between task order and “Most Skewed category” as *effect of task order on conditioning.* W found a significant *evidence of change in conditioning*, showing that participants, overall, showed a higher degree of conditioning than that in their input (β = -2.186, z = 48.73, p <.001). Similarly, the *effect of task order on conditioning* was significant (β =-.705, z =-15.66, p <.001), showing once again a higher degree of conditioning when production came second.

More importantly, as Figure 5.2 shows, the *effect on animacy condition on conditioning* was not significant (β =.056, z =1.22, p =.222). However, we found an interaction between the *effects of task order and of animacy condition on conditioning* (β =.236, z =5.28, p < .001). We ran nested analyses of each of the orders to better understand the nature of the triple interaction. For both task orders, we found *evidence of change in conditioning*. When the production task came first, the degree of conditioning was lower than in the input (F(1, 9850) = 760.977, p<.001), whereas when production came second (F(1, 9850) = 1565.998, p<.001), the degree of conditioning was higher than in the input. Also, contrary to the general collapsed analysis, a significant *effect of animacy condition on conditioning* was present within both task orders, but these followed the opposite patterns in each of the tasks. When the production task came first, participants in the skewed condition conditioned more than their counterparts in the uniform condition (F(1, 9850) = 16.220, p<.001), whilst when it came second the opposite was true: participants in the uniform animacy condition showed a higher degree of conditioning than participants in the skewed animacy condition (F(1, 9850) = 11.239, p<.001). Table 4.5 shows a summary of the results for this section.

Figure 5.2.

*Degree of conditioning by animacy condition and task order in the production task.*



*Note.* The blue dotted line indicates the degree of conditioning in the input.

Table 5.5. Summary of results on conditioning for Chapter 5

| | | | By Task order | |
|---|---|---|---|---|
| Effect | Task | Across task orders | Comp-prod | Prod-comp |
| Presence of conditioning behaviour | Comprehension | Conditioning higher than input | Conditioning lower in this task order, but still higher than input | Conditioning higher in this task order |
| | Production | Conditioning present – difference with input dependent on task order | Conditioning higher than input in this task order | Conditioning lower than input in this task order |
| Effect of animacy condition | Comprehension | Predicted direction –conditioning in skewed > uniform animacy condition | No effect of task order | |
| | Production | Dependent on task order | Opposite direction –conditioning in uniform > skewed animacy condition | Predicted direction –conditioning in skewed > uniform animacy condition |

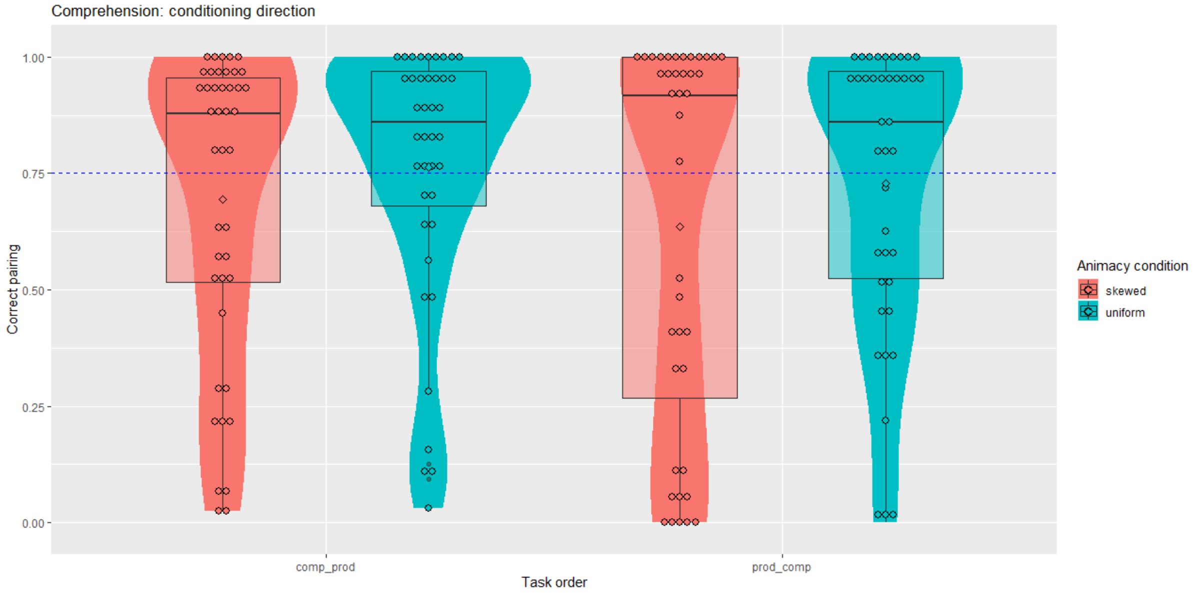
After testing whether participants had conditioned their marker use to animacy category, we wanted to test whether this conditioning followed the patterns that they had encountered in their input. As described before, for each category, and in both conditions, there was a marker that was used more frequently than the other. We labelled those trials in which participants associated a marker to the category it was most frequent with as "correct pairings" and tested whether these differed by animacy condition, as predicted by Sherman et al. (2009)'s account.

### Comprehension task

As Figure 5.3 shows, the direction in which participants conditioned their choice selection was the same one as in the input, but to a higher degree, β = -1.117, z = -28.242, p <.001. As in the previous analyses, we found an effect of animacy condition in the predicted direction, β = .248, z = 6.321, p <.001, and an effect of task order showing that the degree in which participants conditioned in the predicted direction was higher when the comprehension task was first, β = -.142, z = -3.613, p<.001. Once again, we found no interaction between animacy condition, task order, and marker, β = - .023, z =-.585, p = .558.

Figure 5.3.

*Proportion of correct pairings in the comprehension task by animacy condition and task order.*



### Production task

As predicted, and as Figure 5.4 shows, the conditioning behaviour that we found in both conditions was in the same direction as the one in the input, but to a higher degree, β = -1.301, z = 29.804, p <.001. In addition, participants who completed the production task second conditioned in the predicted direction to a higher degree, β =-.375, z =-10.821, p <.001. The interaction between animacy condition and category was not significant, β =.051, z =1.18, p =.236, However, once again, we found a marginally significant triple interaction between category, animacy condition, and task order, β =.067, z =1.925, p = .054. We ran nested analyses of each of the orders to better understand the nature of the triple interaction. When the production task came first, the degree of conditioning was lower than in the input, F(1, 9850) =290.813, p<.001, whereas when production came second, F(1, 9850) = 859.238, p<.001, the degree of conditioning was higher than in the input. Also, contrary to the general collapsed analysis, animacy condition had a significant effect within one of the task orders. When the production task came first there was no difference between animacy condition, F(1, 9850) = 0.085, p=.770, whilst when it came second participants in the uniform animacy condition showed a higher degree of conditioning in the same direction as the input than participants in the skewed animacy condition, , F(1, 9850) = 4.379, p=.036). Table 5.6 shows the summary of the results for this section.

Figure 5.4.

*Proportion of correct pairings in the production task by animacy condition and task order.*

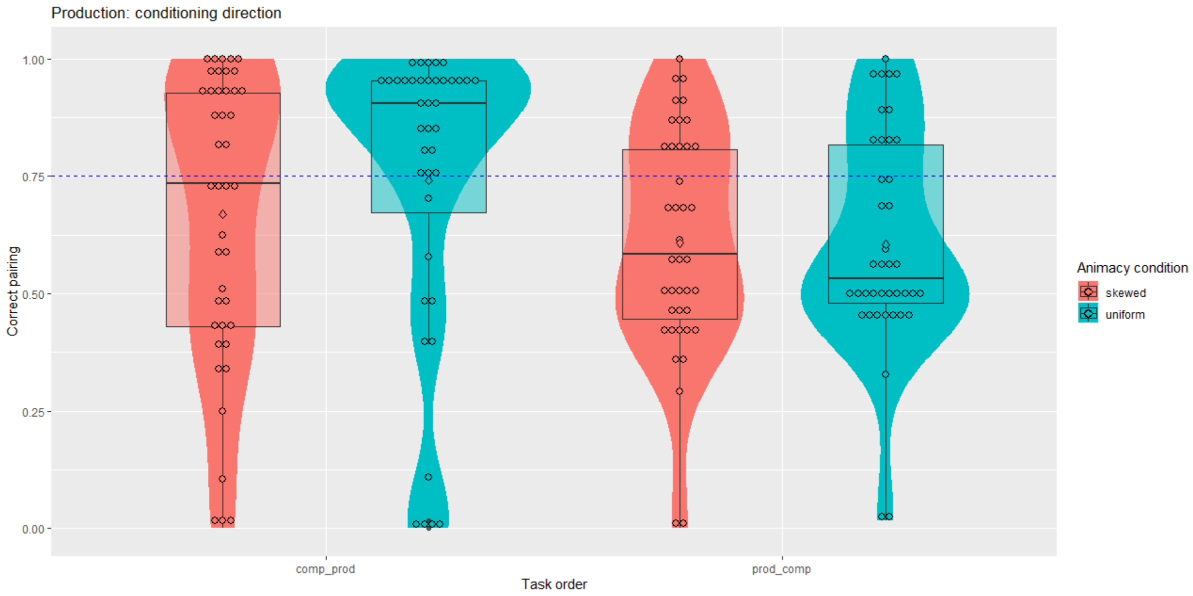


Table 5.6. Summary of results on direction of conditioning for Chapter 5

| | | | By Task order | |
|---|---|---|---|---|
| Effect | Task | Across task orders | Comp-prod | Prod-comp |
| Presence of directed conditioning behaviour | Comprehension | Directed conditioning higher than input | Conditioning lower in this task order, but still higher than input | Conditioning higher in this task order |
| | Production | Conditioning present – difference with input dependent on task order | Conditioning higher than input in this task order | Conditioning lower than input in this task order |
| Effect of animacy condition | Comprehension | Predicted direction –conditioning in skewed > uniform animacy condition | No effect of task order | |
| | Production | Dependent on task order | Opposite direction –conditioning in uniform > skewed animacy condition | No effect of animacy condition within this task order |
| Analysis of symmetry | Evidence for inverse rate effect | | | |

**Conditioning symmetry in the skewed animacy condition**

We aimed to test whether, within the skewed animacy condition, conditioning was symmetrical for minority and majority categories. This would allow us to test the contrasting predictions of different domain-general theories for conditioning behaviour, such as the Attention Theory (Sherman et al., 2009) and the self-categorisation theory (McGarty et al. 1993), which predicted asymmetry in opposite directions, with linguistic accounts of conditioning behaviour, which in principle would not predict any asymmetry. In the comprehension task, if participants had attended to the base rate of the categories and matched the probabilities in their input, we would have expected that, when presented with the majority marker they would choose the majority category 90% of the time, whereas when presented with the minority marker, they would choose the majority and minority categories 50% of the time each (consult the Design section for a visual representation of the probabilities). To test whether they deviated from their input, we ran a linear mixed effects model with marker (majority or minority), task order, and their interaction as the fixed effects, a random intercept for participant and by-participant random slopes for each of the factors. We sum coded both fixed factors and used correct pairings as the outcome variable. This represented the proportion of trials with each marker in which participants had selected the category that was most frequent with the marker. That is, for the majority marker, the correct pairing would represent the proportion of trials in which participants had selected the majority category and vice versa. We did not find a significant effect of category (β = -.005, z =0.557 p = .577), nor task order (β = .030, z =.755, p =.450), or an interaction between task order and category (β = -.007, z = -.766, p = .444). Table 5.7 shows the mean proportions of correct pairing by marker and task order.

Table 5.7.

*Mean proportion of correct pairings in the comprehension task per marker and task order*

| | Comprehension first | | Production first | |
|---|---|---|---|---|
| | N | Mean (SD) | N | Mean (SD) |
| Majority marker | 42 | .702 (.322) | 40 | .635 (.415) |
| Minority marker | 42 | .687 (.346) | 40 | .638 (.407) |

If participants had probability-matched the base rate, we would have expected a significant difference between markers. We ran a series of two-tailed one-sample t-tests to test participants' answers against the input proportion of correct pairings for each of the markers. Since we were running four different tests for the same hypothesis, we applied Bonferroni corrections and established the threshold for significance in .0125. We tested the number of correct pairings for the majority marker against .9 and found a significant difference (t(81) = -5.65, p <.001, d = -.624). Similarly, we tested the correct number of pairings for the minority marker against .5 and found a significant difference (t(81) = 3.92, p <.001, d = .433). Finally, we tested both against .75, which was the frequency with which each of the categories was paired with each of the markers in the input, regardless of the higher base rate of each of the categories, and we found a marginally significant difference for both markers (t(81) = -2.11, p =.038, d = -.233 for the minority marker and t(81) = -1.97, p =.052, d = -.218 for the majority marker). In summary, participants' proportions of correct pairings did not differ by marker nor by task order, and significantly differed from the input proportions bearing in mind the base rate, being closer to the estimations if the base rate was not accounted for.

In the production task, if participants matched the probability in their input, the number of correct pairings should be of .75 for both conditions, and hence, any significant difference between condition would reflect asymmetry. We ran a linear mixed effect model with category (majority vs. minority), task order, and their interaction as fixed effects, both sum coded, a random intercept for participant, and by-participant random slopes for both factors. Our outcome variable was the proportion of correct pairings. We did not find a significant effect of category (β = .008, z =.344, p = .731), task order (β =.046, z =1.530, p =.126), or an interaction between task order and category (β = .012, z = -.518, p = .605). As with the comprehension task, we tested the proportion of correct pairings in each category against .75, which was the input value, using a two-tailed t-test and correcting the threshold for significance to .025. We found a marginally significant difference for the minority category (t(81) = -2.10, p = .039, d = -.232), and no difference for the majority category (t(81) = -1.60, p = .113, d = .177). Table 5.8 shows the mean proportions of correct pairing by category and task order.

Table 5.8.

*Mean proportion of correct pairings in the production task by category and task order*

| | Comprehension first | | Production first | |
|---|---|---|---|---|
| | N | Mean (SD) | N | Mean (SD) |
| Majority category | 42 | .727 (.333) | 40 | .611 (.349) |
| Minority category | 42 | .719 (.277) | 40 | .652 (.345) |

**Regularity and regularisation behaviour**

We aimed to test whether participants in the skewed animacy condition were more regular in their marker and category choices more than those in the skewed animacy condition, aside from what direction they regularised to. For that, we ran two models, one for the comprehension task and one for the production task.

The model for the comprehension task had "Majority category" as the outcome. This was the most frequently selected category for each participant across both markers, regardless of which was the most frequent category in the input, and how much it was conditioned to the marker. If both categories were selected the same number of times, this was assigned randomly. If a participant had clicked on the image corresponding to the "Majority category", this was coded as 1, whereas if they had clicked in the "Minority category" it was coded as 0. The model included animacy condition, task order and their interaction as fixed effects, with random intercepts for participant, noun set, marker set, target marker in the trial, and position of the target image in the screen, as well as by-participant slopes for both fixed factors. Both fixed factors were sum coded.

The model for the production task was identical but had “Majority marker” as the outcome instead, which was defined as whether participants, in a given trial, had selected the marker that they selected most often across the task, or its alternative. The structure of the fixed factors and random effects was also very similar, including animacy condition, task order and their interaction as fixed effects, and random intercepts for participant, noun set, marker set, and target image in the trial, as well as by-participant slopes for both fixed factors. Both fixed factors were sum coded.

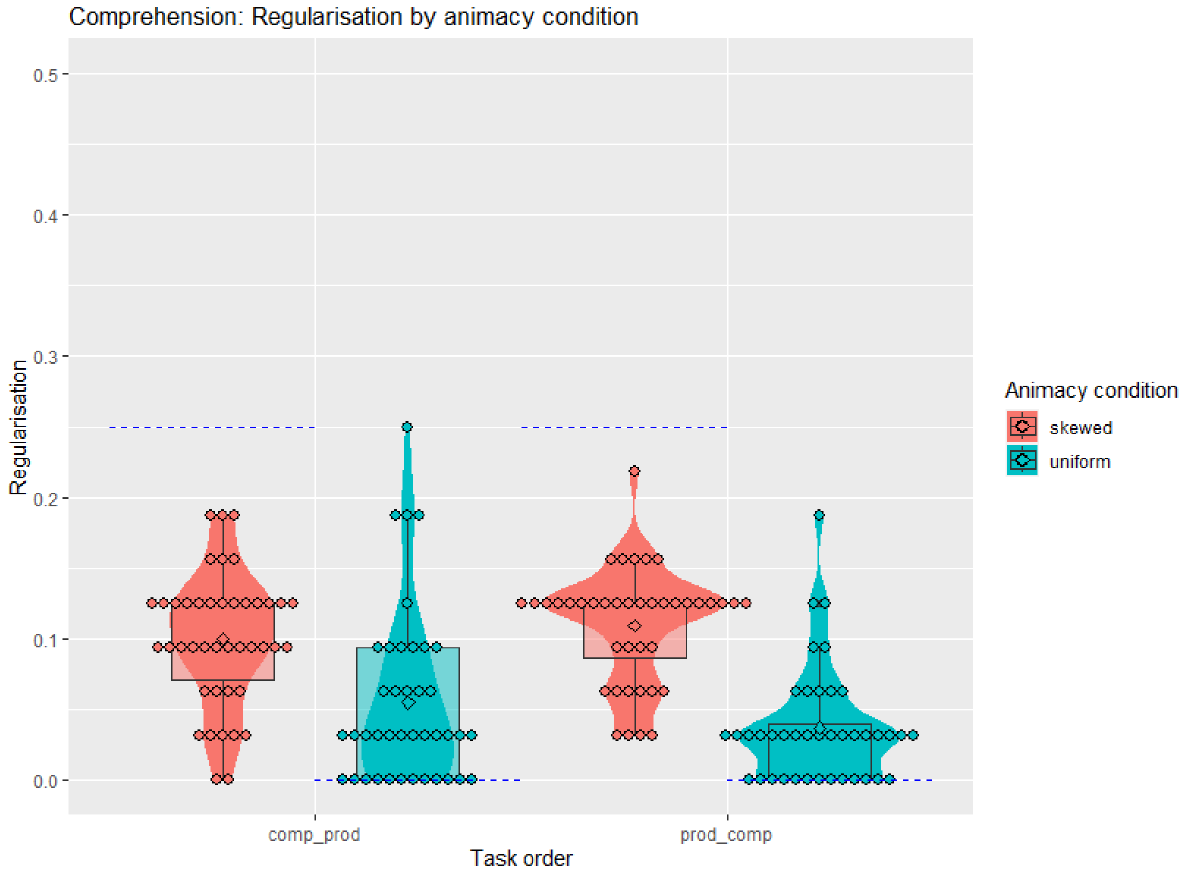
***Comprehension task***

First, we found the intercept to be significant ($\beta$ = .206, z =2.068 , p = .038), showing that participants level of regularisation, across conditions, was different from 0. We found a main effect of animacy condition ($\beta$ = .179, z = 4.040, p <.001), with no effect of task order ($\beta$ = .014, z = .307, p = .759), nor any interaction between task order and animacy condition ($\beta$ =.007, z = .160, z = .873). As Figure 5.5 shows, across both task orders, participants in the skewed animacy condition regularised more than those in the uniform animacy condition. Note however, that none of the participants fully regularised their category selection.

This test compared how regular participants category choice was in the comprehension task compared to 0 and whether there was a difference in regularisation level by condition. However, the regularity in the input of participants in the skewed animacy condition and uniform animacy condition was different. Participants in the skewed animation condition were presented with the majority category in 75% of the trials and with the minority category in the remaining 25%, whereas participants in the uniform animacy condition were presented with both categories with the same frequency. Hence, we run additional analyses testing whether the level of regularisation in each of the conditions differed from their input. This showed a different picture. We ran two one-sample t-tests comparing participants’ average choice of the majority marker to the proportion they were presented with in the input: one per animacy condition. We used the proportion of trials in which they chose the majority category as the output variable and the input proportions (.75 for the skewed animacy condition and .5 for the uniform animacy condition) as the test values. We found that participants in the skewed animacy condition selected the majority category significantly less often than in their input (t(81) = -17.743, p <.001, 95%CI [.522-.568]), whereas participants in the uniform category condition produced one of the categories more often than in their input, (t(79) =2.7689, p =.007, 95%CI [.501-.536]).

