An economy of effort in communication as an influencing variable in the outcomes of naturalistic adult second language acquisition

Mohammad AlabdulRazzaq

PhD

University of York

Education

March 2024

Abstract

Stabilising at a limited end-state (“the basic variety”) is common in naturalistic adult second language acquisition. Usage-based theories attribute this to learned attentional biases shaped by L1-tuned processing routines and the low salience, redundancy, and contingency of many grammatical functions. However, research on L2 end-states, adult associative learning, and psycholinguistics suggests an underlying economy of effort: speakers use what works for communication and revise strategies only when necessary. This tendency may influence native and L2 speakers differently as they balance reducing uncertainty with minimising effort. Despite being widely invoked, the concept of economy of effort remains underexplored.

This thesis investigates two questions: (1) Can economy of effort be experimentally operationalised in communicative interaction? (2) Does it affect native and non-native speakers differently? Three studies examined whether task-based interaction success (or failure) prompts increased communicative effort and whether any resulting changes generalise to new contexts.

Study 1 (n=169 monolingual English) normed stimuli and established a method to manipulate communicative effort using abstract figure descriptions. Literal descriptions (e.g., “a large triangle in the middle…”) are more effortful than figurative ones (e.g., “it looks like…”); thus, conservation of effort manifests as a preference for figurative language. Studies 2 and 3 involved online communicative tasks with an artificial interlocutor (researcher confederate). Study 2 (n=90 monolingual English) confirmed a general tendency toward conservation of effort, with communicative breakdowns prompting more effortful strategies only when these differed from prior language experience. Study 3 (n=90 bilingual L2 English) found similar results but with a smaller effect size for non-native speakers.

Table of Contents

Acknowledgements

First and foremost, all praise to Allah the most generous and most merciful. I am thankful to Allah for this opportunity, and I am grateful to be able to complete my thesis. Alhamdo lilah.

Secondly, I would like to express my deepest gratitude to my supervisor Dr. Cylcia Bolibaugh. She has been extremely kind and generous with me in terms of her help support and guidance throughout my PhD journey. I genuinely feel lucky and blessed to have been supervised by her, and I only hope that she continues to be part of academic journey beyond this thesis. Thank you Dr. Cylcia for everything.

Next, I would like to express my thanks and gratitude to my parents. First to my mother Dr. Wafaa Al-Yaseen, you have been a cornerstone in my education from the first words I read until the final words I write today and beyond. Thank you for your prayers, your support, and your love, thank you. Secondly, to my father Abdullatif AlabdulRazzaq. I still remember the days you went without many things to afford a tutor for me, I still remember every time you wiped a tear from my eyes when I was struggling. To both my parents I love you dearly. Thank you.

Finally, and perhaps most importantly, My wife Kholood. You are beyond my soul mate, you have blessed me with happiness, and you have blessed me with love. Thank you for everything, our amazing family and children, the motivation and support to complete this thesis, and a wonderful future to come. I look forward to every tomorrow with you. Thank you for being my world.

Declaration

I declare that this thesis is a presentation of original work, and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

List of tables

List of figures

Glossary of key terms and definitions

Terminology	Definition
Aspirations levels:	Aspirations levels represent the needed minimum level of performance from a potential solution to a problem (Selten, 1999; Simon, 1972; Weiner, 1995). They function as a stop criteria for the time and cognitive resource consuming search for potentially satisficing solutions, or the *process* of satisficing (Simon, 1972, 1990).
Associative learning:	Associative learning is learning about the predictive relationship between a particular cue and an outcome, and surprisal (i.e. surprisal from discrepancy between expected and actual outcome or prediction error) maximally drives learning (Cintrón & Ellis, 2016; Rescorla, 1988; Rescorla & Wagner, 1972).
Basic variety:	It is the most common form of a limited end-state of adult second language acquisition (Ellis, 2008a, 2008b; Long, 1990). It is characterised by the use of mostly lexical open-class linguistic cues (nouns, verbs, adjectives) in communication while lacking closed-class morphosyntactic cues (Ellis, 2008b; Klein, 1998; Klein & Perdue, 1992, 1997).
Bounded rationality:	Conditions of limited and incomplete information, limited cognitive resources, and limited time (Simon, 1972).
Economy of Effort:	It is a universal tendency to minimise total probable work in achieving objectives, manifest in selecting and reusing the path of least effort; this includes the work of searching for and calculating the accuracy of the path of least effort, as a path that requires a longer search and more exhaustive calculation for accuracy is not considered economical, if this added work is not offset by effort saved in selecting said path (Zipf, 1949).
Satisficing:	It is a heuristic process of searching for possible solutions to problems, amenable to trial-and-error, under conditions of bounded rationality (Simon, 1957, 1972, 1990).
Uncertainty:	Uncertainty is the lack of predictability of outcome (Berger & Calabrese, 1974; Kramer, 1999), due to many, equally probable outcomes being possible (Kaan, 2014).

1. Introduction

Not all language learners are equal in their learning outcomes, adult second language learners typically stabilise at a limited end-state that falls short of native speaker norms. With the most common form of this limited end state being a basic variety of the target language when learning is implicit during naturalistic social interaction (Ellis, 2008a; Ellis, 2008b, Ellis 2008c; Klein, 1998; Klein & Perdue, 1997). This outcome is so common it can be considered a fact of adult second language acquisition (SLA) (Long, 1990), as it is a stage of pragmatic, lexical development that nearly all learners develop (Bardovi-Harling, 2000; Ellis, 2008b; Klein & Perdue, 1997). However, since the adult learner under normal circumstances was invariably the child who learned their native language to a native speaker level of proficiency; the difficulty that the same adult faces in learning a second language is in stark contrast to their ability to learn their native language and why this contrast exists is an important undertaking (Dekeyser, 2005).

Although SLA is a complex and multifaceted process, it is the aim of this thesis to explore the notion that the tendency towards an economy of effort, manifest in the construct of an economy of effort in communication, is one potential variable that may play a role in the contrast between the language acquisition abilities of adults and children. Furthermore, this thesis suggests that such a tendency could influence the trajectory of naturalistic usage-based adult SLA towards a limited end-state i.e. a basic variety of the target language, as well as play a role in adults maintaining and remaining at the proficiency level of such a limited end-state.

According to Zipf (1949), the notion of an economy of effort is defined as a universal tendency to minimise total probable work in achieving objectives, manifest in selecting and reusing the path of least effort (i.e. least effortful means of accomplishing goals); this includes the work of searching for and calculating the accuracy of the path of least effort, as a path that requires a longer search and more exhaustive calculation for accuracy is not considered economical, if this added work is not offset by effort saved in selecting said path (Zipf, 1949). In terms of the construct of an economy of effort in communication, the forthcoming review of literature implies that the construct is defined as the collaborative tendency of both speaker and listener to *economise* the effort exerted in reducing *uncertainty* and achieving *mutual understanding* in social communicative interaction in order to achieve or *satisfice* communicative goals.

The potential role of this construct as a variable in the outcomes of naturalistic usage-based adult SLA is evidenced by the interaction of participants in studies of collaborative

interaction (e.g. Bavelas et al., 2000; Bavelas et al., 2017; Pickering & Garrod, 2004, 2006; Schober & Clark, 1989). These studies show that speaker and listener collaborate to reduce uncertainty of meaning which is the lack of predictability of outcome (Berger & Calabrese, 1974; Kramer, 1999), due to many, equally probable outcomes being possible (Kaan, 2014). This reduction in uncertainty is necessary for successful communication (Ramscar et al., 2010).

Participants in these studies achieved this reduction in uncertainty by exerting communicative effort in the form of words, conversational turns and non-verbal gesticulation. Once uncertainty is reduced to the point that communication is successful, participants can then economise their effort in subsequent interaction by reducing the number of words, conversational turns and non-verbal gesticulation. This is due to their achieving of mutual understanding which is the mutual belief that all interlocutors agree and align upon the meaning of certain linguistic units (Bavelas et al., 2017; Pickering & Garrod, 2004, 2006; Schober & Clark, 1989).

Reaching mutual understanding between interlocutors allows them to reuse words and phrases which they have aligned on the use of in subsequent without the need to renegotiate for alignment on different words and phrases through checks for understanding in conversational turns. This reuse of previously found solutions to recurrent communicative problems is known as satisficing which is a heuristic process of searching for possible solutions to problems, amenable to trial-and-error, under conditions of bounded rationality (Simon, 1957, 1972, 1990). The conditions of bounded rationality are those of limited and incomplete information, limited cognitive resources, and limited time in which to make a rational decision about the choice of satisficing solution (Simon, 1972).

As a result of these limiting conditions, the process of satisficing becomes a process of searching for and using the solution that meets expected minimum level of performance, or aspiration level, needed for the problem being addressed (Selten, 1999; Simon, 1972; Weiner, 1995). These aspiration levels function as a stop criteria for the time and cognitive resource consuming search for potentially satisficing solutions, or the process of satisficing (Simon, 1972, 1990). Once a solution has been found to be at least satisficing for specific aspirations and problem spaces, the process of satisficing forgoes the search for different solutions and reapplies previously used solutions for recurring or similar aspirations and problem spaces.

The characteristics of the process of satisficing are remarkably similar to the definition and description of economy of effort in the process of selecting a path of least effort or solution and reusing it. However, the process of satisficing can also explain the process of selecting and maintaining a path of least effort and forgoing the search for a different path through the use of

aspiration levels. This is to say that these characteristics of the process of satisficing and the tendency towards an economy of effort may be co-dependent processes that work together and influence each other. Where it is possible that an economy of effort selects a path of least effort through the process of satisficing.

The reuse of satisficing solutions for the aspiration levels of recurrent problems points to the learning of an association between problem and solution which implicates associative learning as an underpinning of the development of associations between solutions and the outcome of their usage and developing a predictive relationship between solutions and outcomes. Here associative learning becomes the learning about the predictive relationship between a particular cue (e.g. a recurrent problem) and an outcome (i.e. a solution predicted to satisfice) and the surprisal cause when there is a discrepancy between expected and actual outcome (i.e. prediction error) causing the maximal drive of learning (Cintrón & Ellis, 2016; Rescorla, 1988; Rescorla & Wagner, 1972).

This means that the process of satisficing itself, and subsequently an economy of effort due to their similarity, are likely influenced by associative learning in selecting solutions for reuse. According to Ellis (2008b) and Ellis and Sagarra (2010b, 2011), the basic variety is primarily characterised as a means of satisficing mostly through its lexical repertoire (Klein, 1998; Klein & Perdue, 1997; Trudgill, 2002a, 2002b) in spite of its ungrammaticality due to limited morphosyntactic development. Therefore, adults are likely to remain limited in their morphosyntactic development if they are learning to satisfice their communicative needs despite the consequent ungrammaticality of their basic variety, which is further compounded by the influence of an economy of effort disincentivizing the search for more accurate solutions if the current communicative solution is accurate enough. Especially, when a focus on communicative success rather than language form may be a practical necessity (Ramscar & Gitcho, 2007) in a naturalistic context where the nature of the input that adults encounter is both rapid and complex (Christiansen & Charter, 2016).

However, adults are able to leverage their cognitive and neurobiological development and their ability to selectively attend to language input (Arnon & Ramscar, 2012; Birdsong, 2009; Ramscar & Gitcho, 2007) to successfully communicate in this context. This again implicates associative learning as the following review of literature indicates that the phenomenon of selective attention can be learned and directed towards certain aspects of language input due to their salience. Since unlike children who may have no choice but to start with chunks and develop their linguistic productivity through abstraction of components (Boyd & Goldberg, 2009; Goldberg, 2006; Hudson & Newport, 2005; Tomasello, 2003; McCauley &

Christiansen, 2017); adults have been observed to oversegment their input looking for lexical content rather than grammatical structures under implicit learning conditions (Andringa & Curcic, 2015; Isbilen et al., 2017). This indicates that adults may be learning that lexical items satisfice their communicative needs and pay more attention to them in comprehension to manage difficult input, they may in fact be acquiring linguistic cues that block or prevent the acquisition of morphosyntactic cues.

As shown by associative learning research (e.g. Ellis and Sagarra, 2010b, 2011) order of acquisition can negatively influence subsequent acquisition. However, studies of children's associative learning (e.g. Dye & Ramscar, 2009; Ramscar & Yarlett, 2007) show that children recover from these issues implicitly through exposure and prediction error, while adults do not appear to recover as evidenced by their stabilising at a basic variety of the target language. This indicates that despite the abundance of input available to adults (Ellis, 2006, 2007, 2008b; 2008c), implicitly occurring prediction error is not helping adults recover from blocking of linguistic cues due to order of acquisition; however, as evidenced by the aforementioned studies of adult associative, adults appear to have the necessary morphosyntactic cues in memory for such recovery through implicit prediction error.

Therefore, it may be the case that adult learners cannot focus on language form due to a focus on communicative success which is influenced by a tendency towards an economy of effort as evidenced by studies of collaborative interaction (e.g. Bavelas et al., 2000; Bavelas et al., 2017; Cheng & Warren, 1999; Feng, 2022; Önen & İnal, 2019; Ryan, 2015; Schober & Clark, 1989). If adults are unable to focus on language form and are focused on associatively learning how to satisfice their communicative needs, it is possible that an economy of effort is influencing the level granularity at which associative learning trials occur. Meaning that the success or failure of the utterance used by a speaker, or the lexical item attended to in comprehension are the focus of prediction error and not language form. Essentially, language form is not being tracked for prediction error, and as such morphosyntactic cues that may be available in the mind of the learner are not part of associative learning trials that are occurring on the level of lexical items in comprehension or the overall utterance level in production, and taking part in such trials is a caveat for learning in the Rescorla-Wagner (1972) model. The possibility that language form is not being tracked for associative learning is supported by the observation that the basic variety remains effective for communicative interaction and almost only suffers from communicative breakdowns due to lexical gaps in knowledge (Ellis, 2008b; Klein, 1998; Klein & Perdue, 1997) not morphosyntactic gaps, i.e. prediction error from language form discrepancies may not be occurring.

If an economy of effort influences communication and incentives the reusing and maintaining of satisficing means of communicative interaction, inhibiting associative learning trials from occurring on the level of language form. Then it is possible that manipulating an economy of effort can cause a shift in focus to the more granular level of language forms if success cannot be achieved without them. For example, this manipulation could take the form of limiting the success of a learner's utterances that are missing language form. This would increase communicative effort in repair with every instance of missing language form, which does not necessarily occur in naturalistic interactions (Foster & Ohta, 2005), which deprives learners of important instances for learning to potentially occur (Long, 1985, 1996). This would incentivise an economy of effort to mobilise effort in refining the means of satisficing communicative goals, since reusing utterances that depend on mostly lexical items are likely to take more effort in constant repair than searching for a new way of satisficing communicative goals. This could also potentially enhance the dimension of morphosyntactic cues on a respective level since they have become part of the criteria for successful communication, where they may now start to trigger implicit prediction error regarding discrepancies in language form. This would essentially help learners notice the gaps in knowledge that are the sources of errors (Brown, 2007; R. Ellis, 1997; Sabbah, 2015) in the morphosyntax of their productions.