Figure 5.5.

*Regularisation degree in the comprehension task by animacy condition and task order.*



*Note.* The y-axis represents regularisation. This index is calculated by subtracting .5 from the frequency in which participants selected the "Output majority category" (see Indices for the description). Therefore, it ranges from 0 to .5, 0 representing participants selected both choices with an equal frequency (.5 is subtracted from .5), and .5 representing those participants who selected exclusively one of the categories in all the trials (.5 subtracted from 1). The blue lines represent the input level of regularisation: .25 for the skewed animacy conditions and 0 for the uniform animacy conditions.

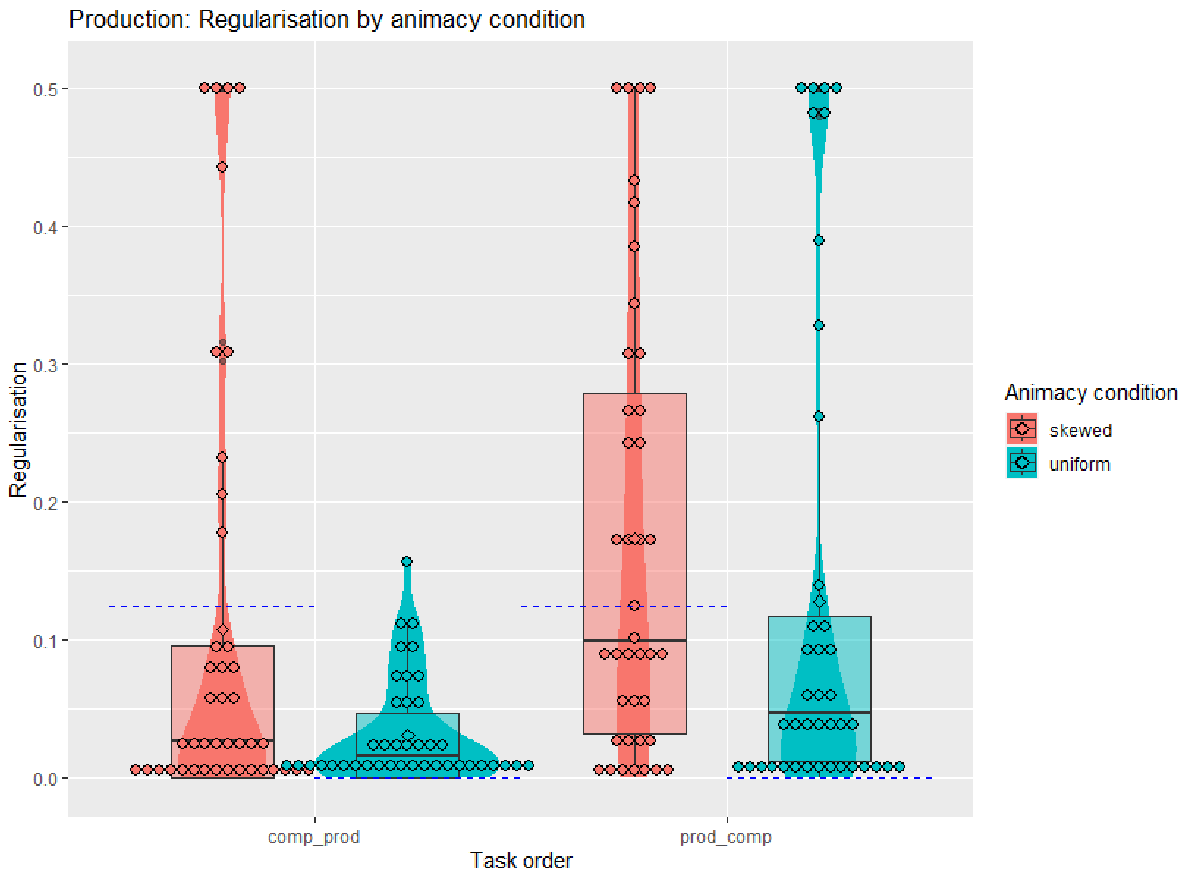
***Production task***

Once again, we found the intercept to be significant (β = .811, z =5.798 , p <.001), showing that participants level of regularisation, across conditions, was different from 0. We also found a main effect of animacy condition (β = .372, z = 4.936, p <.001), as well as a significant effect of task order (β = -.156, z = -2.268, p = .023), and a marginally significant interaction between task order and animacy condition (β = .127, z = 1.849, z = .065). A post hoc nested analysis by task order showed that the effect of animacy condition was present in both task orders (F(1,9853) = 24.447, p <.001 for the comprehension first order, and F(1,9853) = 5.642, p =.017 for the production first order). As Figure 5.6 shows, participants in the skewed animacy condition regularised more than those in the uniform animacy condition, across both task orders, but regularisation was overall higher when the production task was completed first.

As with the comprehension task, this test compared how regular participants marker use was in the production task, disregarding the regularity in the input for different conditions. Participants in the skewed animation condition were presented with the majority marker in 62.5% of the trials, whereas participants in the uniform animacy condition were presented with both markers with the same frequency. We ran two one-sample t-tests comparing participants' average use of the majority marker to the proportion they were presented with in the input: one per animacy condition. We used the proportion of trials in which they used the majority marker as the output variable and the input proportions (.625 for the skewed animacy condition and .5 for the uniform animacy condition) as the test values. We did not find a significant difference between the proportion of trials in which they used the majority marker and that in their input, (t(81) = -1.479, p=.143, 95%CI [.328-.639] for the skewed animacy condition, and t(79) =-1.857 p =.067, 95%CI [.433-.502] for the uniform animacy condition). Finally, we examined the number of full regularisers, which following Hudson Kam (2019) we defined as those participants who had used one of the markers in all or all but one of the trials. Even if the number of full regularisers was higher in the production first task order (10 out of 80) than in the comprehension first task order (4 out of 82), the difference in proportion was only marginally significant ($\chi^2(1) = 2.980$, p =.084). Table 5.9 presents a summary of the results for this section.

Figure 5.6.

*Regularisation degree in the production task by animacy condition and task order.*



*Note.* The y-axis represents regularisation and ranges from 0 to .5, 0 representing participants selected both choices with an equal frequency, and .5 representing participants selecting exclusively one of the categories. The blue dashed lines represent regularisation level in the input: .125 for the skewed animacy conditions and 0 for the uniform animacy conditions.

Table 5.9. Summary of results for regularisation Chapter 5

| | | | By Task order | |
|---|---|---|---|---|
| Effect | Task | Across task orders | Comp-prod | Prod-comp |
| Regularity | Comprehension | Higher than 0 | No effect of task order | |
| | Production | Higher than 0 | Regularisation lower in this task order | Regularisation higher in this task order |
| Effect of animacy condition on regularity | Comprehension | Regularisation higher in skewed > uniform animacy condition | No effect of task order | |
| | Production | Regularisation higher in skewed > | No effect of task order | |

| | | | |
|---|---|---|---|
| | | uniform animacy condition | |
| Effect of animacy condition on regularisation | Comprehension | Lower than input in skewed animacy condition<br>Higher than input in uniform animacy condition | No effect of task order |
| | Production | Same as in input | No effect of task order |

## Discussion

In this study, we aimed to test whether the skewness in a linguistic cue affected participants' conditioning behaviour when a probabilistic association between said linguistic cue and a linguistic element existed. For that, we trained participants in an artificial language in which the frequency of a plural marker vs. the alternative depended on the semantic category (animate or inanimate) of the noun and manipulated the frequency of the nouns belonging to each category (uniform vs. skewed). After training, all participants completed two main tasks: a comprehension task, in which they were shown a marker and asked to predict which noun would follow between two options, one from each category, and a production ask, in which they were shown an image of two items and asked to produce the corresponding description.

### Conditioning behaviour

We predicted that, a) participants in the skewed animacy condition would show a higher degree of conditioning than those in the uniform animacy condition, and b) that participants across conditions and task orders would show a higher degree of conditioning than that in their input. We found the predicted effect in the comprehension task: when the distribution of animacy category was skewed participants conditioned the marker to the category to a higher degree than when the distribution was uniform. This held true regardless of whether they completed the task first or second. The direction of the conditioning was the same to that provided in the input, showing an accentuation of existing differences. Also, participants, across both conditions and task orders, conditioned more than in their input. These results are overall in line with the predictions from the domain-general accounts of category accentuation (Sherman et al., 2009; Tajfel & Wilkes, 1963): participants conditioned more than in their input, and conditioning was higher in the skewed animacy condition than in the uniform animacy condition. When we tested participants by offering them a marker and asking them to predict the semantic category of the corresponding noun, they tended to exaggerate the existing association between marker and category.

In the production task, however, we did not find the predicted effects. When production came first, participants overall conditioned to a lower degree than in their input. When looking at conditioning, regardless of the direction, we found the predicted effect: participants were more likely to condition in the skewed animacy condition. However, the effect was not present when we exclusively looked at conditioning in the input direction. When production came second, participants conditioned to a higher degree overall than in their input, as they had done in the comprehension task. Nevertheless, we found the opposite effect to the predicted one: participants conditioned to a higher degree in the uniform animacy condition than in the skewed animacy condition, both when we took into account the direction of the conditioning and when we did not.

The reversed effect that we find in the production task when it comes first was not predicted by any of the theories that we covered. Furthermore, given the high level of conditioning in both animacy conditions, we cannot argue as we did in Chapter 4, that participants may not have perceived the different categories. We could argue, nonetheless, that participants in the uniform animacy condition, given that there was not a majority marker, reduced the uncertainty and increased the efficiency of the language by conditioning it, whereas some of the participants in the skewed animacy condition could have been more likely to reduce it by regularising to the most frequent marker, which, as described in the introduction, is not compatible with conditioning. We will come back to this point when we review the regularisation results, as well as to the results on the low level of conditioning within the production task.

It is also important to note that, when testing participants using the comprehension task, which is the one most similar to the ones used in domain-general cognitive research, we find results closer to these theories than when testing participants with a linguistic production task. In addition, as discussed in Chapter 4, this task is the one that is most related to the perception of the input in its framing, as participants completing the comprehension task are asked to predict the productions of the *native speaker* that taught them the language.

### Regularisation behaviour

We hypothesised that participants in the skewed animacy conditions would regularise more than those in the uniform animacy condition. We found evidence for this in both tasks and task orders. In addition, we found that participants regularised more in both conditions of the production task when it came first, and that the number of full regularisers was marginally higher in the production first task order.

These results are consistent with the ones we find for the conditioning behaviour, and those we find in Chapter 4: Participants completing the production task first may not have remembered the name of both markers, particularly when they were exposed to one of them in a lower frequency than the alternative. However, even if we excluded full regularisers from the sample, regularisation levels were higher in this task when it came first. These results are consistent with and could be explained by Hudson Kam and Chang's (2009) and Hudson Kam's (2019) findings, which found that an increase in retrieval load led to regularisation. This was the condition in which participants had a higher retrieval load, relative to the comprehension task, and a lower exposure to the language, relative to those who completed the production task second.

Nevertheless, when looking at the change in regularity in relation to their input, we see that in the comprehension task, participants in the skewed animacy condition regularised less than in their input, whereas participants in the uniform animacy condition regularised more than in their input, whereas in the production task, participants from both conditions and in both task orders probability-matched their marker selection to their input. In combination with the results from conditioning behaviour, we can speculate that, in the comprehension task, participants in the skewed marker conditions regularised less than in their input because they were conditioning marker use to category, and those behaviours would not have been compatible. Participants in the uniform animacy condition were also showing more conditioning than in their input in the comprehension task but given that the level of regularisation in their input was 0, any imbalance on the frequency with which they chose different categories could result on a slight increase in regularity. Indeed 53 out 80 participants in the uniform animacy condition chose both categories either at an equal rate, or with a difference of one trial.

In the production task, however, the picture is slightly different. We found that participants in both conditions and task orders, on average, accurately match the probability that they find in their input, as it is often the case in studies in unpredictable variation (e.g. Feher et al., 2016; Hudson Kam, 2019), even if regularisation was higher in the production task first order, as discussed earlier. Conditioning, in contrast, was higher than in the input in the comprehension first order, and lower than in the input in the production first order, with effects in opposite directions in the effect of animacy condition. This comes to show how probability-matching behaviour is not always a reflection of an accurate perception of the statistical properties of the language, as it has been suggested in the past (e.g., Hudson Kam, 2019; Samara et al., 2017; Smith et al., 2017), but that it can hide important underlying patterns of biased behaviour.

See, for example, participants in the uniform animacy condition in this production task. If we look at their regularisation behaviour across both task orders, we can see that it is significantly higher in the production first task order than in the comprehension first task order, but that it does not significantly differ from 0 in any of the cases (see Figure 5.2). Based on this alone we could speculate that the slight difference in task order is fully due to the higher number of full regularisers in the production first task order. However, their behaviour in terms of conditioning is radically different. Participants in the comprehension first task order are nearly at ceiling in conditioning, and therefore, their almost equal number of uses of each of the markers was a result of a highly structured production that amplified the conditioning they found in their input, in accordance with the category accentuation hypothesis. If their regularisation was near 0 this is because the number of items of each category was equal (see Figure 5.1). In contrast, those participants in the production first task order show a lower level of conditioning than that they found in the input, with their marker selection based on the category close to being random. In this case, their probability-matching behaviour was explained by a nearly random selection of markers for each category. In this case, participants seem to have perceived the proportion of use of the markers in their input, but not its relationship to category. As discussed in Chapter 4, this could be due to participants in the production first task order not having perceived animacy categories as such when they were completing the production task, as the comprehension task increased the salience of these categories. In summary, the regularisation results show that participants partially replicated the frequency patterns in their input, but they emphasise the importance of a careful analysis of the regularisation behaviour and its underlying patterns.

### Symmetry in the skewed animacy condition

Another of our aims was to explore whether conditioning in the skewed animacy condition was symmetrical across categories. According to the distinctiveness-based theory (McGarty et al., 1993), participants in the skewed animacy condition would first form a strong association between the majority category and its associated marker, and then, associate the minority category with the minority marker to distinguish it. Hence, we would have predicted participants to use the majority marker with the majority category more than they use the minority marker with the minority category. In contrast, the Attention Theory (Sherman et al., 2009; Medin & Hassle, 1988) predicted that participants would first learn about the majority category and its associated marker and assume them as the norm, to then turn their attention to the minority category and identify its predictor. According to this account, participants would form a stronger bond between the minority category and its associated marker than between the majority marker and its associated marker. The linguistic approaches, such as the communicative efficiency account (Fedzechkina & Jaeger, 2020; Gibson et al., 2019), do not make obvious predictions about the symmetry of this relationship. In any case, we could infer that, given the higher frequency of the majority marker, its association with a marker would have a stronger impact on communicative efficiency than the association between the minority category and its marker.

Our results were mixed. On the one hand, we did not find differences in the proportion of correct pairings between the majority and minority category in any of the tasks. However, based on the statistical distribution of markers and categories in the input language, this could be interpreted differently for the different tasks.

In the production task, the frequency of each category could not affect the results, as participants were provided with an image belonging to a category and asked to produce the corresponding marker, and the frequency with which each category appeared mimicked that in the input language. Therefore, the lack of differences in this task would go against the predictions of the Attention Theory and also the distinctiveness-based theory (McGarty et al., 1993) which would also predict an asymmetry, but in the opposite direction.

In the comprehension task, nonetheless, the skewness in category would affect participants' results should they account for the difference in base rate, as participants were asked to predict a category from a marker. The *inverse rate effect* is the phenomenon the Attention Theory was originally designed to explain (Medin & Edelson, 1988, see Don et al. 2021 for a review). In their original study, Medin and Edelson (1988) used a contingency task, similar to the allergy task described in Chapter 4 (Van Hamme & Wasserman, 1994), in which they presented participants, sequentially, with descriptions of the symptoms of individuals that had been diagnosed with two different diseases. One of the diseases was more frequent than the other. Two symptoms were listed for each patient, one of which was a reliable descriptor of the disease while the other one was common to both diseases. For example, patients with disease A, the frequent one, would be described with symptoms A and B, and patients with the disease B, the infrequent one, would be described with symptoms A and C. When participants in these studies were asked to predict what disease a patient with symptom A, the common one, would be more likely to belong to, they would be more likely to classify them as having the most frequent disease, which is consistent with the base rate, that is, there are overall more patients of the majority group. However, more interestingly, when they were asked to state which disease a patient described with symptoms B and C, each of which was a predictor of a disease, would be most likely to be diagnosed with, participants tended to assign the patient to the infrequent disease, ignoring that, overall, the disease was rare. This phenomenon has also been explored in the context of stereotype formation (Sherman et al., 2009) amongst other fields, but to our knowledge, never before in the context of linguistic biases.

The results in the comprehension task in our study go in line with the predictions of the *inverse rate effect.* Even if the proportion of each of the markers associated with each category was the same for both categories, that is, each associated marker was presented 75% of the time with each of the categories, if participants bore in mind that one category was naturally more frequent than the other, we would have predicted participants to predict the majority category 90% of the time with the majority marker and 50% of the time with the minority marker (refer back to the design section, Table 5.2). Our results, however, showed that participants' underpredicted the majority category with the majority marker and overpredicted the minority category with the minority marker, neglecting the base rate at which the categories appeared in their input.

In summary, once again, the results of the comprehension task match those predicted by the domain-general theories, whereas those in the linguistic production theories did not. As briefly discussed in Chapter 4, there were fundamental differences between these two tasks. One of them refers to their conceptual framing: in the comprehension task we were asking participants to predict the production of another individual, whereas in the production task we asked them to produce descriptions themselves. The first process, prediction, should then be a more accurate reflection of their perception of the statistical properties of the language, whereas the second, production, involves additional processes, such as the cognitive load of retrieving the words in the artificial language (Hudson Kam, 2019), the pragmatic assumptions regarding the nature of the language and the task (Perfors, 2016), or the accounting of communicative efficiency (Gibson et al., 2019). Given that most tasks in the domain-general field exploring these phenomena focused on the perceptions of the statistical properties that participants reported, and that linguistic production entails many other processes, it is not surprising after all that we find that the task for which the results best match those in the domain-general field is that which is methodologically and conceptually most similar. Behaviour in the comprehension and production task is not fully separated, however. In Chapter 6 we will further discuss the relationship between behaviour in these tasks and explicitly reported perceptions of the language.

**Limitations & Conclusion**

Some of the limitations raised in Chapter 4 also apply to this study, such as, for example, that the way in which we recalled linguistic prediction and production differed in format and in level of cognitive demand, the prediction task being presented as a 2AFC question and the production task as a free recall task. This limitation was particularly important in the context of this study, as the characteristics of our contingency table meant that the predictions from category to marker and vice versa were not symmetrical. While this allowed us to compare the results from a task most similar to those found in the domain-general field (the comprehension task) and a task typical of the field of linguistic evolution (the production task), permitting us to test different theoretical predictions, it made it hard to detangle the effects of the cognitive demand of the task from those from the statistical properties of the language.

Our results in the comprehension task supported our hypotheses, but that was not always the case in the production ask. The comprehension task was the closest to those used in cognitive research, but was this the case because these effects only appear in prediction, and not in production, or because they only appear when the cognitive demand of the task is low, as suggested by Hudson Kam & Chang (2009)? Future research could address this issue by separately manipulating task demand (low vs. high) and prediction vs. production.