However, the notion of an economy of effort is underexplored as a variable that has potential influence on naturalistic usage-based adult SLA. While evidence of its influence on communication can be observed through the findings of the aforementioned studies of collaborative interaction, this evidence is a by-product of experimental paradigms not specifically designed to capture this influence. Furthermore, studies of collaborative interaction indicate that non-native speakers are observed to be overexplicit in their communication which appears to be at odds with a tendency towards an economy of effort; indicating potential differences in how an economy of effort influences native vs non-native speakers. This means that evidence of an economy of effort influencing communicative interaction, and how this influence may differ between native and non-native speakers is required before evidence of its influence on naturalistic usage-based SLA can be validly collected and interpreted.

Therefore, before designing an experimental paradigm that can operationalise and manipulate the influence of an economy of effort on naturalistic usage-based adult SLA; an experimental paradigm that proves that this notion influences communication in general and that it can be operationalised and manipulated is required first. Furthermore, this paradigm must then be extended to non-native speakers to compare the influence of an economy of effort between them and native speakers to explore the influence of this notion between both

contexts. As such, designing a paradigm that tests if an economy of effort can be operationalised and manipulated and allows for the comparison of its influence between native and non-native speakers is the gap in literature that this thesis aims to address by answering the following research questions.

- Research question1: Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?
- Research question 2: Does an economy of effort influence communicative interaction differently for native and non-native speakers?

Addressing this gap in literature provides an important contribution to the field of research by allowing the notion of an economy of effort to be further developed and explored as a variable that can influence communicative interaction. Thereby, providing an experimental paradigm that can be further adapted to potentially test the influence of an economy of effort on naturalistic usage-based adult second language acquisition that emerges from communicative interaction.

The following section, Section 2 begins the literature review by covering a range of literature that discusses the influence of adults' cognitive abilities and the complex nature of the linguistic input they receive on the granularity and size of the linguistic units they start with in naturalistic usage-based second language acquisition. Section 3 reviews studies of collaborative interaction for both native and non-native speakers and presents evidence for the influence of an economy of effort observed within these studies for both populations and how it differs between them.

Section 4 reviews studies of both adult and child associative learning studies and the implications of their findings for the influence of an economy of effort on adult associative learning. Section 5 concludes the review of literature by expanding upon the notion of an economy of effort and its relationship to the process of satisficing. Additionally, this section addresses the limitations of adapting previous experimental paradigms directly to investigate the influence of an economy of effort on naturalistic usage-based SLA, and the need to establish an experimental paradigm specifically designed to operationalise, measure, and manipulate this notion.

Section 6 outlines the methodology and results of Study 1 which was the norming of the bespoke visual stimuli commissioned for use in Studies 2 and 3. Section 7 outlines the detailed experimental paradigm and methodology of Study 2 designed to both address the aforementioned gap in literature and answer the aforementioned research questions in terms of native speakers. Section 8 outlines an abbreviated experimental paradigm and methodology for

Study 3, as it is identical to Study 2, with the exception of being conducted on non-native speakers.

Section 9 details the procedure of analysis used for studies 2 and 3, and Section 10 provides a summary of the analysis strategy used for both studies. Section 11 presents the detailed results and summary of results of Study 2 for native speakers and Section 12 presents the detailed results and summary of results of Study 3 for non-native speakers.

Section 13 compares the results of native and non-native speakers and discusses the differences in their performance within the experimental paradigm. Finally, Section 14 is the conclusion of this thesis and draws final conclusions in terms of both research questions and discusses the contributions of this work and its limitations.

2. Cognitive and input differences between adults and children

This section aims to cover a range of literature which shows that adults do not go through the natural pattern of language acquisition (Ramscar & Gitcho, 2007) which is the gradual abstraction of linguistic productivity from stored multi word sequences (Arnon & Ramscar, 2012; McCauley & Christiansen, 2017; Ramscar & Gitcho, 2007). This is due to adults significant cognitive domain advantages, neurobiological advantages, and metalinguistic knowledge (Arnon & Ramscar, 2012; Birdsong, 2009; Ramscar & Gitcho, 2007), background world knowledge (Ellis, 2006), and their better ability to recognizing regularities in the input (Boyd & Goldberg, 2009; Hudson & Newport, 2005). Adults leverage these advantages in dealing with significantly different and more demanding linguistic input that forces them into a now or never bottleneck due to the rapid nature of speech (Christiansen & Charter, 2016).

Here input refers to the set of available linguistic cues that the learner is exposed to, and intake is what the learner has taken in from this set of available cues. In the case of speech, the input is the acoustic signal produced by a speaker or, in the case of writing, it is a graphic object produced by a writer (Badger, 2018). And the input adults encounter is significantly different than the input children encounter. Whereas children encounter highly repetitive input that is suited to more easily become intake (Goldberg & Suttle, 2010), adults encounter less repetitive, longer and less prosodically informative input that is less conducive to learning (Fernald et al., 1989; Fisher & Tokura, 1996). However, unlike children adults are able to leverage the totality of their developmental advantages to selectively attend to particular linguistic cues which influences the size of the linguistic units adults are able to discriminate from the input thereby influencing their intake (Arnon & Ramscar, 2012; Ramscar & Gitcho, 2007).

This means that beginner adult learners start with far more over segmented speech when compared to children resulting in a different order of acquisition that results in different learning outcomes (Ramscar et al., 2010). Children have been observed to treat articles and nouns as single inseparable units as opposed to adults which is considered a natural consequence of the differences in nature of input that children and adults encounter (Carroll 1989; Chevrot et al., 2008; MacWhinney, 1978). This is further supported by the findings of Mariscal (2009), and Pine and Lieven (1997), where initially, children would produce an article-noun pairing without using that article with other nouns, indicating that children's speech is under-segmented and the findings of Bannard and Matthews (2008) in a forthcoming review.

Comparatively, studies have shown that adults tend to oversegment received input. In artificial language learning paradigms they segment letter strings into word-like units, looking for lexical content rather than grammatical structures under implicit learning conditions (Isbilen et al., 2017; Andringa & Curcic, 2015). This behaviour indicates a strategic approach to parsing linguistic input, leveraging their pre-existing linguistic knowledge to isolate meaningful components from complex auditory streams. This can be considered a form of selective attention which allows adults to effectively segment and process language input, tailoring their learning experiences to their cognitive abilities and linguistic tasks at hand (Arnon & Ramscar, 2012).

This means that after parsing through the input adults start with word-like units whereas children start with relatively frozen chunks (i.e., they have little productive knowledge of the components of a chunk) (Lieven et al., 1997), and chunks that are not frozen are only so in terms of simple addition or subtraction of single words (Bannard et al., 2009; Lieven et al., 2003; Lieven et al., 2009). This preservation of over form and tendency to use chunks functionally without manipulation is considered to be lexical conservatism, children are able to understand the function of a chunk or multi-word sequence and are able to use it as a functional chunk (Bannard & Matthews, 2008), but are unable to separate chunks into their individual components (Goldberg, 2006; Tomasello, 2003).