In any case, this research showed that classical cognitive theories can be tested in the context of language learning and learning use. Even if the drive for communicative efficiency is a feature unique to language and have been widely shown to predict the structures that we observe in existing natural languages (Gibson et al., 2019) incorporating the learnings from cognitive psychology into the field of psycholinguistics can open new windows for the understanding of the processes behind language change and make predictions in situations in which there are opposing ways to increase learnability and reduce uncertainty (i.e., regularisation or conditioning), such as the one tested in this study, improving the existing models. Statistical processing has already been widely shown to affect language processing and perception (Aslin et al., 1998; Gerken, 2005; Gomez & Gerken, 1999; Reeder et al., 2013; Saffran et al., 1996; Wonnacott et al., 2008; see Romberg & Saffran, 2010, for a review), so looking at how its known biases could also affect the processing of language and its change, interacting with other forces, is vital if we want to fully understood how languages evolve and come to be the way they are. Finally, this study adds evidence to the claim that regularisation behaviour on its own only shows a partial picture of participants behaviour (Hudson Kam, 2019; Samara et al., 2017, Smith et al., 2017, among others). As we saw in this study, the same regularisation behaviour could hide diametrically different patterns.

One of the issues we did not cover and is usually tested in the domain-general field is whether participants are aware of their behaviour. In Chapter 6 we will examine these results together with participants' reported perception of category and of conditioning.

# Chapter 6: Implicit and explicit perception of statistical linguistic properties

In Chapters 4 and 5 we have covered the main theories for the illusion of causality, illusory correlation, and category accentuation, and we explored how these processes could explain the observations in language acquisition and evolution. As described in previous chapters, these phenomena have been attributed different origins: some authors claim that the biases we observe come from the use of heuristics to process statistical data (Allan & Jenkins, 1980, 1983; Alloy & Abrahamson, 1979; Blanco et al., 2013, 2018; Chow et al., 2019; Fernbach & Van Boven, 2021; Fiedler, 1991; Hamilton & Gifford, 1976; Matute et al., 2015; Msetfi et al., 2005; Sherman et al., 2009; Smith & Alpert, 2007; Tajfel & Wilkes, 1963; Van Dessel et al., 2021; McGarty & Penny, 1988) while others claim that people's perception of statistical data is not biased, and that what we call a bias is actually a rational correction of the observed statistical trends, based on the assumption that the stimuli come from a biased sample (Bott et al., 2021; Costello et al., 2019).

Both approaches, however, share an important assumption: that illusion of causality, illusory correlation, and category accentuation are perceptual phenomena, that is, they are about how participants perceive the statistical relationship between different elements to be. Accordingly, they measure it through various explicit means, whether that be asking participants to report their reported causal relationship between a cue and an outcome, like in the illusion of causality literature (Allan & Jenkins, 1980, 1983; Alloy & Abrahamson, 1979; Blanco et al., 2013, 2018; Chow et al., 2019; Matute et al., 2015; Msetfi et al., 2005), or asking participants to assess different groups for different traits, like in the illusory correlation literature (Fiedler, 1991; Hamilton & Gifford, 1976) as well as in the category accentuation literature (Fernbach & Van Boven, 2021; McGarty & Penny, 1988; Sherman et al., 2009; Tajfel & Wilkes, 1963). In all of these studies, participants are asked to report what they have perceived.

Some studies have used more implicit measurements of participants' perceptions. One of these measures is the classification task. In it, after being presented with, for example, a series of descriptions of people belonging to two different groups, participants are given a description and asked to predict which of the two groups the described individual belongs to (i.e., Hamilton & Gifford, 1976; Sherman et al., 2009; Tajfel & Wilkes, 1963). Other studies use similar prediction tasks, but during the initial presentation of the stimuli. For example, those using variation of the *allergy task*, sometimes present participants with a cue (e.g., an allergen), and ask them to predict the outcome (e.g., whether an allergic reaction will occur). Subsequently, they are told what the outcome was (Tajfel & Wilkes, 1963; Blanco et al., 2013; Van Hamme & Wasserman, 1994). Interestingly, there is some evidence that, when used together, the implicit and explicit measures do not yield the same results, which has led some authors to conclude that the biased perceptions stem from a post-hoc evaluation of the information that participants receive, rather than an implicit process that develops during learning (Perales et al., 2005).

In contrast, regardless of the apparent similarities between conditioning and associative learning described in previous chapters, the process of conditioning in the linguistic domain has not been seen as an outcome of a biased perception. For instance, Less in More theory defines it as a process to facilitate learning (Hudson Kam & Newport, 2005, 2009; Newport, 2020), whereas other approaches describe it as an outcome of limited memory capacity (Hudson Kam, 2019; Hudson Kam & Chang, 2009). Relatedly, the bottleneck theory would explain it as an outcome of a generalisation from a biased sample, due to the quickly fading property of language (Christiansen & Chater, 2008, 2016), whereas the communicative efficiency theory (Kirby et al., 2008; Gibson et al., 2019) would describe it as an outcome of a force for learnability. Whether this process is conscious or not has not been widely explored (Leung & Williams, 2011; Samara et al., 2017). Statistical learning in the context of language has been described as an implicit process (Saffran et al., 1996) with research showing that explicit perception is not necessary for learning (Williams, 2006; Leung & Williams, 2011). The few studies surveying participants about their explicit perceptions, however, show that at least some of them are aware of the statistical patterns they perceive in the language and of their behaviour (Hudson Kam & Newport, 2009; Samara et al., 2017).

As described earlier, the literature finding these biases in associative learning mostly relies on explicit measures (Allan & Jenkins, 1980, 1983; Alloy & Abrahamson, 1979; Blanco et al., 2013, 2018; Chow et al., 2019; Matute et al., 2015; Msetfi et al., 2005; Fernbach & Van Boven, 2021; Fiedler, 1991; Hamilton & Gifford, 1976; McGarty & Penny, 1988; Sherman et al., 2009; Tajfel & Wilkes, 1963). It would be easy to assume, hence, that the results we found in Chapters 4 and 5, which were based on this literature, involved an explicit perception of relationships in the language.

In our studies in Chapters 4 and 5, in addition to the comprehension and production tasks, we also asked participants to provide us with explicit estimations of the statistical properties of the language, similarly to domain-general studies asking participants to report their views on the stimuli they were presented with. Crucially, we asked participants to separately report their perception of the input language (the one that they were presented with) and output language (the one they themselves produced in the tasks).

In this chapter, we revisit our results in Chapters 4 and 5, combining them and comparing them with participants' self-report measures. As stated earlier, theories coming from the domain of cognitive psychology and associative learning, which usually rely on self-report measures, predict that our self-report measures will be sensitive to illusory correlation and category accentuation effects, whether they match the self-report results we found in linguistic production or not (Allan & Jenkins, 1980, 1983; Alloy & Abrahamson, 1979; Blanco et al., 2013, 2018; Bott et al., 2021; Chow et al., 2019; Costello et al., 2019; Fernbach & Van Boven, 2021; Fiedler, 1991; Hamilton & Gifford, 1976; Matute et al., 2015; Msetfi et al., 2005; Sherman et al., 2009; Smith & Alpert, 2007; Tajfel & Wilkes, 1963; Van Dessel et al., 2021; McGarty & Penny, 1988). In contrast, linguistic theories describe these processes of change as an outcome of a bias towards a more learnable language, which does not necessarily stem in an explicit awareness and could be due to a bias to reduce uncertainty (Christiansen & Chater, 2008, 2016; Gibson et al., 2019; Hudson Kam, 2019; Hudson Kam & Chang, 2009; Hudson Kam & Newport, 2005, 2009; Newport, 2020; Kirby et al., 2008). Hence, we could predict that whether participants reported having perceived the language to be conditioned in their input or not would not affect whether they accurately report conditioning in their output.

We analysed the relationship between participants' reported perceptions on the language, the condition they were assigned to, and their behaviours on the tasks. For that, we rely on their numerical estimation of the frequency of different elements in the language, as well as their answers to open-ended questions about category and conditioning perception.

We have two competing accounts for the effect of animacy condition and marker condition over participants' perception of conditioning in their input and in their output:

a) Participants' conditioning behaviour directly stems from a biased perception of conditioning in their input, which they are explicitly are of, and
b) Participants' conditioning behaviour arises from a conscious strategical decision to reduce uncertainty in the face of unpredictable variation.

According to the first account (a) we would expect awareness measures, both about the input language and their own productions, to follow the same patterns as we predicted for the behavioural measures. That would lead to the following hypotheses:

1. Participants in the skewed marker conditions report higher perception of conditioning, both in their input and in their output, than those in the uniform marker conditions.
2. Participants in the skewed animacy conditions of the illusory correlation (Chapter 4) report a higher perception of conditioning, both in their input and in their output, than those in the uniform animacy conditions.
3. Participants in the skewed animacy condition of the category accentuation (Chapter 5) report a higher perception of conditioning, both in their input and in their output, than those in the uniform animacy condition.
4. Animacy condition and marker condition interact with each other in the illusory correlation (Chapter 4) conditions to show an increase perception of conditioning beyond the sum of both effects.

In that line, we predict that participants' behaviour in the comprehension and production tasks will correlate with each other and with their perception of conditioning in both their input and their output, and the difference between their input and output perception will have a low variability and will not correlate with any of the measures.

According to the second account (b), that is, if participants added conditioning to the language as an outcome of a bias to reduce uncertainty, we would expect a dissociation between those awareness measures referring to the perception of the input language, and those referring to the perception of the output language. That leads us to the following prediction:

1. The predicted effect of marker and animacy condition will only be present in the awareness measures regarding the output language, but not the input language.
2. Participants' behaviour in the comprehension and production tasks would correlate with each other and with their perception of conditioning in their output, but not in their input.
3. The difference between participants' perceptions on input and output conditioning would correlate with their conditioning behaviour on the comprehension and production task. That is, the higher the difference between the explicit perception of

conditioning in the input language and the explicit perception of conditioning in the output language, the higher the conditioning behaviour will be.

Finally, following up from the discussions in Chapters 4 and 5, we will explore the relationship between category perception, skewness perception, and conditioning perception, and whether they predict conditioning behaviour. We predict that participants who report having perceived categories will show a higher degree of conditioning, as well as those who report having perceived conditioning. Similarly, we predict that participants who have perceived the distribution of categories in their input as skewed, whether it was skewed or not, will show a higher degree of conditioning in the comprehension and production tasks, and will report a higher degree of conditioning in the input and output languages.

## Methods

The methods and hypotheses were preregistered at https://osf.io/5bjkd/.

### Participants

We used the combined sample of experiments in Chapters 4 and 5. That yielded a total sample of 482 native-English speakers, recruited via Prolific (see Chapters 4 and 5 for a detailed description of the conditions of participation). 278 identified as women, 197 as men, and 5 as non-binary, and their average age was of 36.01 years (SD = 13.00, range 18-77). 293 were monolingual English speakers, 114 spoke two languages, 52 spoke three languages, and 23 of them spoke four or more languages. The most common second language was French, followed by Spanish, German, and Italian.

### Design

The combination of the data from experiments in Chapters 4 and 5 yield a 2*3 design, with two between-subject factors: animacy condition (skewed vs. uniform) and marker condition (conditioned vs. skewed vs. uniform). Our outcome variables were the level of conditioning in the comprehension and in the production task, and the reported perceptions of the language. The latter include the responses given on the perceived level of conditioning in the input, and the perceived level of conditioning in the output, whether participants perceived categories, and the perceived level of skewness in the categories. These measures are further defined in the Measures section.

### Procedure

We analysed the data from experiments 3 and 4, so the procedure was identical to that described in those chapters. To quickly sum up, participants learned an artificial language called Panitok, which contained nouns from two categories (animate and inanimate) and two possible plural markers that varied in frequency depending on the condition. The tasks were: noun training, noun testing, plural training, plural testing (comprehension task and production task, in a counterbalanced order), and a post-test questionnaire. In this Chapter we will be focusing on the responses to the post-test questionnaire.

First, participants were asked about their age and gender, whether they were native speakers of English and which languages they spoke. Then, they were asked whether they had perceived different categories in the set of nouns they had to learn (i.e., "Do you think that the things that you have learned to name in Panitok belong to different groups? If so, how would you describe these groups?"). They also were asked to assess their perception of the frequency distribution of the markers for each category in their training through a slider than ranged from 0 to 100 in increments of 1 (i.e., “When **I** was teaching you the language, how often did **I** use the words "bok" and "nim" to describe animals/vehicles? 0- Always bok, 50-Equally, 100-Always nim”), as well as in their own production (i.e., “When **you** were describing images for me, how often did **you** use the words "bok" and "nim" to describe animals/vehicles? 0- Always bok, 50-Equally, 100-Always nim”). By asking participants to separately estimate marker use for each of the categories, we avoided response biases such as potential misunderstandings or differential scale use to represent conditioning. In addition, to obtain a conditioning value higher than 0, participants had to give a different response to the marker estimation with each of the categories.

Finally, participants were asked about their perception of the skewness in the categories (“Indeed, you learned the names for two groups of things - animals and vehicles. Were there the same number of things in each group? If not, which group had more things?”) and their perception of relationship between the two markers and the different categories in the language (“Would you say the use of bok and nim in Panitok depends on the group of things it describes? In what way?”).

## Results

### Slider judgments

We aimed to test whether marker condition, animacy condition, and task order had impacted participants’ reported perception of conditioning in the input they received, as well as in the output they produced. For that we ran four separate models, two for the illusory correlation conditions (Chapter 4), one for input perception and one for output perception, and two for the category accentuation conditions (Chapter 5), one for input perception and one for output perception.

The outcome variable for all four models was the dummy coded response to the sliders. We did this in order to be able to use the same statistical analysis strategy for the reported perception of frequency distribution as for the behaviour in the trials. For example, if a given participant, when asked about how often the speaker used “nim” vs. “bok” with animals, answered 20/100, and when asked the same question about vehicles, they answered 50/100, we generated 200 trials. A 100 of them would have animate as their category, and “nim” as the marker 20 times and “bok” 80 times, and remaining 100 would have inanimate as their category, and “nim” and “bok” as markers 50 times each.

***Illusory correlation conditions***

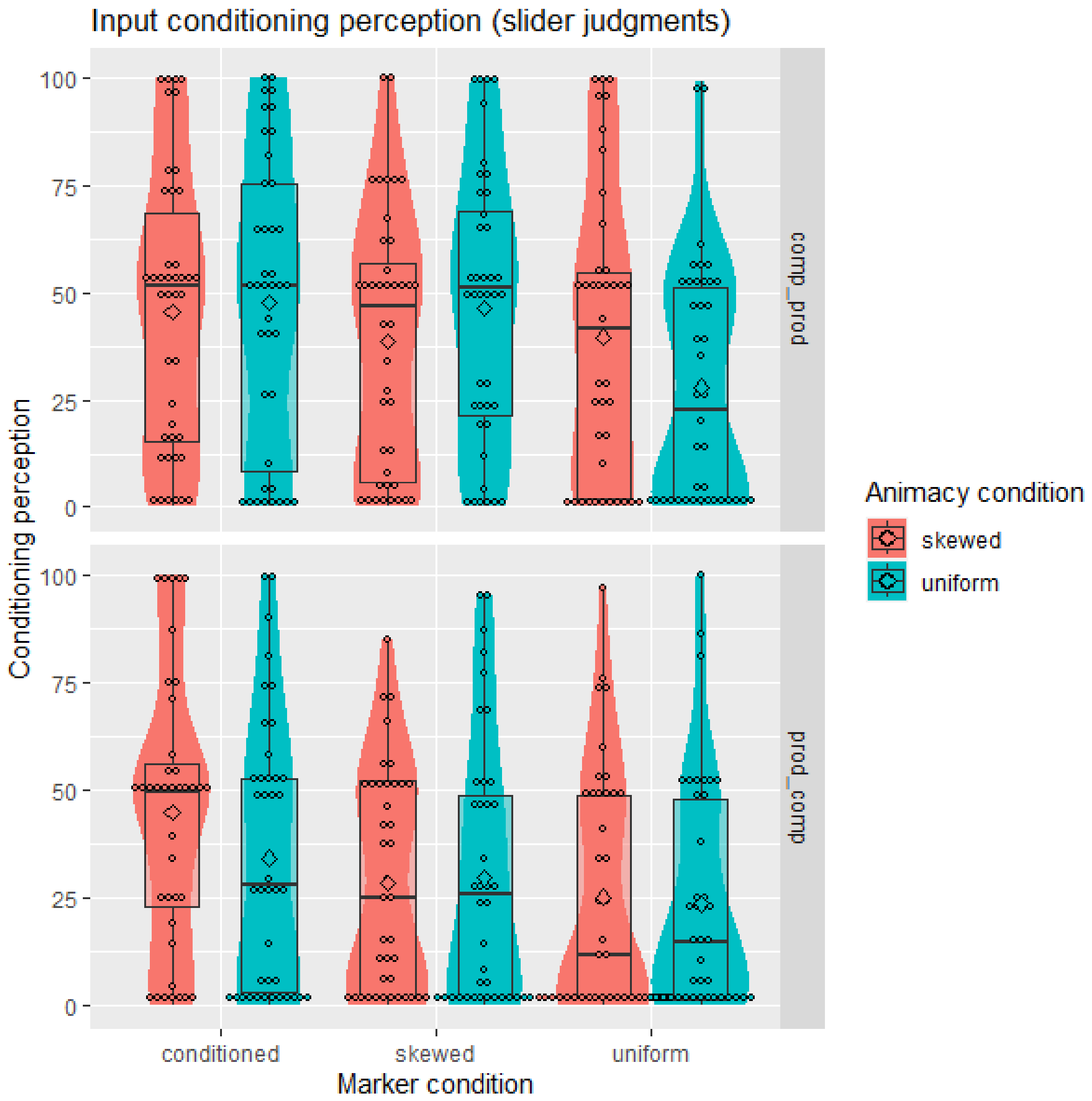
The models for the illusion of causality conditions had category, marker condition, animacy condition, task order, and their interaction as main effects, all sum coded, and random intercepts for participant, marker subset, noun subset, and position of the marker in the scale (which of the markers was on the right or left of the scale) as well as by-participant random slopes for all fixed factors and the position of the marker in the scale. The outcome variable was the dummy coded response to the slider judgment, either for the input language or their output production.

**Input perception.** As it can be observed in Figure 6.1 (right columns), we found a main effect of category (β =- .727, z = -82.39, p <.001), which indicated that participants declared having perceived the use of marker in their input to be conditioned to the category. We also found a significant interaction between marker condition and category, indicating that the conditioning was perceived as being higher in the skewed marker conditions than in the uniform marker conditions (β =-.093, z =-10.61 , p <.001). The interaction between animacy condition and category was however not significant (β = .012, z = 1.34, p =.179). In addition, we found a significant interaction between task order and category, indicating that participants who completed the comprehension task first perceived their input language to be more conditioned than those who completed the production task first (β =.141, z =16.05, p<.001).

We also found a triple interaction between animacy condition, marker condition and category (β = .066, z = 7.54, p<.001), as well as a triple interaction between task order, marker condition, and category (β = -.029, z = -3.25, p <.001), and a four-way interaction between all fixed factors (β = .057, z = 6.41, p <.001). To understand the nature of these complex interactions, we ran follow-up nested analyses by task order using the "joint tests" function of the R package emmeans (Lenth et al., 2022), which corrects for multiple comparisons when looking at nested effects in mixed effect models. In the comprehension first task order, we found a significant interaction between marker condition and category (F(1, 63972) = 91.047, p = <.0001), and a marginally significant interaction between animacy condition and category (F(1, 63972) = 3.151, p = .0759), as well as a triple interaction between the three F(1, 73972) = 92.301, p <.001. A further exploration of this triple interaction, nested by marker condition shows that, within the comprehension first task order, there is a significant effect of animacy condition both within skewed marker condition (F(1, 63972) = 28.673, p <.0001), and within uniform marker condition (F(1, 63972) = 69.641, p <.001), but in opposing directions. Within the skewed marker condition, participants in the uniform animacy condition reported having perceived a higher degree of conditioning than those in the skewed animacy condition, whereas the effect of animacy condition within the uniform marker condition was the predicted one: participants in the skewed animacy condition reported having perceived a higher degree of conditioning than those in the uniform animacy condition. In the production first task order, we found once again a significant interaction between marker condition and category (F (1, 63972) = 28.624, p <.0001), but no effect of animacy condition (F(1, 63972) = 0.006, p =.936), nor any triple interaction between the three F (1, 63972) = .681, p =.409). The two right columns of Figure 6.1 show these results. In summary,  1) participants perceived the marker use in the input language to be conditioned to category use, 2) the predicted effect of marker condition over the conditioning perception in the input language was present across task orders, 3) the predicted effect of animacy condition was only present within the uniform marker condition in the comprehension first task order, 4) we did not find the predicted interactions between marker condition and animacy condition, and 5) participants who completed the comprehension task first reported having perceived a higher degree of conditioning than their counterparts.