Lexical conservatism is a common feature of child language and the input they encounter is similar to the input emphasising verb-specific constraints (i.e. repetitive) (Goldberg & Suttle, 2010). It is not, however, a common feature of the input adults receive nor of their productions. Adult productions are overly-flexible (Fillmore,1979; Pawley & Syder, 1980; Wray, 2002), with little under segmentation (Yorio, 1989), and under or misuse of formulaic language is a common characteristic of adult productions (R.Ellis et al., 2008; Granger, 1998; Wray, 2002, 2004, 2008). This again points to the fact that differences for adults in their starting point, the significantly more complex input they deal with and how they deal with it can and does lead to differences in acquisition outcome when compared to children. However, dealing with input differently than children is not necessarily a conscious choice by adults but a result of the circumstances they find themselves where oversegmenting may be their only option in dealing with the now or never bottleneck (Christiansen & Charter, 2016) nature of input as highlighted by the following quote:

“For example, our language system faces a formidable three-pronged challenge: (a) the speech signal is highly transient (50–100 ms, Remez et al., 2010); (b) normal speech is fast (about 150 words per minute,

Studdert-Kennedy, 1986); and (c) memory for auditory sequences is very limited (between 4 1, Cowan, 2000; and 7 2, Miller, 1956).” (Christiansen, 2019; p427)

This means that for beginner adult learners it may be a practical necessity to leverage their cognitive advantages and undersegment speech in search of word-like units and semantic meaning to satisfice the demands of the communicative context. The following subsection is a review of literature that shows that input is a determining factor in the size of linguistic units learners can discriminate from the input (Wonnacott et al., 2008), that adults tend towards lexical productivity (i.e. generalisation) while children tend towards lexical conservatism even when the input is designed to support the opposite outcome (Bannard & Matthews, 2008; Wonnacott et al., 2008), that the size of the linguistic units that the learner arrives at as intake from input first (i.e. order of acquisition) determines their learning outcomes (Arnon & Ramscar, 2012). This is while of course taking into account that adults' cognitive domain and neurobiological advantages play a significant role in influencing the intake derived from input (Birdsong, 2009; Ramscar & Gitcho, 2007).

2.1 The influence of cognitive development and input characteristics on what becomes intake

As mentioned previously, adults receive linguistic input that is significantly different than the input children receive, and they come to the task of dealing with this input with significant cognitive domain advantages, neurobiological development, meta linguistic knowledge, and background knowledge of the world. The development of cognitive abilities causes a domain general shift towards cognitive control away from children’s imitative unsupervised learning of social and linguistic conventions, giving greater self-direction in adults leading to less conventionalised and more idiosyncratic learning (Ramscar & Gitcho, 2007). Adults are able to self-monitor and select between alternative responses (response conflict processing), goals, and the ability to control their own thoughts (Ramscar & Gitcho, 2007). The result of these increased cognitive abilities is the ability to selectively attend to particular aspects of linguistic input in order to achieve goals they have also selectively attended to (Arnon & Ramscar, 2012; Ramscar & Gitcho, 2007).

As a result, adults have the ability to oversegment the input the receive, whereas children due to their inability to discriminate chunks into their individual components (Boyd & Goldberg, 2009; Goldberg, 2006; Hudson & Newport, 2005; Tomasello, 2003) have no other choice but to start with chunks and develop their linguistic productivity through abstraction of components

(McCauley & Christiansen, 2017). However, given the opportunity adults can and will undersegment if the input they receive supports it (Wonnacott et al., 2008).

In a series of experiments, Wonnacot et al. (2008) aimed to determine if distributional learning mechanisms can acquire types of lexical constraints observed in language acquisition and processing, and how input distribution details affect the balance between applying verb-specific patterns and generalising. Learners across all 3 experiments were adult native speakers of English, experiments 1 and 2 had 14 participants while experiment 3 had 30 participants. Experiment 1 investigated whether participants could acquire verb-specific constraints—that certain verbs are limited to one of two competing constructions—even with the presence of unconstrained 'alternating' verbs that appear in both constructions. This experiment also explored if the learning of these constraints is influenced by the frequency of the verbs, mirroring the modulation seen in natural language learning and processing. Experiment 2 sought to see if adjusting the distribution of verb types across the language could influence the lean towards generalisation over verb-specific constraints. By introducing a language with a broader and more varied class of alternating verbs than in Experiment 1, it examined whether this diversity prompts learners to apply alternating frames more broadly, even to verbs previously understood as constrained. Experiment 3 exposed to languages where both verb-specific and verb-general patterns are probabilistic, examining the effect of these statistics on various language behaviours. This experiment questioned what occurs when these distinct statistics conflict and how such conflict is influenced by the overall distributional properties of the language.

The experiments collectively aimed to assess whether learners can navigate the complex landscape of lexical constraints, differentiating between verbs that are strictly bound to specific constructions and those that can alternate, based solely on the statistical distribution in the linguistic input. Their findings indicated that participants were able to track both verb-specific statistics and verb-general statistics under the condition that the manipulation of frequency favoured that type of verb class, otherwise participants would ignore that verb class if the manipulation did not favour it. This means that while adults are able to leverage their developmental advantages, the nature of the input is also an important factor in influencing learning outcomes, where different types of input lead to different types of intake.

This resulted in learners showing signs of lexical conservatism and overgeneralized to a much lesser degree when the artificial language used was repetitive and verbs were more constrained to specific argument structures meaning that these verb-specific statistics were treated as inseparable multi word units similar to how children would acquire chunks. However, when the

experimental conditions were manipulated to favour verb-general statistics, that is the alternating verb class was much larger than the class of verbs associated with only one argument structure, the results were inverted.

Although Wonnacott et al. (2008) intended to use the adult participants of these experiments as proxies for children learning their native language, the fact that they are adults bringing their developmental advantages to this task cannot be discounted. As Onnis (2012) notes it is common practice for research to treat adult participants as approximations of children in artificial language paradigms, the differences in initial state previously noted make the extension of these results to child language learning tenuous. However, Onnis (2012) also notes that adults in such experiments can be considered approximations of adults learning a second language as artificial language paradigms are designed to control for previous experience, adult participants in these experimental paradigms can be thought of as learning a second language due to them already being proficient speakers of their native language.

The results of Wonnacott et al. (2008) show that differences in input can and do lead to differences in intake. As highlighted by the findings of Bannard and Matthews (2008) when children encounter language input that emphasises verb-specific constraints (i.e. repetitive) (Goldberg & Suttle, 2010), they become lexically conservative and use chunks functionally rather than productively.

Bannard and Matthews (2008) tested the ability of children to acquire multi word sequences under conditions similar to those emphasising verb-specific constraints. The logic of their research was that children are exposed to multi word sequences (2-5 word utterances) as frequently as single word utterances, in addition to some of the most frequent multi word sequences being as frequent as their component words (e.g. a cup of tea vs cup, tea) (Bannard & Matthews, 2008). Their participants were 38 normally developing monolingual English children 17 2-year-olds (mean age 2 years 6 months) and 21 3-year-olds (mean age 3 years 4 months). Using corpus extracted materials, they tested children's ability to repeat either high or low frequency multi word sequences. Their findings were that children were more accurate and faster at repeating high rather than low frequency multi word sequences, with older children being both more accurate and faster than younger children.