Figure 6.1.

*Reported conditioning perception in the input by animacy condition, marker condition and task order.*



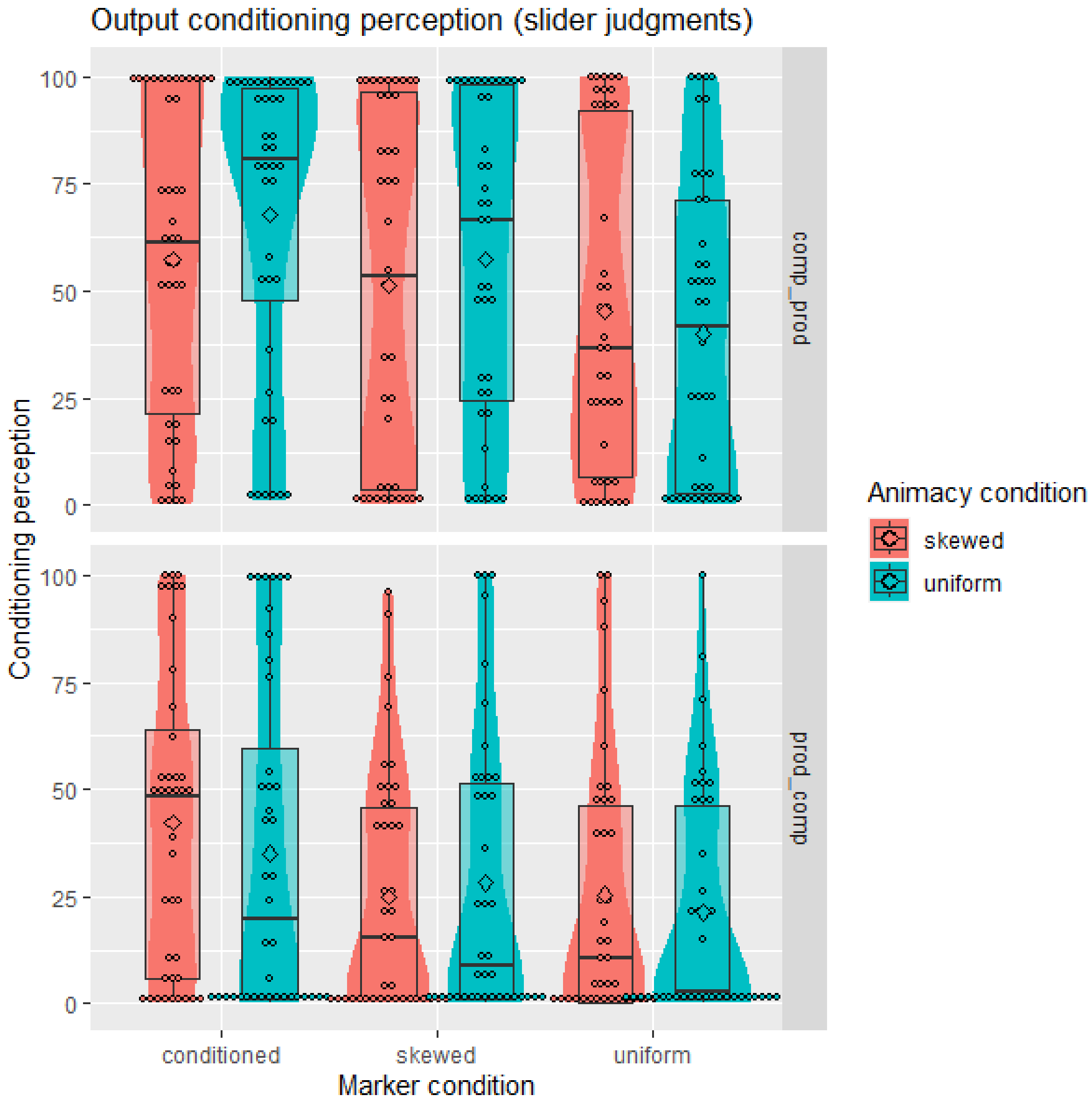
*Note.* The y-axis represents input conditioning perception (see Measures), based on participants' responses to the slider. The column on the left, in which the marker condition is referred to as "conditioned" refer to the category accentuation conditions, whereas the two columns in the right (labelled as "skewed" and "uniform" marker conditions) refer to the illusory correlation conditions. The upper panel presents the results for the comprehension first task order, whereas the lower panel presents it for the production first task order. Colours represent the animacy conditions: orange for skewed and green for uniform.

**Output perception.** Similarly to the results regarding input perception, as Figure 6.2 (right columns) show, we found a main effect of category, β = -.936, z = -95.48, p <.001, which indicated that participants declared having perceived the use of marker in their input to be conditioned to the category, as well as a significant interaction between marker condition and category, indicating that the conditioning was perceived as being higher in the skewed marker conditions than in the uniform marker conditions (β = -.107, z = -10.94, p <.001). The interaction between animacy condition and category was significant for the perception of conditioning in the output (β = -.030, z = -3.11, p =.002): participants in the skewed animacy conditions perceived to have conditioned their input more than those in the uniform animacy condition. In addition, we found a significant interaction between task order and category, indicating that participants who completed the comprehension task first perceived their input language to be more conditioned than those who completed the production task first (β =-.312, z = -31.86, p<.001). Finally, we found a triple interaction between animacy condition, marker condition and category (β = .082, z = 8.36, p <.001), a triple interaction between marker condition, task order, and category (β = -.051, z = -5.17, p <.001), and a four-way interaction between all fixed factors (β = .031, z = 3.16, p = .002).

Once again, we ran nested analyses by task order to better understand these interactions. Within the comprehension first task order, we found a significant interaction between marker condition and category (F(1, 63974) = .115690, p <.001), a marginally significant interaction between animacy condition and category (F(1, 63974) = 3.712, p = .054), and a triple significant interaction between animacy condition, marker condition, and category (F(1, 63974) = 59.147, p <.001). A further follow-up nested analysis by marker condition showed that the interaction between animacy condition and category was present both within the skewed marker condition (F(1, 63974) =15.452, p <.001), and the uniform marker condition (F(1, 63974) =50.002, p <.001), but in the opposite direction. Within the skewed marker condition, opposite to what we predicted, participants in the uniform animacy condition perceived to have conditioned their output more than those in the skewed animacy condition, whereas we found the predicted effect within the uniform marker condition. Within the production first task order, we found a significant interaction between marker condition and category (F(1, 63974) =6.355, p =.012), between animacy condition and category (F(1, 63974) =18.972, p <.001), and, once again, a triple interaction between the three (F(1, 63974) =15.403, p <.001). A further nested analysis by marker condition showed that the interaction between animacy condition and category was only present within the uniform marker condition (F(1, 63974) = 21.718, p <.001), but not within the skewed marker condition (F(1, 63974) =.944, p =.331). The right panels on Figure 6.2 present these results.

Figure 6.2.

*Reported conditioning perception in the output by animacy condition, marker condition and task order.*



*Note*. The y-axis represents output conditioning perception (see Measures), based on participants' responses to the slider. The column on the left, in which the marker condition is referred to as "conditioned" refer to the category accentuation conditions, whereas the two columns in the right (labelled as "skewed" and "uniform" marker conditions) refer to the illusory correlation conditions. The upper panel presents the results for the comprehension first task order, whereas the lower panel presents it for the production first task order. Colours represent the animacy conditions: orange for skewed and green for uniform.

In summary, participants' reported perceptions of the level of conditioning in their productions followed a similar pattern to the reported perceptions in the linguistic input: 1) participants reported having conditioned the marker use on categories, 2) participants in the skewed marker conditions were more likely to report a higher level of conditioning, 3) We only found the predicted effect of animacy condition within the uniform marker conditions in both task orders, but not within the skewed marker conditions, 4) participants reported a higher degree of conditioning when they completed the comprehension task first.

***Category accentuation conditions***

The models for the category accentuation conditions had category, animacy condition, task order, and their interaction as main effects, and random intercepts for participant, marker subset, noun subset, and position of the marker in the scale (which of the markers was on the right or left of the scale) as well as by-participant random slopes for all fixed factors and the position of the marker in the scale. The outcome variable was the dummy coded response to the slider judgment, either for the input language or their output production. We sum coded animacy condition and task order, and we used custom coding for category. This entered the contrast to the input conditioning rate in the hypothesis matrix and allowed us to compare participants' degree of conditioning with the input one. As the hypothesis matrix was not centred, we used its generalised inverse as our contrast matrix. The resulting contrast matrix was the same as the one presented in Chapter 5 (p.150) containing {1,1} in the intercept column, and {1.2, -.4} in the contrast column. The first row was for minority category and the second for majority category.

**Input perception.** As predicted, we found a main effect of category, showing that participants declared to have perceived the input to be conditioned on category less than it was in the input (β = -1.182, z=-74.33, p<.001). We also found an effect of animacy condition (β =-.126, z =-7.91, p< .001) in the predicted direction: participants in the skewed animacy condition declared perceiving the level of conditioning to be higher than those in the uniform animacy condition. However, we also found a triple interaction between task order, animacy condition, and category (β = .062, z =3.91, p<.001). A nested analyses by task order showed that the effect of animacy was in the predicted direction only within the production first task order (F(1, 32783) = 69.616 , p <.001), but it was in the opposite direction within the comprehension first task order (F(1, 32783) = 8.053, p = .004). Finally, as in the illusory correlation conditions, participants that completed the comprehension task first perceived the input to be more conditioned than those who completed the production task first (β = -.070, z =-4.74, p<.001). The left panel in Figure 6.1 shows these results.

**Output perception.** As with input perception, we found the predicted main effect of category, showing that participants declared to have perceived the input to be conditioned on category less than it actually was (β = -1.485, z=-85.42, p<.001). In addition, we found an effect of animacy condition (β =-.092, z =-5.30, p< .001) in the predicted direction: participants in the skewed animacy condition declared perceiving the level of conditioning to be higher than those in the uniform animacy condition. However, we also found a triple interaction between task order, animacy condition, and category (β = .040, z =2.32, p=.020). A nested analyses by task order showed that the effect of animacy was in the predicted direction only within the production first task order (F(1, 32783) = 31.128 , p <.001), but it was in the opposite direction within the comprehension first task order (F(1, 32783) = 4.173 p = .041). Finally, as in the illusory correlation conditions, participants that completed the comprehension task first perceived the input to be more conditioned than those who completed the production task first (β = -.304, z =-17.54, p<.001). The left panel in Figure 6.2 shows these results.

### Correlation between behaviour and perception

We wanted to test whether, across all conditions, participants' perceptions of conditioning in their input and in their production correlated with their behaviour in the tasks. We used five different measures:

- Input conditioning perception: as described before, this measure was based on participants' answers to the slider question asking them about the frequency with which each of the markers was used with each of the categories in the language they were taught. Precisely, it was based on the absolute difference between the estimation of use of one of the markers for one of the categories and for the other. This ranged from 0 (no difference; no conditioning) to 100 (full conditioning). For instance, if a participant answered that, in the language they were taught, the native speaker used marker 1 30% of the times with animals, and 50% of the times with vehicles, their score would be 20.

- Output conditioning perception: this measure followed the same logic as the one on the perception of conditioning in the input language but was based on the question regarding participants' own behaviour.

- Input-output difference: This measure is the subtraction of the "input conditioning perception" from "output conditioning perception". and it ranges from -100 to 100. A value of 100 means that participants "input conditioning perception" was of 0 and their output conditioning perception of "100", that is, they perceived the marker not to be conditioned to category in the language they were presented with, but they perceived their own output to be conditioned. A value of -100 means that participants "input conditioning perception" was 100 and their output conditioning perception was 0, that is, they perceived the marker to be conditioned to the category in their input, but not in the output.
- Comprehension task conditioning: this is a measure ranging from 0 to 1 based on participants' behaviour in the comprehension task. In this task, participants were shown a marker and asked to predict which of two images (an animal or a vehicle) would come next. We calculated the absolute difference between the proportion of trials in which they chose an animal with each of the markers. For example, if a participant always clicked on the image with an animal with both markers, the value of conditioning would be 0, whereas if they clicked on the animal a 100% of the times with one of the markers and only 30% of the times with the alternative marker, the value would be .7.

- Production task conditioning: this measure also ranged from 0 to 1 and was based on participants' behaviour in the production task. Participants were presented with an image containing two animals or two vehicles and asked to describe it. We recorded the proportion of trials in which they used each of the markers with each of the animacy categories. This measure represented the absolute difference in the proportion of trials in which participants used one of the markers between categories. For instance, if a participant used marker "bok" 60% of the times with both animals and vehicles, their score would be 0, whereas if they always used "bok" for animals and "hap" for vehicles, their score would be of 1.

We first checked for the distribution of our measures, to see if they followed a normal distribution using Shapiro-Wilks tests, and if they were variable. Table 6.1 shows the descriptives for each of these measures. Aside from those metrics, it is important to note that regardless of the wide range and standard deviation of the values for input-output difference, 38.8% of participants had a value that ranged between -5 and 5, and hence reported very similar perceptions of input and output.

Table 6.1.

*Descriptives for measures of conditioning behaviour and perception.*

| | Mean | Median | SD | Min | Max | W | p |
|---|---|---|---|---|---|---|---|
| Input conditioning perception | 35.929 | 34.000 | 31.234 | 0 | 100 | 0.898 | <.001 |
| Output conditioning perception | 41.463 | 36.000 | 37.691 | 0 | 100 | 0.855 | <.001 |
| Input-output difference | 5.533 | 0.500 | 30.518 | -97 | 100 | 0.962 | <.001 |
| Comprehension task conditioning | 0.595 | 0.646 | 0.347 | 0 | 1 | 0.884 | <.001 |
| Production task conditioning | 0.439 | 0.296 | 0.387 | 0 | 1 | 0.840 | <.001 |

*Note.* The range for "Input conditioning perception" and "Output conditioning perception" was 0 to 100, whereas the range for "Input-output difference" was -100 to 100, and the range for "Comprehension task conditioning" and "Production task conditioning" was 0 to 1.

We used Spearman's rho as the correlation index, as none of the measures followed a normal distribution, and applied Bonferroni adjustment for multiple comparisons, establishing the significance threshold as .005. Sample size was 482 for all correlations. As Table 6.2 shows, there are moderate to strong correlations between these measures. The strongest correlations are between input and output conditioning perception, and between output conditioning perception and conditioning in the production task. It is important to note that participants' perceptions of their own behaviour (output conditioning perception) correlated more strongly with their behaviour in both tasks than participants' input conditioning perception, and that the difference between the perceived level of conditioning in input and output moderately correlated with participants' behaviour in the production task and had a small correlation with their behaviour in the comprehension task.

Table 6.2.

Correlations between perception and behavioural measures.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. Input conditioning perception | - | | | | |
| 2. Output conditioning perception | .588** | - | | | |
| 3. Input-output difference | -.204** | .598** | - | | |
| 4. Comprehension task conditioning | .372** | .421** | .171** | - | |
| 5. Production task conditioning | .441** | .713** | .441** | .343** | - |

+p<.010, *p<.005, ** p<.001

Finally, given the conceptual differences between the illusory correlation conditions and the category accentuation conditions, we split the correlations by experiment. As Table 6.3 shows, the correlation indexes were similar for both experiments. However, we found some notable differences. The correlation between input-output difference and comprehension task conditioning was not significant for illusory correlation conditions, but it was for category accentuation conditions, whereas the correlation between input-output difference and production task conditioning was moderate and significant for both experiments. Perhaps relatedly, the correlation between comprehension task and production task conditioning was only small for the illusory correlation conditions (rho = .205) and moderate for the category accentuation conditions (rho = .460).

Table 6.3.

Correlations between perception and behavioural measures divided by condition.

| Category accentuation conditions<br>Illusory correlation conditions | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1. Input conditioning perception | - | 571** | -.219* | .324** | .455** |
| 2. Output conditioning perception | .586** | - | .605** | .436** | .768** |
| 3. Input-output difference | -.252** | .601** | - | .197* | .463** |
| 4. Comprehension task conditioning | .351** | .362** | .104 | - | .460** |
| 5. Production task conditioning | .391** | .649** | .422** | .205** | - |

+p<.010, *p<.005, ** p<.001. Upper diagonal panel represents the correlation indexes for the category accentuation condition (N = 162) and the lower diagonal panel for illusory correlation conditions (N = 320).

## Open-ended questions

### *Category perception*

We asked participants to answer to the question: “Do you think that the things that you have learned to name in Panitok belong to different groups? If so, how would you describe these groups?". This was a text-based open-ended question. We coded their responses according to the categorising factor that they identified. The resulting codes, and a few examples of responses that fell into each code, as well as the number of responses under the code, can be seen in Table 6.4 Note that some answers fell under more than one code.

Table 6.4.

*Coding system for responses in category perception*

| N (%) | Code | Examples |
|---|---|---|
| 353 (73.24%) | animacy | “Vehicles/living creatures”<br>“Animals and Vehicles”<br>“nim is methods of transportation / hap describes animals” |
| 51 (10.58%) | prefix | “There are the 'Da' group that take 'Nim' as 'two', and the 'Te' group that take 'Bok' as 'two”<br>“There are the 'Da' group that take 'Nim' as 'two', and the 'Te' group that take 'Bok' as 'two” |
| 42 (8.71%) | gender | “I think the words are gendered”<br>“Male and female” |
| 33 (6.84%) | no | “No”<br>“I wouldnt be able to say. I found it tricky to remember”<br>“Nope, there were no trends or groups from what I could see” |
| 7 (1.45%) | orientation | “Whether they are facing left or right” |
| 1 (0.21%) | colour | “black and white or coloured images?” |
| 1 (0.21%) | size | “I think Bok is to describe things that are bigger” |
| 58 (12.03%) | NA | “N/A”<br>“?”<br>“unsure”<br>“similar to english maybe” |

Three of the categorising factors (animacy, prefix, and orientation) correctly distinguished between our two categories of items (animate and inanimate). All the nouns for animals started with the prefix “da-“ and were facing left, whilst all the vehicles started with “te-“ and were facing right. For that reason, we categorised participants as having shown awareness of the animacy categories if their answers were coded as “animacy”, “prefix” or “orientation”, and as not having shown awareness of animacy categories if their responses did not contain any of these codes. Table 6.4 shows the proportion of participants having shown awareness per condition and task order.

Table 6.5.

*Number of participants who showed awareness of the category by condition.*

| | Comprehension fist | | Production first | |
|---|---|---|---|---|
| | Uniform animacy | Skewed animacy | Uniform animacy | Skewed animacy |
| Uniform marker distribution | 31 (77.5%) | 33 (82.5%) | 25 (62.5%) | 31 (77.5%) |
| Skewed marker distribution | 31 (77.5%) | 28 (70%) | 32 (80%) | 23 (57.5%) |
| Conditioned marker distribution | 31 (77.5%) | 32 (76.19%) | 32 (80%) | 31 (77.5%) |

As Table 6.5 shows, the percentage of participants who had shown awareness of the categories varied somewhat across condition (between 57.5% and 83.5%). We used Chi-square Test of Independence to check whether the proportion of participants who showed awareness of the categories differed by animacy or marker condition. However, we did not find marker condition ($\chi^2$(2, 1) = 1.827, p = .271), animacy condition ($\chi^2$(1, 1) = .331, p = .565), nor task order ($\chi^2$(1, 1) = 1.211, p = .271), to be reliable predictors of awareness. Finally, we tested whether awareness of the categories predicted conditioning behaviour or reported perception of conditioning. We ran five t-tests with each of the five measures of conditioning as outcome variables and category perception as the predictor. We used Bonferroni correction to our p-value, which we established at .01 given that we ran four separate tests. As Table 6.6 shows, across conditions, participants who reported having perceived the categories showed a higher degree of conditioning than those who did not, both in the comprehension and production tasks, and they reported having perceived a higher degree of conditioning both in the input language, and in their own productions. Interestingly, the effect size was the biggest for conditioning in the production task and output conditioning perception. We also found a marginally significant effect of category perception on input-output conditioning difference (p = .020).

Table 6.6.

*Conditioning behaviour and conditioning perception by category perception*

| Outcome variable | Category perception | N | Mean | SD | t | Cohen's d |
|---|---|---|---|---|---|---|
| Input conditioning perception | no | 122 | 28.418 | 28.325 | 3.10** | .325 |
| | yes | 360 | 38.475 | 31.797 | | |
| Output conditioning perception | no | 122 | 28.385 | 31.716 | 4.52*** | .474 |
| | yes | 360 | 45.894 | 38.556 | | |
| Comprehension task conditioning | no | 122 | 0.501 | 0.340 | 3.50*** | .366 |
| | yes | 360 | 0.626 | 0.344 | | |
| Production task conditioning | no | 122 | 0.298 | 0.343 | 4.78*** | .501 |
| | yes | 360 | 0.487 | 0.390 | | |
| Input-output conditioning difference | no | 122 | -0.033 | 26.3 | 2.34+ | .245 |
| | yes | 360 | 7.42 | 31.6 | | |

+p<.05, ** p <.0025, *** p <.001. *Note.* The range for "Input conditioning perception" and "Output conditioning perception" was 0 to 100, whereas the range for "Input-output difference" was -100 to 100, and the range for "Comprehension task conditioning" and "Production task conditioning" was 0 to 1.

***Category skewness perception***

We checked whether participants perceptions of the proportion of nouns in each category was accurate. For those participants in the uniform animacy conditions, there was an equal number of nouns in each category, whereas within those in the skewed animacy conditions, the language was skewed towards animates for 124 of them and towards inanimate for the remaining 118.

This was also an open-ended question, and we coded the responses as "majority animates" (e.g., "animals", "4 animals and 2 vehicles"), "majority inanimate" (e.g., "more vehicles", "objects"), "equal number" (e.g., "same number") or NA if they had not answered the question or their response could not be understood. Table 6.7 shows the number of participants the number of participants within each of the conditions.

Table 6.7.

*Category skewness perception in relation to category skewness in the input*

| | Category skewness perception | | | |
|---|---|---|---|---|
| Condition | Majority animate | Majority inanimate | Equal number | NA |
| Majority animate | 103 (83.06%)[a] | 0[b] | 9 (7.26%)[c] | 12 (9.68%) |
| Majority inanimate | 0[c] | 101 (85.69%)[a] | 12 (10.17%)[b] | 5 (4.24%) |
| Uniform animacy | 56 (23.33%)[b] | 29 (12.08%)[c] | 134 (55.83%)[a] | 21 (8.75%) |

[a] Correct estimation, [b] Overestimate animate, [c] Overestimate inanimate.

Participants were, overall, accurate in their reporting of the number of items in each category. However, within the uniform animacy condition, only 55.83% accurately reported that there was the same number of items per category, with a 35.41% either overestimating the number of nouns of one of the categories. Across all conditions, participants were more likely to overestimate the number of animate nouns (14.11%) than the number of inanimate nouns (7.88%).

As discussed in Chapter 4, one of our potential explanations for the unexpected results (and lack thereof) regarding the effect of animacy condition was that participants may have had an inaccurate perception of the animacy distribution. For that reason, we tested whether there was a difference in our four outcome variables depending on participants skewness perception. We classified participants as "skewed animacy" if they reported having perceived more animals than vehicles or vice versa, and as "uniform animacy" if they reported having perceived an equal number of animals and vehicles, regardless of the actual animacy category they were assigned to. We discarded those participants who did not answer the question on the perception of skewness. As Table 6.8 shows, after applying Bonferroni correction and establishing the significance threshold at .0125, none of the differences between conditions were significant.

Table 6.8.

*Conditioning behaviour and conditioning perception by skewness perception*

| Outcome variable | Skewness perception | N | Mean | SD | t | p-value | Cohen's d |
|---|---|---|---|---|---|---|---|
| Input conditioning | Skewed | 289 | 36.360 | 31.426 | .384 | .728 | .087 |
| perception | Uniform | 155 | 35.284 | 30.350 | | | |
| Output conditioning | Skewed | 289 | 43.304 | 37.071 | .874 | .382 | .035 |
| perception | Uniform | 155 | 40.039 | 38.353 | | | |
| Comprehension task | Skewed | 289 | 0.619 | 0.338 | 1.769 | .078 | .176 |
| conditioning | Uniform | 155 | 0.558 | 0.356 | | | |
| Production task | Skewed | 289 | 0.438 | 0.386 | -.703 | .482 | -.07 |
| conditioning | Uniform | 155 | 0.465 | 0.394 | | | |

*Note*. The range for "Input conditioning perception" and "Output conditioning perception" was 0 to 100, whereas the range for "Input-output difference" was -100 to 100, and the range for "Comprehension task conditioning" and "Production task conditioning" was 0 to 1.

***Conditioning perception***

We asked participants whether they had perceived the marker use to be conditioned to anything in an open-ended question (i.e., "Would you say the use of bok and nim in Panitok depends on the group of things it describes? In what way?"). Once again, we coded their responses depending on what factor they believed they were conditioned on. Table 6.9 shows the frequency of each code. Less than half of the participants reported having perceived conditioning to animacy, with almost a third claiming that they had not. A few participants claimed that they had conditioned animacy to marker in their productions but did not think they were conditioned in their input.

Table 6.9.

*Coding system for responses in conditioning perception*

| **N (%)** | **Code** | **Examples** |
|---|---|---|
| 219 (45.44%) | animacy | "bok seemed to be for objects and hap for animals"<br>"Bok was more for animals hap was more for objects" |
| 131 (27.18%) | no | "It seemed to make no difference" |

| | | |
|---|---|---|
| | | “No”<br>“they seem fairly interchangeable” |
| 14 (2.90%) | animacy/prod | “It varied in the images but my brain defaulted to using bok for vehicles and nim for animals.”<br>“”I don’t think it did, but I put them into two groups.” |
| 10 (2.07%) | gender | “describes 2 things, there could be a gender element but I couldn’t see what” |
| 8 (1.66%) | orientation | “I thought it was to do with if they were facing to the right or the left” |
| 8 (1.66%) | Other factor | “I assumed hap and bok were maybe left and right”<br>“I feel that the words are interchangeable. One is perhaps more formal than the other, however it wasn’t obvious which was which.”<br>“hap tended to be used more for words ending in a/as and bok seemed to be more for I and o” |
| 5 (1.04%) | prefix | “I thought it was hap when the word started with T and bok if the word started with D” |
| 87 (18.05%) | NA | “Possibly i assumed it meant 2 and Pair”<br>“It know”<br>“I think ’o, I'm not really sure” |

As with the responses with animacy condition, we classified the answers as reporting to have perceived conditioning between animacy (or any correlating factors, namely prefix and orientation) and marker, or not. Table 6.10 shows the number of participants that had by condition.

Table 6.10.

*Proportion of participants per condition who report having perceived conditioning in the input*

| | Comprehension first | | Production first | |
|---|---|---|---|---|
| | Uniform animacy | Skewed animacy | Uniform animacy | Skewed animacy |
| Uniform marker distribution | 8 (20%) | 21 (52.5%) | 16 (40%) | 13 (32.5%) |

|  |  |  |  |  |
|---|---|---|---|---|
| Skewed marker distribution | 27 (67.5%) | 18 (45%) | 19 (47.5%) | 19 (47.5%) |
| Conditioned marker distribution | 24 (60%) | 26 (65%) | 15 (37.5%) | 26 (65%) |

The proportion of participants who declared having perceived category as a predictor did not vary by task order ($\chi^2(1) = 1.638$, p = .201), nor by animacy condition ($\chi^2(1) = 1.205$, p =.272). However, it varied by marker condition ($\chi^2(2) = 14.142$, p <.001). A post-hoc Fischer-test, Bonferroni corrected for multiple comparisons, revealed that difference was between the uniform marker condition and both the skewed marker condition and conditioned marker condition, but there was no difference between the latter two. Once again, we tested whether those participants who declared having perceived category also showed a higher level of conditioning behaviour and a higher perception of conditioning. As Table 6.11 shows, this difference was significant for all measures of conditioning after applying Bonferroni correction for multiple comparisons and establishing the significance threshold at .0125, but not for input-output conditioning difference. Interestingly, the effect size was the biggest for input conditioning perception.

Table 6.11.

*Conditioning behaviour and conditioning perception (sliders) by conditioning perception (open-ended question)*

| Outcome variable | Conditioning perception | N | Mean | SD | t | Cohen's d |
|---|---|---|---|---|---|---|
| Conditioning perception in the input | no | 250 | 24.0 | 26.9 | 9.48*** | .865 |
|  | yes | 232 | 48.8 | 30.5 |  |  |
| Conditioning perception in the output | no | 250 | 28.3 | 32.9 | 8.51*** | .776 |
|  | yes | 232 | 55.6 | 37.5 |  |  |
| Comprehension task conditioning | no | 250 | 0.479 | 0.346 | 8.06*** | .735 |
|  | yes | 232 | 0.719 | 0.302 |  |  |
| Production task conditioning | no | 250 | 0.320 | 0.329 | 7.39*** | .674 |
|  | yes | 232 | 0.568 | 0.404 |  |  |
|  | no | 250 | 4.34 | 26.3 | .88 | .081 |

| | | | | |
|---|---|---|---|---|
| Input-output conditioning difference | yes | 232 | 6.81 | 34.5 |

*** p <.001. *Note.* The range for "Input conditioning perception" and "Output conditioning perception" was 0 to 100, whereas the range for "Input-output difference" was -100 to 100, and the range for "Comprehension task conditioning" and "Production task conditioning" was 0 to 1.

## Discussion

In this chapter, we aimed to explore if the manipulations used in Chapters 4 and 5 affected participants' explicit perception of conditioning, and how these related to participant's behaviours in the comprehension and production tasks. The purpose was better understanding the cognitive processes behind conditioning and comparing the results we found using linguistic measures (production and prediction) with the results found in explicit measures, which are closer to the measures used in the fields of illusory correlation, illusion of causality, and category accentuation (Allan & Jenkins, 1980, 1983; Blanco et al., 2013, 2018; Chow et al., 2019; Matute et al., 2015; Fernbach & Van Boven, 2021; Fiedler, 1991; Hamilton & Gifford, 1976; McGarty & Penny, 1988; Sherman et al., 2009; Tajfel & Wilkes, 1963).

We asked participants to estimate on a slider the frequency of each of the markers with each of the semantic categories, both in the input language and in their output language. We then calculated the difference in marker use between categories, obtaining two estimates of perceived conditioning, one for the input language (the language they had been taught) and another one for the output language (the one that they themselves had produced).

When looking at illusory correlation conditions, our results greatly converged with those we found in Chapter 41) participants' responses showed that they had perceived marker to be conditioned to category both in their input and in their output, 2) marker skewness had an impact on conditioning perception, 3) conditioning perception was higher when the comprehension task came in first place, 4) the effect of animacy skewness was limited to some of the conditions, and 5) we did not find the predicted interaction between marker condition and animacy condition. There was a clear parallel between the conditions within which we found the predicted effect of animacy condition vs. the opposite effect and the results in the comprehension and production tasks: input perceptions followed the pattern we found in the comprehension task, whereas output perceptions followed the pattern in the production task.

Similarly, when looking at input and output conditioning perception in the category accentuation conditions, we found that the patterns matched those found in the comprehension and production tasks in Chapter 5. Across both input and output conditioning perception, we found an overall effect of task order, participants that completed the production task first tended to underestimate the level of conditioning both in their input and in their output, whereas those who completed the comprehension task first tended to overestimate the level of conditioning in their input and output. In addition, as it had been the case both in the comprehension and the production task in Chapter 5, we only found the predicted effect of animacy in the production first task order, even if the degree of conditioning perception within these participants was overall lower than that in the input.Similarly, as it was the case in the comprehension task described in Chapter 5, we found the opposite effect of animacy to the predicted one when comprehension task was presented first, but only in the input perception measure:: participants in the uniform animacy condition had perceived conditioning in the input language to be higher than in those participants in the skewed animacy condition.

Next, as predicted, we found that the input and output perceptions correlated with each other and with participants' behaviour in the comprehension and production tasks. The correlation between input and output perception was higher for the production task than for the comprehension task, and across the comprehension and production tasks, the correlation between perceptions and behaviour was higher for output perception than input perception. This goes in line with the results described earlier, which show a higher parallelism between the behavioural results in the production tasks and the reported perception, particularly when it came to input perceptions. Finally, the correlation between input and output perception was moderate for both illusory correlation conditions and category accentuation conditions, and the correlation between conditioning in the comprehension task and in the production task was low for participants in the illusory correlation conditions, and moderate for participants in category accentuation conditions.

This convergence between output conditioning perception and conditioning in the production task, but not in the comprehension task, both in terms of the patterns that we identify and in the strength of the correlations, can be easily explained by the phrasing of the question we used when measuring output conditioning perception: "When you were describing images for me, how often did you use the words "bok" and "nim" to describe animals/vehicles?”. This more closely related to participants behaviour in the production task, and also shows how participants were aware of their own behaviour and were sensitive to nuances in the wording of the question. In addition, this is in line with previous research, which showed that nuances in the wording of the question in contingency tasks, such asking participants if a cue predicted or caused an outcome, changed participants’ judgments in similarly designed slider scales (Vadillo et al., 2011; Collins & Shanks, 2006; Perales & Shanks, 2008).

Finally, the input-output conditioning difference measure showed some interesting results. The variability of the measure shows that participants effectively discriminated between the statistical properties of the language they were taught and the one they produced, with values ranging from -97 to 100. A value of -97 would reflect the case of a participant who perceived marker to be fully conditioned to animacy in the input, but whose marker use did not depend on animacy in their output, either because they selected markers randomly, or because they fully regularised their use, that is, the only used one of the markers in the production task. A value of 100 would reflect the case of a participant who perceived marker use not to be conditioned to category in the input, but that imposed full conditioning in their output, that is, they use each of the markers with one of the animacy categories even if they perceived that marker use was independent from category in their input. We found a positive correlation between this input-output conditioning difference measure and participants’ behaviour on the production task both in the illusory correlation conditions and in the category accentuation conditions. However, the correlation between this index and the conditioning in the comprehension task was only significant for participants in the category accentuation conditions.

The last of our analyses involved participants’ answers to the open-ended questions. We found those participants who declared having conditioned the marker use to category had on average shown a higher degree of conditioning on both the comprehension and the production task, and in their perception of the input and output conditioning. Furthermore, we found that the proportion of participants who declared having perceived conditioning in the input language was higher in the skewed marker conditions than in the uniform marker conditions. Similarly, those participants who showed evidence of having explicitly perceived animacy as a category showed a higher degree of conditioning in their behaviour in the comprehension and the production task, and declared having perceived a higher degree of conditioning in the input and in the output. Finally, having perceived the animacy distribution as skewed (regardless of whether it was indeed skewed or not) did not have a significant effect on conditioning behaviour nor conditioning perception.

Based on the results presented above, we can identify a few key insights:

1) Explicit measures of conditioning were correlated with participants' behaviour, showing that participants are to some extent aware of their linguistic behaviour and that, these measures, commonly used in the field of contingency learning (Allan & Jenkins, 1980, 1983; Alloy & Abrahamson, 1979; Blanco et al., 2013, 2018; Chow et al., 2019; Matute et al., 2015; Msetfi et al., 2005) or social psychology (Fiedler, 1991; Hamilton & Gifford, 1976; Fernbach & Van Boven, 2021; McGarty & Penny, 1988Tajfel & Wilkes , 1963) can also capture participants' perceptions of statistical properties of the language. Statistical learning is considered implicit (Aslin et al., 1998; Saffran et al., 1996), and explicit awareness does not seem a requisite in the perception of patterns (Williams, 2006; Leung & Williams, 2011). However, some previous studies have shown that participants can describe the statistical patterns they perceive in the language and in their own linguistic behaviour (Hudson Kam & Newport, 2009; Samara et al., 2017). Our results match these findings and expand them, showing correlations between participants' frequency estimations of different aspects of the language and their own behaviour, and comparing them to participants' own explicit assessment of conditioning. In the linguistics literature, participants conditioning behaviour has been described as a bias to reduce uncertainty, whether it was deemed conscious or unconscious (Kirby et al., 2008; Christiansen & Chater, 2008, 2016; Hudson Kam, 2019; Hudson Kam & Chang, 2009; Samara et al., 2017). However, the underlying perceptions of statistical patterns in the language has not been considered. Through the gathering of explicit perceptions of statistical patterns, we can better understand the origins of conditioning behaviour and how different forces interact in its development, establishing links with domain-general cognitive processes.

2) The outcome-density effect that we find in the literature of illusion of causality, which states that a higher frequency of the outcome will lead individuals to associate it with a cue (Allan & Jenkins, 1980; Alloy & Abrahamson, 1979; Msetfi et al., 2005) can account for some of the results we find in conditioning. In our study, the outcome (the

variable element that needed to be predicted) was the plural marker and the cue (the potential predictor of the outcome) was the animacy category. In Chapter 4, we saw that skewness in marker distribution in the absence of a real relationship between marker and category led participants to show more conditioning behaviour, both in the comprehension task and in the production task, and consistently across task orders. In this chapter we observed that participants' reported perception of conditioning matched their behaviour in this matter. Participants in the skewed marker conditions were more likely to report that each of the markers was used with a different frequency with each of the categories in their input (input conditioning perception), as well as in their output (output conditioning perception). In addition, when directly asked whether animacy category had any relationship with marker use, participants in skewed marker conditions were more likely to respond affirmatively than participants in the uniform marker conditions. Responding affirmatively to this question (reporting having perceived conditioning) and showing evidence of having identified animacy as a category both predicted a higher degree of conditioning in the linguistic behaviour. However, only the former (reporting to have perceived conditioning) varied depending on marker skewness, adding evidence to the hypothesis that the mechanism behind the effect of marker skewness of conditioning was related to a biased perception of the probabilities, as predicted by the outcome-density effect.

3) When faced with unpredictable variation, some participants' behaviours and answers suggest that they might be consciously use conditioning as a strategy, particularly when cognitive demands are higher, such as in the production task. These participants reported - using the perception sliders - that the conditioning in the output they produced was higher than that in their input. Some of them explicitly reported, in the open-ended question, to have imposed conditioning. For example, one participant who had shown a high level of conditioning and a high input-output conditioning difference score responded to the open-ended question about conditioning as follows: "During the learning phase, both [markers] seemed to be used interchangeably for different items. So I applied a [conditioning] rule when I was speaking.". Equally, other participants declared having perceived some level of conditioning in the input and having amplified it in their output. For instance, another participant who also

showed a high level of conditioning behaviour and a high input-output conditioning difference, declared: “I tried to work out a pattern for which word was used for which group of things but I don't think there was much of a rule to it. I'd noticed that hap seemed to be used more for vehicles, and nim for animals, so it made sense to me that when I reproduced those phrases, I would always stick with that as my rule.” There was a total of 14 participants (out of 482) who spontaneously provided statements like the former. Despite the low overall number, we must remember that they were not explicitly asked whether they had intentionally added structure to the language, which suggests that this may be more than an anecdotal observation. In addition, as described earlier, the input-output conditioning difference shows similar results, with participants who have a higher value in this measure showing overall higher levels of conditioning, particularly in the production task. These results are consistent with previous observations (e.g., Perfors, 2016) and with accounts of conditioning that describe it as a process for reducing uncertainty and increasing learnability in the face of high cognitive demands (Ferdinand et al., 2019; Hudson Kam & Chang, 2009; Samara et al., 2017); they also show that this process can be conscious.