Bannard and Matthews (2008) noted that these results were indicative of whole form storage due to the gains in accuracy and processing speeds observed as a result of exposure to high frequency multi word sequences. Their findings also point to the effects of input type, specifically verb-specific constraints leading to lexical conservatism which is a common characteristic of child language productions (Goldberg & Suttle, 2010). This is supported by the

observations that children's chunks are relatively frozen (i.e. they have little productive knowledge of the components of a chunk) (Lieven et al., 1997), and chunks that are not frozen are only so in terms of simple addition or subtraction of single words (Bannard et al., 2009; Lieven et al., 2003; Lieven et al., 2009). Therefore, children are able to understand the function of a chunk or multi word sequence and are able to use it as a functional chunk (Bannard & Matthews, 2008), but their lexical conservatism is due to their inability to separate chunks into their individual components (Goldberg, 2006; Tomasello, 2003). The notion that lexical conservatism is due to the inability to discriminate a chunk or multi word sequence into its individual components is supported by findings that suggest that children are simply not as good as adults at recognizing regularities in the input (Boyd & Goldberg, 2009; Hudson & Newport, 2005).

However, children do move on from lexical conservatism and frozen chunks to linguistic productivity and the gradual extraction of single lexical items from stored chunks (McCauley & Christiansen, 2017), as their cognitive abilities develop into adulthood through utilising their cognitive domain advantages which influences the size of linguistic units adults are able to discriminate from the input (Arnon & Ramscar, 2012; Ramscar & Gitcho, 2007). This brings the focus back to the fact that learner abilities also play an important role in what is derived as intake from input and not only the nature of input itself. This is a key tenant of a constructionist account of language acquisition is that language learning is based on the intake derived from the positive (i.e. available) input, and what is derived from the input is affected by domain general cognitive processes, namely attentional-biases, principles of cooperative communication, general processing demands, and processes of categorization (Goldberg & Suttle, 2010). In addition to these cognitive advantages, adults are also better at recognising regularities in the input when compared to children (Boyd & Goldberg, 2009; Hudson & Newport, 2005). These factors combined all point to the ability of adults to deal with input that is more complex than the input children receive in a vastly more sophisticated manner.

However, based on the experimental design of Wonnacott et al. (2008) it does not appear that adults consciously decide between selectively attending to chunks or lexical items or verb-specific vs verb-general constraints. I.e., they do not appear to leverage their overall developmental advantages at will but rather as a response to the nature of the input itself. In a naturalistic SLA context, it may however be a practical necessity to leverage these advantages in order to deal with input that is rapid and less suited for induction (Christiansen & Charter, 2016; Ellis, 2008b). The leveraging of these advantages and selectively attending to specific aspects in linguistic input such as the search for word-like lexical units to achieve

communicative success can lead adults to oversegmenting the input they receive and starting with single lexical items rather than chunks or stored multi word sequences which has implications for the ability of adults to acquire a second language to the level of native speaker proficiency (Arnon & Ramscar, 2012).

This means that the input adults receive may not allow them to go through the native pattern of acquisition or the gradual abstraction of linguistic productivity (McCauley & Christiansen, 2017). This can influence the adult learner's ability to predictively process input (Ramscar & Gitcho, 2007), as the differences in predictive processing between natives and non-natives are due to differences in language experience not differences in the mechanism that underlies predictive processing (Kaan, 2014). This means that whereas native speakers learned language as children through the process of abstracting linguistic productivity gradually from stored multi word sequences, adult learners can skip this process and start with single lexical items; resulting in native speakers treating article+noun pairings as more cohesive in predictive processing as a consequence of how they abstracted them from stored multi word sequences as children, while non-native adult learners do not as a consequence of starting with single lexical items via selectively attend to particular aspects of the input (Arnon & Ramscar, 2012; Ramscar & Gitcho, 2007).

Essentially, there is a dichotomy of outcomes between starting with stored multi word sequences and starting with single lexical items when it comes to achieving native like proficiency. Arnon and Ramscar (2012) compared the learning outcomes of adults learning the predictive relationship between articles (grammatical gender) and nouns when controlling for the order of acquisition of differing unit sizes (i.e. article and noun first vs noun only first). As the study focused on the predictive relationship between articles and nouns the authors adopted the use of the Rescorla-Wagner (R-W) model (1972) to formally examine the effects of order of exposure to different unit size on language acquisition. The R-W model simulates learning as changes in associative strength between individual cues and outcomes as the result of discrete learning trials, and error in this model is the result of failed prediction (Arnon & Ramscar 2012). The results of this study showed that when both their simulations and participants were exposed to less segmented sequences first conditions (article + noun) they were better able to use articles in prediction of the noun associated with that article (i.e. the grammatical gender of the article allowed the prediction of the subsequent noun), and vice versa, the results of noun first showed that participants were not using articles to predict nouns.

The dichotomy of outcome between sequence first and noun first conditions in this study support the notion that differences between the learning outcomes of children and adults is one

based on differences in input, and how their differences in initial state enables them to derive intake from the input. The development of cognitive abilities causes a domain general shift from children's imitative unsupervised learning of social and linguistic conventions to cognitive control, giving greater self-direction in adults leading to less conventionalised learning and more idiosyncratic (Ramscar & Gitcho, 2007). Adults are able to self-monitor and select between alternative responses (response conflict processing), goals, and the ability to control their own thoughts (Ramscar & Gitcho, 2007). The result of these increased cognitive abilities is the ability to selectively attend to particular aspects of linguistic input to achieve goals that have been selectively attended to (Ramscar & Gitcho, 2007). One of the findings of Arnon and Ramscar (2012) from participants learning under noun first conditions which simulate selectively attending to particular aspects of the linguistic input was the blocking of the acquisition of articles. Blocking is a statistical outcome of learning to reduce uncertainty of outcome, once a certain outcome can be fully predicted by a cue learning about additional cues becomes unnecessary (Arnon & Ramscar, 2012).

The results of Arnon and Ramscar (2012), and the differences in input, initial state, and the effects of cognitive development on adults ability to self-monitor and selectively attend to particular aspects of input and particular goals (Ramscar & Gitcho, 2007) all point to associative learning and the phenomenon it subsumes (e.g. selective attention and blocking) as being part of the cause of a limited end-state outcome for adult SLA. However, as it is the aim of this paper to propose that the construct of economy of effort is an additional factor that influences the outcome of adult SLA, the following section covers a range of studies of adult collaborative interaction that feature native and non-native speakers. This review provides insight into the incidental evidence of an economy of effort, its characteristics in communication for both native and non-native speakers, and its implications for associative learning before discussing associative learning and the overall implications of associative learning and an economy of effort for naturalistic usage-based adult SLA.

3. Studies of collaborative interaction for native and non-native speakers

The aim of this section is to review a series of studies that although not designed to explicitly capture the influence of an economy of effort in communication, do in fact show evidence of the potential role of an economy of effort in communication and the modulation of effort in the attempts of participants to successfully convey meaning by the reduction of uncertainty. The forthcoming review of native speaker focused studies (Bavelas et al., 2000; Bavelas et al., 2017; Schober & Clark, 1989) shows that this modulation of effort in the reduction of uncertainty results in a potentially positive correlation between an economy of effort and uncertainty, as uncertainty is reduced so is the amount of effort needed over the course of developing mutual agreement on meaning. While the review of non-native speaker focused studies (Cheng & Warren, 1999; Feng, 2022; Önen & İnal, 2019; Ryan, 2015) show that the modulation of effort may in fact result in a negative correlation between an economy of effort and uncertainty, where non-native speakers focus on success (Ryan, 2015) and err on the side of over explicitness; i.e. expending more effort per utterance or turn to avoid the overall more effortful breakdown and repair cycle which entails attending to and interpreting the listeners call for clarification, formulating and articulating a corrected response and then attending to the listener for an indication of understanding and acceptance of the correction before proceeding with the communicative interaction. Essentially, this cycle represents the act of doing mutual understanding through the recalibration of meaning as further defined below.