Taken together, these results suggest that both a drive for reduction of uncertainty and increase of communicative efficiency and general cognitive biases are at play when individuals are presented with a language containing unpredictable variation. Based on these results, we can picture how these pressures interact in a two-step process:

First, when participants are presented with a language, they try to infer its structure. As proposed by the bottleneck theory (Culbertson et al., 2016) the fading nature of language leads to participants making quick inferences on its structure and meaning. At that point, they are sensitive to domain-general biases in statistical processing (Ellis, 2006; Ramscar et al., 2013). Such biases led our participants in the illusory correlation conditions to perceive a conditioning that was not present in their input - when the distribution of the marker was skewed, following our prediction regarding outcome-density bias (Allan & Jenkins, 1980; Alloy & Abrahamson, 1979; Msetfi et al., 2005), and to report that this conditioning was indeed present in their input. Equally, this would explain how in the comprehension task, participants predicted different categories based on the markers, that is, why this bias showed not only in their production, but also when they were predicting the linguistic behaviour of the *native speaker* that had taught them the language.

Second, when participants are asked to produce the language, these perceived differences get accentuated. When participants do not have enough experience with the language, such as those who were presented with the production task first, they are more likely to fully regularise, instead of maintaining the conditioning, in line with numerous studies of regularisation (Ferdinand et al., 2019; Hudson Kam, 2019; Hudson Kam & Newport, 2009; Hudson Kam & Chang, 2009). In contrast, those participants who have more experience with the language further reduce the variability by increasing the level of conditioning, as we see in Smith and Wonnacott (2010) or Kirby et al., (2015). The correlation that we find between the input-output conditioning difference measure and conditioning behaviour both in the comprehension task and in the production task in the illusory correlation conditions goes in line with this account. In addition, having an explicit awareness of the categories led to a higher conditioning, particularly when looking at the measures of output conditioning perception and conditioning in the production task, and it also seems to affect the input-output conditioning difference, which is our implicit measurement of conscious conditioning. This suggests an involvement of *top-down* processes in conditioning: if participants identify animacy categories they can consciously identify that they are conditioning their marker use to them, and amplify this behaviour, leading to a more efficient language. Across different measures, our results suggest that those participants with a better awareness of the language and its structure were more likely to present conditioning behaviour. For example, those participants who were able to accurately identify the skewness in the artificial language also showed a higher level of conditioning. Equally, explicitly stating to have perceived conditioning in the input did not affect the input-output conditioning difference measure, that is, declaring to have perceived conditioning in the input did not affect later reports about the presence or absence of a difference between the level of conditioning in the input and in the output. This further suggests that the effect of illusory correlation over conditioning and the conscious adding of structure are two independent processes. These results also match those in a recent study by Johnson et al. (2020), in which they found that best learners were the ones who added structure to an artificial language.

A similar process could be described for the category accentuation conditions. In this case, participants perceive an existing conditioning, which is heightened by the simple presence of categories (Tajfel & Wilkes, 1963). In line with predictions from Sherman et al. (2009), participants in the skewed animacy conditions perceive a higher level of conditioning, and this is reflected in their results in the comprehension task. When confronted with a higher demand task, like the production task, however, participants in the uniform animacy condition reduce uncertainty by increasing the level of conditioning whereas participants in the skewed animacy condition, seem to split between regularising or conditioning, leading to an overall lower level of conditioning but a higher level of regularisation in this condition. The correlation between the input-output difference measure and conditioning in the production task - but not the comprehension task - in the category accentuation experiment further supports this account. It suggests that conscious conditioning behaviour did not play a role in the comprehension task, in which we found that predicted effect of skewed animacy, but it did play a role in the production task, where we see a surge in conditioning behaviour for participants in the uniform animacy conditioning. Of course, these results are based on correlations between measures in a single study, and further research would be needed to test whether this hypothesis on the mechanism of conditioning holds true. However, we found this explanation to be the best fitting for our results.

The inconsistency in the effect of animacy condition in all our measures is hard to explain. Further analysis in this chapter classifying participants according to the skewness they perceived in the language rather than the actual distribution of categories did not yield any differences between conditions, discarding the potential explanation developed in Chapter 4, and the presence of the predicted effect of animacy category in the comprehension task could well be an artifact of the task. In this task, participants had to choose between two options, an animate and an inanimate image, on each trial. For participants in the skewed animacy conditions, one of the options would always be one of the two items from the minority category. This could have drawn attention to the categories, and as we have seen, an explicit awareness of categories could have been linked to strategic conditioning behaviour. Alternatively, as discussed in Chapter 4, the prior beliefs on the frequency of these categories in the natural world could be behind the lack of effect of category. Unfortunately, the slider and open-ended question results did not help clarify this matter, as we did not find animacy condition to have any effect on any of the measures, other than the output conditioning perception, which was tightly linked to the actual conditioning behaviour observed and discussed in Chapters 4 and 5.

This is, as stated in previous chapters, one of the main limitations of this study, together with the lack of systematic manipulation of task characteristic (prediction vs. production) and task demand. In addition, as discussed earlier, the phrasing of our output perception question was biased towards production. Further studies could include a question directed towards participants' estimations of their performance in the comprehension task. Finally, even though we worked on improving clarity following the pilot studies, the phrasing of some of our open-ended questions seemed to be confusing for a portion of our participants, who gave answers that did not relate to them.

This study is also of course limited in its scope and generalisation of the results. Natural language learning does not normally happen sequentially, and the data we obtained is heavily limited and determined by the material and tasks we used. Future research can expand on these results using paradigms such as the one described in Chapter 3 that better capture more realistic scenarios. This paradigm proposes a modification of the director-matcher task that allows the learning of artificial languages to happen implicitly and through interaction, instead of through direct instruction. That would have allowed us to observe the changes in marker use over the course of language acquisition and gain a better understanding of these processes in a more naturalistic situation. This setting would have been more similar to those in which the first observations in language change appeared and in which language is used while it is being acquired (Senghas et al., 2004; Singleton & Newport, 2004). Alternatively, the results from this study can be complemented with observations from natural languages, as it has been tradition in the field of language evolution (Goldin-Meadow & Mylander, 1983, Goldin-Meadow et al., 1995, Haviland, 2013). However, we believe that the careful design and testing of the materials, the well-controlled and randomised manipulations directed at testing specific hypotheses, the preregistration of the methods and analysis, and our big and diverse sample give this study series of experiments strength and reliability.

Finally, it is important to acknowledge how the discussion in this chapter can lead to question the ecological validity of the results and their relevance for language learning, particularly regarding the involvement of top-down processes, as well the fact that the results are more in line with those in the cognitive domain when it comes to the comprehension task.

As often discussed across this dissertation, language learning, and particularly first language acquisition, is a complex and messy process, where many rules are acquired at once. Could then the results based on adults learning a simplified language in a limited context reflect the reality of the processes of language learning? And more importantly, could these linguistic changes observed in the context of these studies be generalised to the processes of language evolution?

Many of the existing literature on language evolution has used paradigms similar to the one in this study (see for example, Fedzechkina et al., 2012; Feher et al., 2016, 2019; Ferdinand et al., 2019; Hudson Kam, 2019; Hudson Kam & Newport, 2005, 2009; Hudson Kam & Chang, 2009; Samara et al., 2017; Smith et al., 2017), and even if it is clearly acknowledged that language evolution is a complex multilayered process, the results of these studies are used nevertheless to understand and explain some of the phenomena observed in the change of natural language. The addition of measurements of explicit awareness and open-ended questions and the subsequent results, however, could make us question the validity of these paradigms and crucially, the conclusions that are reached in the cited studies. Previous research, such as Perfors (2016), has shown that differences about the experimental setting and participants' understanding of the goal of the study and the context can affect the results.

Of course, a single set of studies, such the ones in this dissertation, and the interpretation of these results, are not enough to question the research methods of the field, but they highlight however the importance of including these additional measures and understanding the cognitive processes that participants undergo when completing the studies and of bearing these in mind when interpreting the results. As proposed earlier, a way of increasing the ecological validity of these studies in relation to the fields of first language acquisition and language evolution would be including implicit learning of the artificial language, rather than explicit instruction.

In addition, even if the findings in Chapter 6 may question the generalisability of the results to first language acquisition, the learning process in this experiment would be more similar to the one in second language acquisition with explicit instruction, where adult learners are being presented with limited subsets of the language in a systematic manner and top-down processes are known to have an important role (see Moskovsky et al., 2015).

In conclusion, even though these are not the first series of experiments trying to combine insights from cognitive science, and particularly associative learning, and psycholinguistics (see Ellis, 2006; Ramscar et al., 2013) our studies and findings show how it is possible to develop and adapt existing paradigms from both fields, and how the interaction between cognitive biases and linguistic theories can help understand the variability we find in the studies in psycholinguistics, improve the existing models and make new predictions.

# Chapter 7: General discussion and conclusion

This dissertation used artificial language learning paradigms in innovative ways to answer questions about the effect of interaction and domain-general biases in language learning, and language change. This chapter summarises the main results and discusses their contribution to the general field of psycholinguistics.

**Summary of the results**

In Chapter 2, we presented a series of three experiments that introduced and tested an adaptation of the director-matcher paradigm that allowed participants to learn an artificial language implicitly, without direct instruction. This paradigm allowed us to test what aspect of interaction helped language learning. We focused in interactivity, defined as defined as the reception of feedback after the production of an utterance, and tested whether it had an effect on language acquisition and found that asking participants guess the meaning of an utterance before being provided with it boosted the accuracy of their semantic learning but did not have an impact on their grammatical learning or the learning of the wordform. The effect persisted in a retest 24-48h later. We showed that simple interactivity, without real interaction, can have long-term effects in language learning. We discussed these results in the context of second language learning, virtual learning research, and drew parallels with some domain-general research findings, such as the generation effect in memory. We also proposed additional uses for this paradigm, which we believe has the potential to help advance research in psycholinguistics, as it allows to closely observe the process of language acquisition through implicit means in a highly controlled environment.

In Chapter 3, we presented a further development on the paradigm in Chapter 3 that allowed for real interaction. In this paradigm, two participants and an experimenter play an interactive game together, which serves as a means for the two participants to learn an artificial language, used for communication during the game. We proposed a study design that manipulated the game setting so participants either learn by observing, by interacting with the experimenter, or by both interacting with the experimenter and observing their peer's interactions. This study also contained an element of unpredictable variation, with the aim of testing how it was treated depending on the modality of learning. We aimed to test the impact of modality in acquisition and in the treatment of unpredictable variation.

From those results, we planned to use the paradigm to test further hypotheses on the effect of communication constraints on how languages change during the process of acquisition through interaction. Unfortunately, due to the constraints imposed by the COVID-19 pandemic we did not collect data for this study, but the chapter discusses the multiple extensions and manipulations that this paradigm allows and how they can contribute to research in the field of psycholinguistics.

Chapters 4, 5, and 6 focused on the effect of statistical learning biases on linguistic conditioning. Chapter 4 tested whether the domain-general biases for illusion of causality and illusory correlation could generalise to the linguistic field. We trained participants in an artificial language containing unpredictable variation (two variants to mark plurality), and a linguistic cue (animacy category) that were not contingent on each other. We manipulated the skewness in the frequency of the markers and of categories. As predicted by the domain-general *outcome-density effect*, when the frequency of the markers was skewed, participants were more likely to condition their marker use to the animacy category. Similarly, Chapter 5 tested the predictions from the category accentuation phenomenon in language acquisition and change. In this study, participants learnt the same artificial language as in Chapter 4, but in this case, the marker use was probabilistically conditioned to categories. We manipulated the skewness of the frequency of the categories, and found that, as predicted, participants in both conditions increased the conditioning degree relative to their input, and that when the animacy frequency distribution was skewed, the average degree of conditioning was higher than when it was uniform. We discussed the potential impact of both of these biases on the patterns we observe in language evolution.

Finally, Chapter 6 revisited the results of Chapters 4 and 5 and contrasted participants linguistic behaviour with their reported perceptions on the statistical frequency of different elements in the language. We showed that the change from the input language to the output language that we observed in Chapters 4 and 5 is likely due to a combination of domain-general biases in perception of the statistical regularities of the languages and a conscious effort to reduce uncertainty.

## Contributions

### *Methodological contributions*

Across this thesis, we used artificial language learning paradigms in novel and flexible ways that allowed us to bridge gaps between disciplines. In Chapters 2 and 3, we presented two extensions of the director-matcher paradigm that allow for implicit learning through interaction. Research in statistical learning has long used paradigms in which participants learn an artificial language implicitly (Aslin et al., 1996; Saffran et al., 1996). However, these studies have been focused on the perception of statistical regularities in the language and have not explored the effect of interaction. Many of those studies that do focus on interaction have relied on either observation (see Newport, 2020 for a review) or have used experiments with natural languages (Anderson & Pembek, 2005; Roseberry et al., 2009, 2015). Syntactic priming studies have used artificial languages or interaction but did not focus on evolution (Weatherlotz et al., 2014). Finally, studies focusing on the effect of interaction in language evolution either come from the field of experimental semiotics and have used communication systems created from scratch by participants (Fay et al., 2010; Garrod et al., 2010; Motamedi et al., 2019), or they start by directly instructing participants in the artificial language (Feher et al., 2016, 2019).

Language acquisition has been described as one of the drivers of language evolution, and interaction as an important means for language acquisition. With the paradigms that we presented in this thesis, we can closely observe and monitor language acquisition through interaction and how it impacts linguistic structure, bringing together different lines of research to advance in the understanding of how these forces interact with each other.

In Chapters 4, 5, and 6, we used a paradigm based on that used in associative learning research (Van Hamme & Wasserman, 1994; Ramscar et al., 2013; Saldaña et al., 2019b) to explore how the biases of statistical learning could affect language change. Importantly, we also used self-report measures based on those in domain-general cognitive sciences to gather participants' perceptions on the linguistic behaviour. This is not the first study in psycholinguistics using self-report measures (i.e., Ferdinand et al., 2018; Samara et al., 2017), but the combination of measures of linguistic behaviour and participants self-report measures played a central role in the discussion of our findings and allowed us to test contrasting theoretical predictions on the drivers of language change. This approach to the collection and analysis of the data also raised important questions on the validity of the methods of other studies in the field. Based on the results of these studies, we believe that the use of measures regarding the explicit perceptions of linguistic behaviour can have the potential to expand our understanding of language change.

### *Interaction and language learning*

The second chapter of the thesis focused on the effect of interactivity in the speed of language learning. It is clear that the main way in which we learn languages is by using them in interactions with other people, even if we are able to learn language through direct instruction or by hearing others use it. However, understanding how interaction affects language learning has been an important question in psycholinguistics for decades. Since Skinner (1957) argued that operant learning and conditioning could explain all linguistic learning, and Chomsky (1959) argued back that an innate grammar was necessary to explain how language was acquired, researchers have found and proposed a myriad of ways in which interactions with others affect language learning, whether interaction was considered the source of all learning (Skinner, 1957), just a mechanism to unlock innate grammar knowledge (Chomsky, 1959), or a combination of both, in which a general ability to understand communicative intention was combined with linguistic input acquired through interaction to construct linguistic knowledge (Tomasello, 2001). The research aimed at testing these accounts has allowed us to learn that interaction does help language learning (Clark, 2018; Kartushina et al., 2022; Weisleder & Fernald, 2013), and identified a multitude of potential mechanisms for this effect, from attention (Ataman-Devrim et al., 2023; Tomasello & Todd, 1983; Pruden et al., 2006), to motivation, to contingent response (Roseberry et al., 2014).

Given the relevance of this question for the instruction of languages, educational research in second language learning has strived to find ways to leverage the positive effect of interaction to improve teaching methods (see Elabdali, 2021 for a review). The findings of this field, often based on the testing of different instruction methods in the classroom, have also helped us understand the processes underlying the positive effect of interaction. In addition, this positive effect of interaction in learning is not exclusive to language, having shown a boost of the learning of other skills (De Felice et al., 2021; Myers et al., 2017).

The mechanisms by which interaction helps language learning seem to be multiple, and the existing methods do not allow for the exploration of the relative contribution of each and how they interact with each other. Our research, with the aim of detangling the different contributors to the effect of interaction in language learning, showed how adding an interactive element to an online learning task boosted the learning of semantics, but not of phonology nor grammar. Our results were also concurrent with research on the "generation effect" in memory (Bertsch et al., 2007) and consolidated in a retest one day later. Aside from the limitations discussed in Chapter 2, and in combination with the methodological innovations described above, we believe that this study has served as a first step to build a better understanding of how interaction shapes the process of language learning, by presenting a paradigm that can be easily expanded to test the effect of different contributors.

Finally, this research has very clear practical implications. On the one hand, the learning from this research can be easily and directly translated to the rising field of virtual language learning, showing an easy way to boost the learning of word meaning. On the other hand, we live in a world where Large Language Models such as ChatGPT are becoming excellent at mimicking human interactions, and in which human-robot interactions will become more common, potentially shaping first and second acquisition processes. We know that interacting with a human and with a robot affects language in a different way (Branigan et al., 2011; Feher et al., 2016; Vollmer et al., 2018), but the mechanisms behind it are not clear. It is therefore vital to gain an understanding of the specific ways in which different types of interactions can affect us.

### *Domain-general biases in acquisition*

The last three chapters of the thesis focused on the potential impact of domain-general biases in language acquisition and language change. The impact of domain-general processes in language acquisition and language evolution is well known; for example, the effect of memory retrieval has been used to explain some of the observations in language evolution (e.g., Hudson Kam, 2019). In addition, statistical learning has been shown to play an important role in language learning (Aslin et al., 1996; Saffran et al., 1996; Gerken, 2005; Gomez & Gerken, 1999; Reeder et al., 2013; Wonnacott et al., 2008)

We know from domain-general research in cognitive science that the perception of statistical regularities in humans is biased. Particularly, we are biased to perceive contingencies and correlations where they are none (Blanco, 2017; Haselton & Nettle, 2007; Matute et al., 2011) and to amplify whichever associations we perceive (Sherman et al., 2009; Tajfel & Wilkes, 1963). In the process of language emergence in the natural world, we see that any unpredictable variation is gradually reduced until reaching a structured language that follows clear grammatical rules. In experimental contexts, researchers have observed how participants sometimes reduce unpredictable variation by imposing rules on when to use different variants, in a process referred to as *conditioning* (Samara et al., 2017). This has been interpreted as an outcome of an effort to reduce uncertainty (Vujovic et al., 2021). However, we wondered if the general bias to perceive illusory relationships and to amplify them could be contributing to language change. To test this, we taught participants a language that contained unpredictable variation and manipulated the frequency of different elements. We found that, as it happens in the domain-general field, a skewness in the frequency of one of the variants predicted linguistic conditioning. Next, we ran an experiment in which we taught participants an artificial language including probabilistic variation of a plural marker and observed that, overall, participants tended to increase the level of conditioning relative to the input. Interestingly, we gathered self-report data on participants' perceived frequencies of different elements in the input language, and in their output language. Even if not all of our hypotheses were confirmed, the combination of linguistic data and self-report data allowed us to gain an understanding of the process of how perceptive biases and the bias for communicative efficiency interact with each other to reduce unpredictable variation in language.

We found evidence that, as predicted, those domain-general biases in statistical learning extended to language. Then, starting from a biased perception, participants showed a tendency to consciously increase the efficiency of the language by reducing variation. These results contribute to our understanding of language acquisition and language evolution in two main ways. First, they show that, though we are good at perceiving statistical relationships, we are not free of bias, and that this shapes the way we perceive linguistic stimuli as much as any other stimuli. This calls for further research on biases in language perception. Second, we show how domain-general processes of perception and language-specific biases for communication are not necessarily opposing explanations for the phenomena we observe, but they interact with each other to produce the changes that we observe in language.

**Scope and future research**

Although this dissertation covers a wide range of literature in the fields that contribute to our understanding of language, its scope does not expand to other important areas of psycholinguistic research, such as that of experimental pragmatics (Clark, 1996; Noveck & Sperber, 2006; Rubio-Fernández, 2023). Aside from the forces for language acquisition and language change that we have discussed, such as interaction or transmission, language happens and is shaped by the communicative context in which it occurs (e.g., Perfors, 2016). In addition, as much as we talk about language evolution processes, none of the studies in this dissertation included transmission (Reali & Griffiths, 2009; Smith & Wonnacott, 2010; Kirby et al., 2014). Future research could look at how the statistical biases we identified work in transmission or how they interact with pragmatic forces. For example, Perfors (2016) showed that pragmatic inferences about variation changed participants' behaviour. Further, Loy and Smith (2019) found that participants alignment behaviour varied depending on their interaction partner's accent. If we told participants that the linguistic input that they receive was produced by another participant with a partial mastery of the language, how would it affect their perception of the language? Would they be motivated to identify non-existent patterns? And to replicate them in their linguistic output? Equally, Rubio-Fernández (2023) proposed that language and communicative social cognition coevolved and that they are acquired in development together through a positive feedback loop. Further manipulations to the paradigm developed in Chapter 3 could help test the predictions of this hypothesis and how pragmatics interacts with other forces, for example, by exploring how children and adults treat unpredictable variation when acquiring a language depending on their assumptions about the other speakers' knowledge.

Furthermore, the research in this dissertation is focused on grammatical variation. We know that the acquisition of sociolinguistic variation is slightly different to that of grammatical variation and has different characteristics, such as the tendency to be probabilistic, rather than categorical (Nardy et al., 2013). Even though the domain-general biases that we describe have been observed in the perception of individuals of different groups (Hamilton & Griffiths, 1976; Tajfel & Wilkes, 1963), the subsequent process of reducing variation could be different in the realms of sociolinguistic variation.

Finally, the methodological advances that we propose open multiple avenues for future research. For example, we believe that the paradigm described in Chapter 3 may allow us to answer a multiplicity of questions about how different forces affect the process of language acquisition, and how this leads to language evolution, offering opportunities to observe the impact of learning through observation or interaction, or from sources of different level of authority on how language is acquired, used, and changed in the process.

## Concluding remarks

This dissertation tries to build bridges between the different subdisciplines studying language learning and language evolution by extending and adapting the existing experimental methods and uses of artificial language learning. Despite the multiple limitations of the studies presented and the disruption in data collection imposed by the pandemic, this work shows how integrating the knowledge from different and often disconnected lines of research can yield productive results in understanding language.

# References

Akhtar, N., Jipson, J., & Callanan, M. A. (2001). Learning words through overhearing. *Child development*, *72*(2), 416-430.

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8*(2), 199-209.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternatives. *Canadian Journal of Experimental Psychology, 34*, 1–11. doi: 10.1037/h0081013

Allan, L. G., & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation, 14*, 381–405. doi: 10.1016/00239690(83)90024-3

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: sadder but wiser? *Journal of Experimental Psychology: General, 108,* 441–485. doi:10.1037/0096-3445.108.4.441

Allport, F. H., & Lepkin, M. (1945). Wartime rumors of waste and special privilege: Why some people believe them. *The Journal of Abnormal and Social Psychology*, *40*(1), 3.

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

Anderson, D. R., & Pempek, T. A. (2005). Television and very young children. *American Behavioral Scientist*, *48*(5), 505–522. https://doi.org/10.1177/0002764204271506

Aslin, R. N., & Newport, E. L. (2012). Statistical Learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*(3), 170–176. https://doi.org/10.1177/0963721412436806

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

Ataman-Devrim, M., Nixon, E., & Quigley, J. (2023). Joint attention episodes during interactions with fathers but not mothers at age 2 years is associated with expressive language at 3 years. *Journal of Experimental Child Psychology, 226,* 105569.

Austin, A. C., Schuler, K. D., Furlong, S., & Newport, E. L. (2022). Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, *18*(3), 249-277.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*(2), 395–418.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.

Barr, R., & Wyss, N. (2008). Reenactment of televised content by 2-year olds: Toddlers use language learned from television to solve a difficult imitation problem. *Infant Behavior and Development*, *31*(4), 696-703.

Barron, A., & Schneider, K. P. (2009). Variational pragmatics: Studying the impact of social factors on language use in interaction. *Intercultural Pragmatics*, *6*(4), 425–442. https://doi.org/10.1515/IPRG.2009.023

Bartlett, F.C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, England: Cambridge University Press

Bentz, C., & Christiansen, M. H. (2013). Linguistic adaptation: The trade-off between case marking and fixed word orders in Germanic and Romance languages. In G. Peng & F. Shi (Eds.), *Eastward flows the great river: Festschrift in honor of Prof. William SY. Wang on his 80th birthday* (pp 48-56). Hong Kong: City University of Hong Kong Press.

Bernard, A., & Onishi, K. H. (2023). Novel phonotactic learning by children and infants: Generalizing syllable-position but not co-occurrence regularities. *Journal of Experimental Child Psychology*, *225*, 105493.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Bickerton, D. (1981) *Roots of language*. Karoma

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and brain sciences*, *7*(2), 173-188.

Bickerton, D., & Givón, T. (1976). Pidginization and syntactic change: From SXV and VSX to SVX. In S. B. Steever, C. A. Walker, & S. Mufwene (Eds.), *Papers from the parasession on diachronic syntax*, (pp. 9–39). Chicago: Chicago Linguistic Society

Blanco, F., Matute, H., & Vadillo, M. A. (2011). Making the uncontrollable seem controllable: The role of action in the illusion of control. *Quarterly Journal of Experimental Psychology*, *64*(7), 1290–1304. https://doi.org/10.1080/17470218.2011.552727

Blanco, F., Matute, H., & Vadillo, M. A. (2013). Interactive effects of the probability of the cue and the probability of the outcome on the overestimation of null contingency. *Learning & Behavior, 41(*4), 333-340. doi: 10.3758/s13420-013- 0108-8

Blanco, F. (2017). Positive and negative implications of the causal illusion. *Consciousness and Cognition, 50,* 56-68. doi: 10.1016/j.concog.2016.08.012

Blanco, F., Gómez-Fortes, B., & Matute, H. (2018). Causal illusions in the service of political attitudes in Spain and the United Kingdom. *Frontiers in psychology, 9*, 1033.

Bott, F. M., Kellen, D., & Klauer, K. C. (2021). Normative accounts of illusory correlations. *Psychological review*, *128*(5), 856.

Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 468.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition, 121*(1), 41-57.

Brooks, P. J., & Kempe, V. (2019). More is more in language learning: Reconsidering the less-is-more hypothesis. *Language Learning, 69*, 13-41.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., ... & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, *9*(2), 378-400.

Brooks, G. (2015). *Dictionary of the British English spelling system* (p. 522). Open Book Publishers.

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69-B77.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review*, *113*(2), 234.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and control*, *2*(2), 137-167.

Chomsky, N. (1965). Persistent topics in linguistic theory. *Diogenes*, *13*(51), 13-20.

Chomsky, N. (1984). On language and culture. *Contrasts: Soviet and American thinkers discuss the future*, 95-101.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. Greenwood Publishing Group.

Chow, J. Y. L., Colagiuri, B., & Livesey, E. J. (2019). Bridging the divide between causal illusions in the laboratory and the real world: the effects of outcome density with a variable continuous outcome. *Cognitive Research: Principles and Implications, 4*(1), 1–15. https://doi.org/10.1186/s41235-018-0149-9

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and brain sciences*, *31*(5), 489-509.

Christiansen, M., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences, 39*, E62. doi:10.1017/S0140525X1500031X

Clark, E. V. (2009). *First language acquisition.* Cambridge University Press.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., & Wilkers-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

Clark, E. V. (2018). Conversation and language acquisition: A pragmatic approach. *Language Learning and Development, 14*(3), 170-185.

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*(2), 120-131.

Collins, D. J., & Shanks, D. R. (2006). Short article: conformity to the power PC theory of causal induction depends on the type of probe question. *Quarterly Journal of Experimental Psychology*, *59*(2), 225-232.

Corneille, O., Huart, J., Becquart, E., & Brédart, S. (2004). When memory shifts toward more typical category exemplars: accentuation effects in the recollection of ethnically ambiguous faces. *Journal of Personality and Social Psychology*, *86*(2), 236.

Costello, F., & Watts, P. (2019). The rationality of illusory correlation. *Psychological review*, *126*(3), 437.

Culbertson, J. (2012). Typological Universals as Reflections of Biased Learning: Evidence from Artificial Language Learning. *Linguistics and Language Compass, 6*(5), 310–329. https://doi.org/10.1002/lnc3.338

Culbertson, J., Jarvinen, H., Haggarty, F., & Smith, K. (2019). Children's sensitivity to phonological and semantic cues during noun class learning: Evidence for a phonological bias. *Language*, *95*(2), 268-293.

Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, *139*, 71–82. https://doi.org/10.1016/j.cognition.2015.02.007

Culbertson, J., & Newport, E. L. (2017). Innovation of word order harmony across development. *Open Mind*, *1*(2), 91-100.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329. https://doi.org/10.1016/j.cognition.2011.10.017

Culbertson, J., Smolensky, P., & Wilson, C. (2013). Cognitive Biases, Linguistic Universals, and Constraint-Based Grammar Learning. *Topics in Cognitive Science*, *5*(3), 392–424. https://doi.org/10.1111/tops.12027

Dąbrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics, 16*(3), 437–474. https://doi.org/10.1515/cogl.2005.16.3.437

Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00852

Dale, R., & Christiansen, M. H. (2004). Active and passive statistical learning: Exploring the role of feedback in artificial grammar learning and language. *Proceedings of the 26th Annual Conferenc eof the Cognitive Science Society*, *September*, 262–267.

De Felice, S., Vigliocco, G., & Hamilton, A. F. D. C. (2021). Social interaction is a catalyst for adult human learning in online contexts. *Current Biology*, *31*(21), 4853-4859.

DeGraff, M. (2007). Kreyòl ayisyen, or haitian creole (Creole French). *Comparative creole syntax: Parallel outlines of*, *18*, 101-126.

DeGraff, M. (Ed.). (1999). *Language creation and language change: Creolization, diachrony, and development*. Bradford Books.

Díaz-Lago, M., Blanco, F., & Matute, H. (2023). Expensive seems better: The price of a non-effective drug modulates its perceived efficacy. *Cognitive Research: Principles and Implications*, *8*(1), 8.

Don, H. J., Worthy, D. A., & Livesey, E. J. (2021). Hearing hooves, thinking zebras: A review of the inverse base-rate effect. *Psychonomic Bulletin & Review*, 1-22.

Elabdali, R. (2021). Are two heads really better than one? A meta-analysis of the L2 learning benefits of collaborative writing. *Journal of Second Language Writing*, *52*, 100788.

Ellis, R. (2001). Introduction: Investigating form-focused instruction. *Language learning*, *51*, 1-46.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied linguistics*, *27*(1), 1-24.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences, 32*(5), 429-448.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science, 34*(3), 351–386. https://doi.org/10.1111/j.1551-6709.2009.01090.x

Fedzechkina, M., Chu, B., & Jaeger, T. F. (2018). Human Information Processing Shapes Language Change. *Psychological Science*, *29*(1), 72–82. https://doi.org/10.1177/0956797617728726

Fedzechkina, M., Hall Hartley, L., & Roberts, G. (2022). Social biases can lead to less communicatively efficient languages. *Language Acquisition*, 1-26.

Fedzechkina, M., & Jaeger, T. F. (2020). Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition*, *196*(October 2019), 104115. https://doi.org/10.1016/j.cognition.2019.104115.

Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences of the United States of America, 109*(44), 17897–17902. https://doi.org/10.1073/pnas.1215776109.

Fedzechkina, M., Newport, E. L., & Jaeger, T. F. (2017). Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive science*, *41*(2), 416-446. doi:10.1111/cogs.12346

Fedzechkina, M., & Roberts, G. (2020). *Learners sacrifice robust communication as a result of a social bias*. https://doi.org/10.31219/osf.io/usfhz

Fehér, O., Ritt, N., & Smith, K. (2019). Asymmetric accommodation during interaction leads to the regularisation of linguistic variants. *Journal of Memory and Language*, *109*, 104036

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158-180.

Fehér, O., Ljubičić, I., Suzuki, K., Okanoya, K., & Tchernichovski, O. (2017). Statistical learning in songbirds: from self-tutoring to song culture. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1711), 20160053.

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, 184, 53-68.

Fernbach, P. M., & Van Boven, L. (2022). False polarization: Cognitive mechanisms and potential solutions. *Current Opinion in Psychology*, *43*, 1-6.

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality and social psychology*, *66*(4), 633.

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology*, *60*(1), 24.

Finley, A., & Penningroth, S. (2015). Online versus in-lab: Pros and cons of an online prospective memory experiment. *Advances in psychology research*, *113*, 135-162.

Floor, P., & Akhtar, N. (2006). Can 18-month-old infants learn words by listening in on conversations? *Infancy*, *9*(3), 327-339.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: a framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 234.

Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2016). Using statistics to learn words and grammatical categories: How high frequency words assist language acquisition. In *38th Annual Meeting of the Cognitive Science Society (CogSci 2016)* (pp. 81-86). Cognitive Science Society.

Galantucci, B. (2009). Experimental semiotics: A new approach for studying communication as a form of joint action. *Topics in Cognitive Science*, *1*(2), 393-410.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in human neuroscience*, *5*, 11.

Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental semiotics. *Language and Linguistics Compass*, *6*(8), 477-493.

Gampe, A., Liebal, K., & Tomasello, M. (2012). Eighteen-month-olds learn novel words through overhearing. *First Language, 32*(3), 385-397.

Gardiner, J. M., & Hampton, J. A. (1985). Semantic memory and the generation effect: Some tests of the lexical activation hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 732.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive science*, *31*(6), 961-987.

Garrod, S., Fay, N., Rogers, S., Walker, B., & Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols? *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *11*(1), 33–50. https://doi.org/10.1075/is.11.1.04gar

Garrod, S., Fay, N., Rogers, S., Walker, B., & Swoboda, N. (2010). Can iterated learning explain the emergence of graphical symbols?. *Interaction Studies*, *11*(1), 33-50.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of child language*, *32*(2), 249-268.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, *23*(5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003

Givón, T. (1985). Language, function and typology. *Journal of Literary Semantics, 14*(2), 83-97. https://doi.org/10.1515/jlse.1985.14.2.83

Goldin-Meadow, S., Brentari, D., Coppola, M., Horton, L., & Senghas, A. (2015). Watching language grow in the manual modality: Nominals, predicates, and handshapes. *Cognition, 136*, 381-395.

Goldin-Meadow, S., & Mylander, C. (1983). Gestural communication in deaf children: Noneffect of parental input on language development. *Science*, *221*(4608), 372-374.

Goldin-Meadow, S., Mylander, C., & Butcher, C. (1995). The resilience of combinatorial structure at the word level: Morphology in self-styled gesture systems. *Cognition*, *56*(3), 195-262.

Goldin-Meadow, S., So, W. C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, *105*(27), 9163-9168.

Goldin-Meadow, S. (2014). How gesture works to change our minds. *Trends in neuroscience and education*, *3*(1), 4-6.

Goldin-Meadow, S. (2020). Discovering the biases children bring to language learning. *Child development perspectives*, *14*(4), 195-201.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109-135.

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87-92.

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

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392-407.

Haselton, M. G., & Buss, D. M. (2000). Error management theory: A new perspective on biases in cross-sex mind reading. *Journal of personality and social psychology, 78*(1), 81.

Haselton, M. G., & Nettle, D. (2006). The paranoid optimist: An integrative evolutionary model of cognitive biases. *Personality and Social Psychology Review, 10*(1), 47–66.

Haviland, J. B. (2013). The emerging grammar of nouns in a first generation sign language: Specification, iconicity, and syntax. *Gesture*, *13*(3), 309-353.

Healey, P. G., Swoboda, N. I. K., Umata, I., & Katagiri, Y. (2002). Graphical representation in graphical dialogue. *International Journal of Human-Computer Studies*, *57*(4), 375-395.

Hiver, P., Al-Hoorie, A. H., Vitta, J. P., & Wu, J. (2021). Engagement in language learning: A systematic review of 20 years of research methods and definitions. *Language Teaching Research*, 13621688211001289.

Hoonhorst, I., Medina, V., Colin, C., Markessis, E., Radeau, M., Deltenre, P., & Serniclaes, W. (2011). Categorical perception of voicing, colors and facial expressions: A developmental study. *Speech Communication*, *53*(3), 417-430.

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 815.

Hudson Kam, C. L. & Newport, L. (2005). Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning & Development, 1*(2), 151-195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59,* 30–66. doi:10.1016/j.cogpsych.2009.01.001

Hudson Kam, C. L. (2019). Reconsidering Retrieval Effects on Adult Regularization of Inconsistent Variation in Language. *Language Learning and Development*, *15*(4), 317–337. https://doi.org/10.1080/15475441.2019.1634575

Iacozza, S., Meyer, A. S., & Lev-Ari, S. (2019). How In-Group Bias Influences the Level of Detail of Speaker-Specific Information Encoded in Novel Lexical Representations. *Journal of Experimental Psychology: Learning Memory and Cognition*. https://doi.org/10.1037/xlm0000765

Ibsen-Jensen, R., Tkadlec, J., Chatterjee, K., & Nowak, M. A. (2018). Language acquisition with communication between learners. Journal of the Royal Society, Interface, 15(140), 20180073. https://doi.org/10.1098/rsif.2018.0073

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1-17.

Johnson, T., Siegelman, N., & Arnon, I. (2020). Individual Differences in Learning Abilities Impact Structure Addition: Better Learners Create More Structured Languages. *Cognitive Science*, *44*(8). https://doi.org/10.1111/cogs.12877

Kartushina, N., Mani, N., Aktan-Erciyes, A., Alaslani, K., Aldrich, N. J., Almohammadi, A., ... & Mayor, J. (2022). COVID-19 first lockdown as a window into language acquisition: associations between caregiver-child activities and vocabulary gains. *Language Development Research*, *2*, 1-36.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. *Simulating the evolution of language*, 121-147.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology, 28*, 108-114

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681-10686.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016

Kocab, A., Senghas, A., & Snedeker, J. (2016). The emergence of temporal language in Nicaraguan Sign Language. *Cognition*, *156*, 147-163.

Kolinsky, R., Navas, A. L., de Paula, F. V., de Brito, N. R., de Medeiros Botecchia, L., Bouton, S., & Serniclaes, W. (2021). The impact of alphabetic literacy on the perception of speech sounds. *Cognition*, *213*, 104687.

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of personality and social psychology, 4*(3), 343-346.

Krishnan, S., Sellars, E., Wood, H., Bishop, D. V., & Watkins, K. E. (2018). The influence of evaluative right/wrong feedback on phonological and semantic processes in word learning. *Royal Society Open Science, 5*(9), 171496.

Krueger, J., & Rothbart, M. (1990). Contrast and accentuation effects in category learning. *Journal of Personality and Social Psychology*, *59*(4), 651.

Krueger, J., Rothbart, M., & Sriram, N. (1989). Category learning and change: Differences in sensitivity to information that enhances or reduces intercategory distinctions. *Journal of Personality and Social Psychology*, *56*(6), 866.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 3.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 830.

Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language, 83,* 152–178. https://doi.org/10.1016/j.jml.2015.03.003

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*, 1-26.

Labov, W. (1986). The social stratification of (r) in New York City department stores. In *Dialect and language variation* (pp. 304-329). Academic Press.

Labov, W. (2006). A sociolinguistic perspective on sociophonetic research. *Journal of phonetics*, *34*(4), 500-515.

Lai, W., Rácz, P., & Roberts, G. (2020). Experience With a Linguistic Variant Affects the Acquisition of Its Sociolinguistic Meaning: An Alien-Language-Learning Experiment. *Cognitive Science*, *44*(4). https://doi.org/10.1111/cogs.12832

Lee, J. C., & Lovibond, P. F. (2021). Individual differences in causal structures inferred during feature negative learning. *Quarterly Journal of Experimental Psychology*, *74*(1), 150-165.