Throughout the review evidence is presented that adult speakers are not autonomous speakers and listeners in communication but rather they are collaborative co-participants in the development of dialogue (Schober & Clark, 1989) and that they are susceptible to each other's contributions to conversation, narration, and discourse in general (Bavelas et al., 2000; Schober & Clark, 1989). Furthermore, these studies indicate that the correlation between effort and uncertainty is quantifiable in terms of the number of turns and words used to achieve communicative success, although it must be stressed that this correlation appears to move in different directions for native and non-native speakers.

3.1 Studies of native speaker collaborative interaction

The studies of native speaker communicative collaboration show that the positive correlation in the reduction of effort in relation to the reduction of uncertainty is an automatic process

that favours an overall reduction in the number of words and utterances used in communication when uncertainty is successfully reduced through the processes of alignment and calibration of mutual understanding. Where alignment is a simple, automatic process by which interlocutors reach a mutual understanding through reusing linguistic representations (i.e. what was said), manifest in the reuse of words and grammatical structures, and allows for deictic and elliptical reference to previously used linguistic representations (Pickering & Garrod, 2004, 2006). Mutual understanding is seen as a means of indicating understanding of dialogue between interlocutors and is considered to be an active process undertaken by interlocutors (Bavelas et al., 2017). The process of doing mutual understanding is defined by Bavelas et al. (2017) as a micro calibration process that is an automatic form of alignment. Where new information introduced by the speaker in utterance [A] is acknowledged by the listener in a [B] backchannel response, and the speaker acknowledges this acknowledgement with a final [C] response to indicate to the listener that their understanding of the utterance was sufficient.

Bavelas et al. (2017) analysed conversational data from video archives of previous studies for evidence of this calibration process. The data set yielded 2128 usable utterances. They outlined 15 communicative functions that should, could, or are unlikely to initiate a calibration sequence highlighted in Figure 1. With the criteria for “should” functions being utterances that introduce new information to the dialogue, 1175 fit this criteria. 97% of these 1175 utterances were completed with 3 step calibrations, 74% of 127 “could” function utterances were treated as introducing new information and completed with 3 step calibrations. Their findings indicated that calibration was efficient, as utterances could play the role of multiple steps in the process (i.e. an utterance can function as an [A, B, C] utterance simultaneously), with 62% of utterances serving more than one role in the calibration process. It was also found that the calibration process was continuous and cumulative as 64% of 1175 A initiations included deictic or elliptical reference to previously introduced information.

Table 1
Six utterance functions that conveyed new information and SHOULD therefore initiate a 3-step calibration sequence.

Utterance function	Operational definition	Examples
CONTRIBUTING new topical content	An utterance that functions to present topical content that is new to this particular conversation (i.e., presenting facts, opinion, or descriptions; characterizing, specifying, elaborating).	• I'm from Kelowna • I'm going into commerce • I'm just in first year, I don't know (furrows brow) any what I am doing
REQUESTING new topical content	An utterance that functions to invite new topical content that has not yet come up in the conversation.	• Where are you from? • What year are you- are you in first year? • How was that?
PROPOSING something	An utterance that functions to manage the conversation at a meta-level, above the level of topical content (e.g., to start the conversation, end the conversation, shift topics, end current topic, change speakers). PROPOSING can be a familiar discourse shift marker such as "anyways" or "so, yeah", often left hanging prosodically.	• Do you think we're done? • Should we move on? • But. . . anyways. . . (often an implicit way to shift the topic) • You wanna go first?
ALERTING that repair is needed	An utterance that functions to signal that there is some trouble in hearing or understanding. There are two different occasions for alerting: Alerting that *I did not hear (or understand) you.* The person who is alerting indicates that he or she needs to hear the utterance again or needs a rephrased version because it was not understandable in its current form. Alerting that *you did not understand me.* In this case, the person doing the alerting is saying that he or she has been misunderstood and will usually repeat or rephrase his or her contribution so that the other person will understand.	• Pardon? • Sorry?
RE-INTRODUCING information	An utterance that functions to invite confirmation that this information was previously calibrated. It initiates a topic shift to the information being re-introduced.	After calibrating on the fact that the boyfriend of one of the participant's lives in Kelowna, they talk about other things. Then later, the other participant says "And your boyfriend lives in Kelowna right now?" as an introduction to suggesting which university she thinks her interlocutor could consider applying to.
SCRIPTING initiating a script	An utterance that functions as an invitation to start a conversational routine or script without inviting or contributing new topical content. These should have a predictable or stereotypical response.	• How's it goin' (predictable response = some variant of "fine") • Nice to meet you (predictable response = some variant of "you, too")

Figure 1: Communicative functions that should initiate calibrations (Bavelas et al., 2017: p100)

The finding that calibration is an automatic and cumulative process that also allows for deictic or elliptical reference is an indication of an underlying learning process and has the potential to be subject to associative learning. The findings and results of Schober and Clark (1989) provide evidence of such cumulative learning, and the automatic, opportunistic nature of calibration and alignment.

In their experiment, Schober and Clark (1989) had 10 pairs of students who could not see each other play the roles of director and matcher in a matching game where the director informs the match which order to place 12 figures in. This study was designed to test the ability of overhearers (listeners not involved in the communicative exchange) to correctly accomplish the same task as the matcher; however, the results of interest for this review of literature was the change in number and size of conversational turns taken.

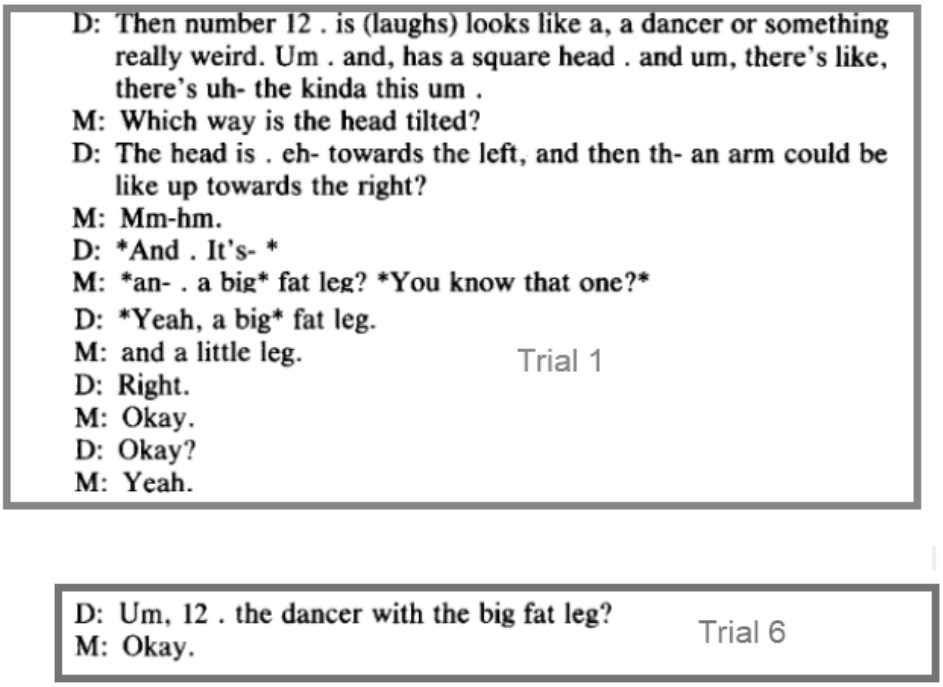


Figure 2: Sample exchange (Schober & Clark, 1989; p216-217)