Lenth, R. V., Buerkner, P., Herve, M., Jung, M., Love, J., & Miguez, F. (2022). Emmeans: estimated marginal means, aka least-squares means. *R package version*, *1*(5), 6.

Leung, J. H., & Williams, J. N. (2011). The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition*, *33*(1), 33-55.

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Long, M. H., & Porter, P. A. (1985). Group work, interlanguage talk, and second language acquisition. *TESOL quarterly*, *19*(2), 207-228.

Loy, J. E., & Smith, K. (2019). *Syntactic adaptation depends on perceived linguistic knowledge: native English speakers differentially adapt to native and non-native confederates in dialogue.* PsyArXiv. doi: 10.31234/osf.io/pu2qa

Little, H., Eryılmaz, K., & de Boer, B. (2017). Signal dimensionality and the emergence of combinatorial structure. *Cognition, 168,* 1–15. https://doi.org/10.1016/j.cognition.2017.06.011

Lytle, S. R., & Kuhl, P. K. (2018). Social interaction and language acquisition: Toward a neurobiological view. In E. M. Fernández & H. S. Cairns (Eds.), *The handbook of psycholinguistics* (pp. 615–634). Wiley Blackwell.

Lytle, S. R., Garcia-Sierra, A., & Kuhl, P. K. (2018). Two are better than one: Infant language learning from video improves in the presence of peers. *Proceedings of the National Academy of Sciences, 115*(40), 9859-9866.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77-80.

Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *British Journal of Psychology*, *102*(3), 392-405.

Matute, H., Blanco, F., Yarritu, I., Díaz-Lago, M., Vadillo, M. A., & Barberia, I. (2015). Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology, 6,* 888.

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

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental science*, *11*(1), 122-134.

McArthur, L. Z. (1980). Illusory causation and illusory correlation: Two epistemological accounts. *Personality and Social Psychology Bulletin*, *6*(4), 507-519.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review*, *27*, 1139-1165.

McGarty, C., & Penny, R. E. C. (1988). Categorization, accentuation and social judgement. *British Journal of Social Psychology*, *27*(2), 147-157.

McGarty, C., Haslam, S. A., Turner, J. C., & Oakes, P. J. (1993). Illusory correlation as accentuation of actual intercategory difference: Evidence for the effect with minimal stimulus information. *European Journal of Social Psychology*, *23*(4), 391-410.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*(1), 68.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phones arise spontaneously? *Cognition*, *7*(4), 323-331.

Morais, J., Bertelson, P., Cary, L., & Alegria, J. (1986). Literacy training and speech segmentation. *Cognition*, *24*(1-2), 45-64.

Moskovsky, C., Jiang, G., Libert, A., & Fagan, S. (2015). Bottom-up or top-down: English as a foreign language vocabulary instruction for Chinese university students. *Tesol Quarterly, 49*(2), 256-277.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964.

Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive realism and outcome density bias in contingency judgments: the effect of the context and intertrial interval. *Journal of Experimental Psychology: General, 134*(1), 10-22. doi: 10.1037/0096-3445.134.1.10

Mufwene, S. (2007). Population movements and contacts in language evolution. *Journal of language contact*, *1*(1), 63-92.

Mulligan, N. W., & Peterson, D. J. (2015). The negative testing and negative generation effects are eliminated by delay. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 1014.

Muylle, M., Bernolet, S., & Hartsuiker, R. J. (2021). On the limits of shared syntactic representations: When word order variation blocks priming between an artificial language and Dutch. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(9), 1471.

Myers, L. J., LeWitt, R. B., Gallo, R. E., & Maselli, N. M. (2017). Baby FaceTime: Can toddlers learn from online video chat? *Developmental Science*, *20*(4), e12430.

Newmeyer, F. J. (2008). Universals in syntax. In *Linguistic Review* (Vol. 25, Issues 1–2, pp. 35–82). Walter de Gruyter GmbH. https://doi.org/10.1515/TLIR.2008.002

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. Language sciences, 10(1), 147-172.Newport, E. L. (1999). Reduced input in the acquisition of signed languages: Contributions to the study of creolization In DeGraff M (Ed.), *Language creation and language change: Creolization, diachrony, and development* (pp. 161–178).

Newport, E. L. (2020). Children and Adults as Language Learners: Rules, Variation, and Maturational Change. *Topics in Cognitive Science*, *12*(1), 153–169. https://doi.org/10.1111/tops.12416

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

Nölle, J., & Galantucci, B. (2022). Experimental Semiotics: past, present, and future. In A. M. García & A. Ibañez (Eds.). *The Routledge Handbook of Semiosis and the Brain. Routledge.*

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, *181*, 93-104.

Noveck, I. A. & Sperber, D. (2006). *Experimental Pragmatics*. Basingstoke: Palgrave.

O’Brien, G. E., McCloy, D. R., Kubota, E. C., & Yeatman, J. D. (2018). Reading ability and phoneme categorization. *Scientific Reports*, *8*(1), 16842.

O'Grady, W. (1996). Language acquisition without Universal Grammar: a general nativist proposal for L2 learning. *Second Language Research*, *12*(4), 374-397.

Perales, J. C., & Shanks, D. R. (2008). Driven by power? Probe question and presentation format effects on causal judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1482.

Perales, J. C., Catena, A., Shanks, D. R., & González, J. A. (2005). Dissociation between judgments and outcome-expectancy measures in covariation learning: a signal detection theory approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 1105.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language*, *67*(4), 486-506.

Perfors, A. (2016). Adult regularization of inconsistent input depends on pragmatic factors. *Language Learning and Development*, *12*(2), 138-155.

Philp, J., Adams, R., & Iwashita, N. (2013). *Peer interaction and second language learning*. Routledge.

Pica, T. (1992). Communication with second language learners: What does it reveal about the social and linguistic processes of second language learning. *Georgetown University round table on languages and linguistics*, 435-464.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language learning, 44*(3), 493-527.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2)*,* 169-225. https://doi.org/10.1017/s0140525x04000056

Pruden, S. M., Hirsh-Pasek, K., & Golinkoff, R. M. (2006). The social dimension in language development: A rich history and a new frontier. *The development of social engagement: Neurobiological perspectives*, 118-152.

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect*? Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(6), 1023.

Potts, R., & Shanks, D. R. (2014). *The benefit of generating errors during learning. Journal of Experimental Psychology: General, 143*(2), 644.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of" mouses" in adult speech. *Language*, 760-793.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339-1362.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164. https://doi.org/10.1016/j.cognition.2018.09.010

Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The role of social network structure in the emergence of linguistic structure. *Cognitive Science*, *44*(8), e12876.

Raviv, L., de Heer Kloots, M., & Meyer, A. (2021). What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability. *Cognition*, *210*, 104620.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317-328.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, *66*(1), 30-54.

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(1), 75.

Roberts, G., & Clark, R. (2020). Dispersion, communication, and alignment: an experimental study of the emergence of structure in combinatorial phonology. *Journal of Language Evolution*, *5*(2), 121-139.

Roberts, G., & Fedzechkina, M. (2018). Social biases modulate the loss of redundant forms in the cultural evolution of language. *Cognition, 171*(November 2017), 194–201. https://doi.org/10.1016/j.cognition.2017.11.005

Rodríguez-Ferreiro, J., & Barberia, I. (2021). Believers in pseudoscience present lower evidential criteria. *Scientific Reports, 11*(1), 24352.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1,* 906-914.

Roseberry, S., Hirsh-Pasek, K., & Golinkoff, R. M. (2014). Skype me! Socially contingent interactions help toddlers learn language. *Child development, 85*(3), 956-970.

Roseberry, S., Hirsh-Pasek, K., Parish-Morris, J., & Golinkoff, R. M. (2009). Live action: Can young children learn verbs from video?. *Child development*, *80*(5), 1360-1375.

Ross, D. S. (2001). *Disentangling the nature-nurture interaction in the language acquisition process: Evidence from deaf children of hearing parents exposed to non-native input.* University of Rochester.

Rubio-Fernandez, P. (2023). Cultural Evolutionary Pragmatics: Investigating the Codevelopment and Coevolution of Language and Social Cognition. *Psychological Review*. Advance online publication. https://dx.doi.org/10.1037/rev0000423

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saldaña, C., Kirby, S., Truswell, R., & Smith, K. (2019a). Compositional hierarchical structure evolves through cultural transmission: an experimental study. *Journal of Language Evolution*, *4*(2), 83-107.

Saldaña, C., Loy, J., & Smith, K. (June, 2019b).  Language users can learn probabilistic conditioned variation better when conditioning factors appear with different frequencies [Poster presentation]. *Interdisciplinary Advances on Statistical Learning, Donostia-San Sebastian.*

Saldaña, C., Claidière, N., Fagot, J., & Smith, K. (2022). Probability matching is not the default decision making strategy in human and non-human primates. *Scientific Reports, 12*(1). https://doi.org/10.1038/s41598-022-16983-w

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology, 94,* 85–114. https://doi.org/10.1016/j.cogpsych.2017.02.004

Sankoff, D., & Laberge, S. (1978). The linguistic market and the statistical explanation of variability. *Linguistic variation: Models and methods*, *239*, 250.

Schapiro, A., & Turk-Browne, N. (2015). Statistical Learning. In *Brain Mapping: An Encyclopedic Reference* (Vol. 3, pp. 501–506). Elsevier Inc. https://doi.org/10.1016/B978-0-12-397025-1.00276-1

Schoot, L., Menenti, L., Hagoort, P., & Segaert, K. (2014). A little more conversation–the influence of communicative context on syntactic priming in brain and behavior. *Frontiers in psychology*, *5*, 208.

Schuler, K. D., Yang, C., & Newport, E. L. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, *38*, 2321–2326.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological science*, *12*(4), 323-328.

Senghas, A., Kita, S., & Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science, 305*(5691), 1779-1782.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge University Press.

Sherman, J. W., Kruschke, J. K., Sherman, S. J., Percy, E. J., Petrocelli, J. V., & Conrey, F. R. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of personality and social psychology, 96*(2), 305.

Sinclair, A. J., Ferreira, R., Gašević, D., Lucas, C. G., & Lopez, A. (2019). I wanna talk like you: Speaker adaptation to dialogue style in L2 practice conversation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11626 LNAI*, 257–262. https://doi.org/10.1007/978-3-030-23207-8_48

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive psychology*, *49*(4), 370-407.

Skinner B. F. (1938). *The behavior of organisms: An experimental analysis*. Prentice Hall; Englewood Cliffs, New Jersey.

Skinner, B. F. (1957). *Verbal behavior.* Appleton-Century-Crofts. https://doi.org/10.1037/11256-000

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, *4*(6), 592.

Smith, M. R., & Alpert, G. P. (2007). Explaining police bias: A theory of social conditioning and illusory correlation. *Criminal justice and behavior*, *34*(10), 1262-1283.

Smith, K., & Culbertson, J. (2018). Does learning favour communicative efficiency? *Proceedings of the 12th International Conference on the Evolution of Language (Evolang12),* 1–3. https://doi.org/10.12775/3991-1.115

Smith, K. & Wonnacott, E. (2010). *Eliminating unpredictable variation through iterated learning. Cognition, 113*(6), 444-449.

Smith, K., Fehér, O., & Ritt, N. (2014). Eliminating unpredictable linguistic variation through interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Smith, K., Perfors, A. Feher, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. *Transactions Royal Society B, 372,* 1-12. doi:10.1098/rstb.2016.0051

Smith, K. (2022). How language learning and language use create linguistic structure. *Current Directions in Psychological Science*, *31*(2), 177-186.

Strouse, G. A., & Samson, J. E. (2021). Learning from video: A meta‐analysis of the video deficit in children ages 0 to 6 years. *Child Development*, *92*(1), e20-e38.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgement. *British journal of psychology, 54*(2), 101-114.

Thiessen, E., & Erickson, L. (2015). Perceptual development and statistical learning. *The handbook of language emergence*, 396-414.

Thorne, S. L., & Lantolf, J. P. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press.

Tomasello, M. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development*, *103130*, 103-130.

Tomasello, M. (2000). First Steps toward a Usage-Based Theory of Language Acquisition. *Cognitive Linguistics, 11,* 61-82. doi:10.1515/cogl.2001.012.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Tomasello, M., & Farrar, M. J. (1986). Joint attention and early language. *Child development*, 1454-1463.

Tomasello, M., & Todd, J. (1983). Joint attention and lexical acquisition style. *First language*, *4*(12), 197-211.

Torres, M. N., Barberia, I., & Rodríguez-Ferreiro, J. (2022). Causal illusion in the core of pseudoscientific beliefs: The role of information interpretation and search strategies. *Plos one*, *17*(9), e0272201.

Vadillo, M. A., Musca, S. C., Blanco, F., & Matute, H. (2011). Contrasting cue-density effects in causal and prediction judgments. *Psychonomic Bulletin & Review*, *18*, 110-115.

Van Dessel, P., Ratliff, K., Brannon, S. M., Gawronski, B., & De Houwer, J. (2021). Illusory-correlation effects on implicit and explicit evaluation. *Personality and Social Psychology Bulletin*, *47*(10), 1480-1494.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and motivation, 25*(2), 127-151.

Vicente, L., Blanco, F., & Matute, H. (2023). I want to believe: Prior beliefs influence judgments about the effectiveness of both alternative and scientific medicine. *Judgment and Decision Making*, *18*, e1.

Vollmer, A. L., Read, R., Trippas, D., & Belpaeme, T. (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science robotics, 3*(21), eaat7111.

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

Walton, G. M., Cohen, G. L., Cwir, D., & Spencer, S. J. (2012). Mere belonging: the power of social connections. *Journal of personality and social psychology*, *102*(3), 513.

Weatherholtz, K., Campbell-Kibler, K., & Jaeger, T. F. (2014). Socially-mediated syntactic alignment. *Language Variation and Change, 26*(3), 387–420. https://doi.org/10.1017/S0954394514000155

Weber, K., Christiansen, M. H., Indefrey, P., & Hagoort, P. (2019). Primed From the Start: Syntactic Priming During the First Days of Language Learning. *Language Learning, 69*(1), 198–221. https://doi.org/10.1111/lang.12327

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological science*, *24*(11), 2143-2152.

Williams, J. N. (2005). Learning without awareness. *Studies in second language acquisition*, *27*(2), 269-304.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive psychology*, *56*(3), 165-209.

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

# Appendix A

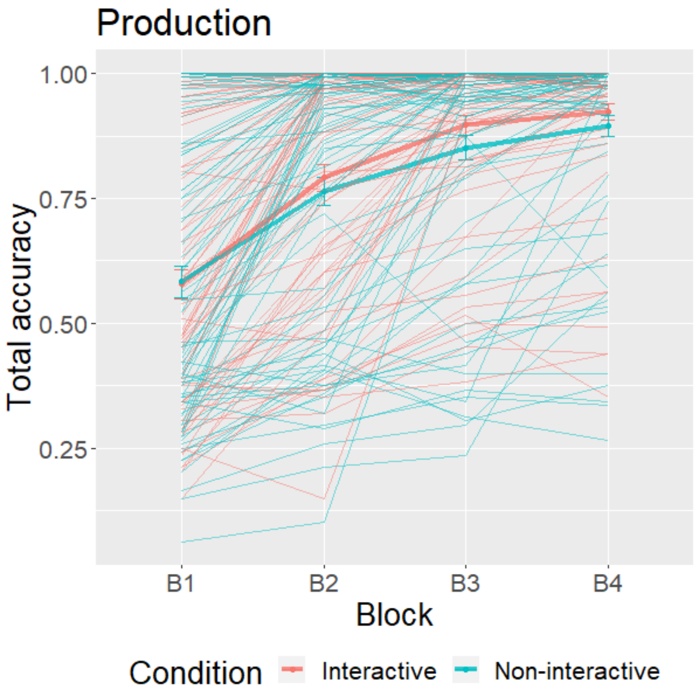
**Main results for Study 3 in Chapter 1 before excluding participants who did not complete Day 2**

**Participants**

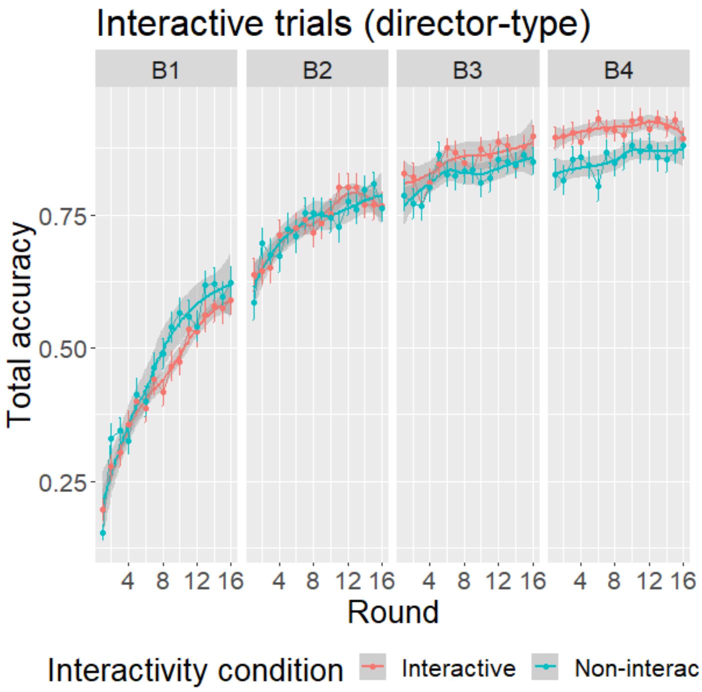
150 undergraduate students took part on Day 1 of the study, 81 in the non-interactive condition and 69 in the interactive condition.

**Results for Day 1**

Total accuracy in the production task by block and interactive condition.



Total accuracy in the training trials by block and interactive condition.



# Appendix B

## Transcription rules

| Transcription | Phonemes in IPA | Examples |
|---|---|---|
| a | ae<br>ʌ<br>ə | c**a**t<br>b**a**t<br>wat**er** |
| b | b | **b**ind |
| ch | tʃ | **ch**at |
| d | d<br>ð | **d**ay<br>**th**is, mo**th**er |
| e | e<br>ɜ: | b**e**d<br>l**ea**rn |
| f | f | **f**an |
| g | g | **g**ive, fla**g** |
| h | h | **h**ello |
| i | ɪ<br>i: | h**i**t<br>h**ea**t, b**ee** |
| j | j<br>dʒ | **y**es, **y**ellow<br>**j**elly, py**j**amas |
| k | k | **c**at, **k**ilo, ba**ck** |
| l | l | **l**eg |
| m | m | **m**elon |
| n | n | **n**o |
| o | ɒ<br>ɔ | h**o**t, r**o**ck<br>c**a**ll, f**ou**r |
| p | p | **p**et, ma**p** |
| r | r | **r**ed, t**r**y |

|  |  |  |
|---|---|---|
| s | s | **s**un, **s**ee |
| sh | ʃ<br>ʒ | **sh**e, cra**sh**<br>plea**s**ure, vi**s**ion |
| t | t | **t**ea, ge**tt**ing |
| th | θ | **th**ink, bo**th** |
| u | ʊ<br>u | p**u**t<br>bl**ue** |
| v | v | ca**v**e |
| w | w | **w**ater |
| ks | ks | e**x**pire, ta**x**i |
| z | z | **z**ebra, la**zy** |
| ai | aɪ | five, eye |
| au | aʊ | n**ow**, **ou**t |
| ei | eɪ | s**ay**, **ei**ght |
| ou | oʊ | g**o**, h**o**me, d**ou**gh |
| oi | ɔɪ | b**oy**, j**oi**n |
| ia | ɪə | h**e**re, n**ea**r |
| iu | ʊə | p**u**re, t**ou**rist |

Examples:

Try → trai // Run → ran // Phoneme → fonim // Sound → saund // Elephant → elefant