As illustrated by the above trials in Figure 2, over the course of six trials the number of words per turn and the number of turns decreased markedly. The average number of turns for directors and matchers needed to organise each figure decreased from 8 turns to 1 turn; while the number of words used decreased from an average of 73 words to 13 per turn for the director and from an average of 39 words to 3 per turn for the matcher, additionally the time needed for place each figure dropped significantly from 39 seconds to 6 seconds by the last trial (Schober & Clark, 1989). The quantitative decrease in the number of turns and words per turn exchanged in addition to the qualitative changes in communication highlighted by Figure 2 demonstrate the role of uncertainty in the construct of economy of effort. Due to the abstract nature of the figures used by Schober and Clark (1989) in this experiment, directors expended more effort in early trials trying to explain the features of each figure in order to discriminate them from each other and selecting, according to them, the most discriminating feature to reduce uncertainty for the listener. To some degree this is error driven learning between features and lexical outcomes akin to the wugs and nizes experiments of Ramscar et al. (2010) featured in the forthcoming section on associative learning. The exchanges between participants taking part in this experiment indicate that speakers and listeners are indeed collaborating towards achieving mutual understanding, and that listeners are helping speakers learn what enables them to successfully convey their intended meaning, providing collaborative assistance through suggestions and indication of understanding. As the director and matcher in Figure 2 collaborate

to narrow down the features of the figure they need to place they both learn the predictive value of “dancer with a big fat leg” as a set of features to reliably predict the figure they need. Finally, by the end of the experiment, one single turn is taken, and the matcher indicates their understanding with a single word “okay”.

The implications of this exchange is that it is possible it is in the interest of the listener to indicate understanding to the speaker in order to conserve their own effort in attending to the extended utterances of the speaker that is doing their best to resolve potential uncertainty. In other words, it does appear to confirm that the construct of an economy of effort in communication is indeed a collaborative tendency of both speaker and listener to modulate the effort exerted in reducing uncertainty and achieving mutual understanding in social communicative interaction in order to achieve or satisfice communicative goals. The speaker exerts effort in reducing uncertainty and is sensitive cues from the listener that they have understood, and modulates their effort accordingly, while the listener attends until they are no longer uncertain and indicate their understanding to conserve their own effort in listening.

Although Zipf (1949) posits that in communication the economy of effort of the speaker (the force of unification) is in conflict with that of the listener (Force of diversification). However, this view does not take into account the evidence put forth on the collaborative nature of an economy of effort in communication and that the roles of speaker and listener are interchangeable, that just as a speaker would wish to reduce their effort in maintaining and articulating a large vocabulary, the same applies to that person when they are the listener, in the form of minimising effort in listening to and comprehending utterances. As such a true economy of effort for a single individual that fills both roles in a conversation would lie in conveying and receiving information with as little effort as possible. That is to say that both speaker and listener would prefer the force of unification manifested in maintaining and using as small a vocabulary as possible.

Therefore, when interlocutors are successful in communication, we see a marked decrease in the number of conversational turns and words used to convey meaning in a positive correlation with the required reduction of uncertainty, relative to the increase in cumulative alignment and mutual understanding. However, by the same token, there should be a marked increase in effort in the form of more turns and words used to convey meaning when there is a failure to calibrate mutual understanding indicated by the speaker. Evidence from the experiment of Bavelas et al. (2000) supports the notion that speakers depend on the indications of understanding by the listeners in modulating the amount of effort they exert, and that more effort is exerted when speakers interpret lack of understanding.

In their study Bavelas et al. (2000) compared the effects of attentive and distracted listeners on narrative quality, the results found a significant deterioration of the quality of a speaker's narrative when paired with a distracted listener. Testing the effects of listeners' contributions on the quality of narrative storytelling, Bavelas et al. (2000) had participants take turns in telling close call stories. The experiment proposes listener back channels or responses the listener makes while listening contribute to and affect the quality of the speakers narration. These responses were split into two categories, the first category was generic responses consisting mostly of continuers such as "mhm" "uh-huh" and "yeah", these responses do not contain narrative specific information, but serve to indicate comprehension of the listener which the speaker tracks and makes corrections when necessary (Bavelas et al., 2000). The second category was specific responses, these responses do contain narrative specific responses, and are tied to the moment-by-moment changes in the narration, they occur later in the listening process (Bavelas et al., 2000) and possibly a deeper understanding by the listener and great reduction of uncertainty. The experiment controlled for listeners' abilities to make generic and specific back-channel responses by splitting listeners into different experimental conditions. Listeners were asked to either only listen, listen in order to summarise the story, listen in order to be able to retell the story in as much detail as possible, or to count words that begin with T. The results of the experiment found that specific responses by listeners occurred significantly later than generic responses, indicating that specific responses occurred as the listeners had more knowledge and less uncertainty about the narrative. Furthermore, the experimental condition of counting T-words was found to be detrimental to the production of both generic responses (80% less than other conditions) and specific responses (95% less than other conditions) (Bavelas et al., 2000).

In terms of the effects of the experimental conditions of the listeners on the quality of narration, these effects were analysed on the scales of pace, continued elaboration, disfluency and noticeable gaps, and justification of danger in a close call story (Bavelas et al., 2000). It was found that when listeners were counting T-words their reduced generic responses and almost lack of specific responses negatively affected the quality of narration. Narrators were more likely to elaborate, pause, self-repeat, attempt to justify why their story was dangerous and a close call, and when they fail to make the story relevant to the listener, they abruptly end the story. This was only the case however, after the point where specific responses began to emerge in the other experimental listening conditions, in other words the decline in narration quality occurred at the point where the speaker was expecting more specific rather than generic responses. In other words, the negative qualities mentioned began to appear in the narration of

the story teller in response to perceived uncertainty as the speaker's potential prediction of specific responses indicating understanding did not occur. In response to this failed prediction of outcome, it is possible that narrators began to expend more effort in conveying the gravity of the close call situation through elaboration, justification, and abruptly ending the story when they realised their efforts were in vain. The following example in Figure 3 from (Bavelas et al., 2000; p949) illustrates these potential points of exerted effort in repairing the communicative breakdown.

So this tree's falling, falling, falling. And he was ahead of me, and I was behind him, and *just* the end of the tree clipped my foot. And it felt like, like a *whip* hitting my foot. And so ah after I, I mean, I saw it fall and we both go *diving* into the thing cause we knew—I mean, I don't know how exciting that is but afterwards, ah, I mean, we chuckled about it at lunch. Cause it's always funny if you don't get landed on, sure it was a hoot, but (stylized laugh). Um. I just thought that was, ah, that was funny that, ah. Like *usually*, the easy way to go out is go to either side, and that way it'll fall and you're on either side. But since we had no escape route, we knew it was comin' at us, so we had to run for our lives basically, which puts a little excitement into the job too, cause it's fun, rappelling down trees and stuff and, and what-not. So . . . that's all!

Figure 3: Sample narrative breakdown (Bavelas et al., 2000; p949)

At this point the narrator had explained that they were stuck in a narrow corridor and the tree they had chopped down was falling on them, and with no choice but to try and outrun the falling tree. Bavelas et al. (2000) note that throughout the story the T-word counting listener only nodded and occasionally smiled, and that the climax of the story was “like a whip hitting my foot”. At this point it is possible that generic responses to the climax of the story was not the outcome predicted by the narrator. If through the process of associative learning and from previous experience of communicative social interaction the speaker had learned that specific responses were indication of understanding; then the failure of this predicted outcome to occur is potentially interpreted as a communicative breakdown. As illustrated above, the speaker begins to elaborate, while repeating themselves and pausing. Adding the phrase “I don’t know how exciting that was” a possible indication of their story failing to be as exciting as they had predicted, justifying the danger of the story by emphasising and explaining that they had no choice but to outrun the tree (which is a possible indication of the part of the story the listener might have misunderstood). Finally, the narrator reiterates the excitement in running for their lives and abruptly ends the story with the paused phrase “So . . .that’s all!” indicating a possible

final attempt to reduce the uncertainty surrounding the excitement of the story before deciding it was not worth the added effort and giving up.

The speaker in this trial exerts a substantial and quantifiable amount of effort in an attempt to repair the perceived communicative breakdown. This effort was manifest in the significantly large number of words used across multiple elaborations and justifications after the climax of the story, before ultimately giving up their attempt to repair the communicative breakdown. This is an important example of the significant amount of effort exerted by people in their attempts to find satisficing solutions for their communicative aspirations.

The failure of the listener to continue with the calibration of mutual understanding can be seen as a form of other-initiated repair, as the speaker showed an implicit understanding of the situation and took on the effortful burden of repair. Other-initiated repair or alerting to the need for repair, is one of the “should” functions presented in Figure 1, which brings the attention of the speaker to the need for repair, occurs during or immediately after the problematic conversational turn, putting the burden of repair on the speaker, and defers whatever information was due until after the problem was repaired or abandoned (Schegloff, 1997). This process of requesting repair is inherently an automatic process; however, as it functions to defer the course of dialogue, it initiates a controlled serialisation of information processing, as the speaker must identify the source of the communicative breakdown before attempting repair (Musslick et al., 2016a, 2016b; Schegloff, 1997).

As such, it appears that it is generally the case for native speakers to collaboratively build towards mutual understanding resulting in a marked reduction in the number of turns and words used to successfully communicate intended meanings, where speaker and listener assist each other in providing the necessary reduction of uncertainty and indications of successfully doing so respectively. This process continues and mutual understanding persists and a positive correlation between effort and uncertainty is only interrupted when the need for repair is initiated resulting in an increase in effort over the course of a controlled serialisation of information in the repair cycle.

This added effort in such a repair cycle may be why a potentially negative leaning correlation between effort and uncertainty is observed for non-native speakers relative to native speakers. If non-native speakers are focusing on communicative success due to their cognitive development (Ramscar & Gitcho, 2007) or as an inherent tendency or strategy a more novice language user would use (Ryan, 2015); it becomes possible that non-native speakers may take on more of the burden in reducing uncertainty through various forms of over explicitness in communication (e.g. repetition, use of full noun phrases) increasing effort per utterance but

avoiding overall increased effort in avoiding the repair cycle itself. The following subsection covers four studies that show evidence of an economy of effort in non-native communicative interaction and reveals some of the differences in this tendency when compared to native speakers while discussing why these differences may occur.

3.2 Studies of non-native speaker collaborative interaction

Based on the following review of studies of non-native speaker interaction (Cheng & Warren, 1999; Feng, 2022; Önen & İnal, 2019; Ryan, 2015) it is evident that non-native speakers tend to be overly explicit in their L2 referencing, which suggest focus on maintaining successful communication and avoiding communicative breakdowns. This means that non-native speakers opt to increase their effort per utterance for communicative success and clarity, which may appear to run counter to a tendency towards an economy of effort in communication. However, this increase in effort per utterance is in fact inline with an overall economy of effort in avoiding the cost of repeated repair cycles; It is likely the case that non-native speakers are aware of their status as a novice user of the target language, and the increased likelihood of communicative breakdowns due to this status from their meta linguistic knowledge of communication from L1 experience (Arnon & Ramscar, 2012; Birdsong, 2009); thus they leverage this added effort as a strategy to reduce total effort in communication.

Therefore, the context of non-native communicative interaction causes an observable difference in the manifestation of an economy of effort in communication and its influence on how non-native speakers economise their effort in terms of word count and possibly the number of turns they use to achieve or satisfice their communicative goals. Although it may indeed be the case that these same non-native speakers would show a similar tendency towards reducing overall turns and word count when communicating in their native language, such as the native speakers of English covered in the previous subsection, the current set of studies reveals that in communicating in a non-native language they do in fact increase their number of words used.

In their study Cheng and Warren (1999), found that non-native speakers were more explicit than native speakers in naturalistic conversation, exhibiting a higher word count per conversational turn in pursuit of clarity. Employing qualitative and quantitative methods of analysis, they analysed 29 naturalistic conversations (ten hours in 84,000 words) involving 76 participants (42 non-native speakers and 34 native speakers) for differences in inexplicitness (e.g. use of anaphora or zero anaphora, and context-based

referencing) between both groups. Their results showed that non-native speakers were generally more explicit in their utterances which lead to a greater word count compared to their native speaker counterparts due to repetitive speech patterns, limited linguistic competence, and native language transfer. The results found that non-native speaker utterances used 25-30% more content words compared to native speakers and were 2.5 times more likely to repeat part or all of their utterances, while native speakers were more likely to employ inexplicitness strategies (ellipsis, substitution, deixis, and reference) than non-native speakers.

In a similar study Önen and İnal (2019) complied and analysed a corpus (Corpus IST-Erasmus) comprising 29 interviews and 25 focus group meetings, resulting in 93,913 words of transcribed data focusing on identifying patterns of explicitness (e.g. lack of anaphora, zero anaphora, repetitiveness, and use of full noun phrases) within the collected corpus data. The results showed that non-native speakers were indeed showing patterns of over explicitness for the sake of the listener through repetition and over explicit forms (e.g. black colour rather than just black). These findings further suggest that over-explicitness is a characteristic feature of non-native speakers' utterances employed for the sake of the listener to ensure clarity and mutual understanding. Overall, both studies revealed a characteristic tendency for non-native speakers to be over explicit in their naturalistic use of the target language. However, non-native speakers are also tolerant of over-explicitness and a preference for over-informativeness and redundancy over ambiguity and uncertainty to similar degrees as native speakers on a receptive level (Feng, 2022).

In their study, Feng (2022) explores non-native speakers' tolerance for pragmatic violations in ad hoc implicatures (listener's interference of speaker's intended meaning based on context regarding ambiguous statements) and contrastive inference through an experimental paradigm focused on sentence judgement. 21 native speakers of English and 49 non-native speakers of English (L1 Mandarin Chinese) judged sentences for their naturalness on a 7-point Likert scale, where 1 indicated unnatural and 7 indicated natural for their levels of informativeness (over vs under informative statements). The results of the study showed that both native and non-native speakers on a receptive level judged and rated utterances, similarly, showing a clear preference for over explicit and informative utterances over under explicit and under informative ones. These findings suggest that for the population of participants in this study, regardless of their native language, had a similar tolerance pattern when it came to processing and evaluating the informativeness of statements within the context of the study's experimental paradigm.

Therefore, if both native and non-native speakers tolerate over-explicitness similarly on a receptive level, then why do they differ on a productive level as shown by the previous review of corpus-based studies? Addressing a similar question, one study (Ryan, 2015) examines if this tendency towards over-explicitness represents a transitory developmental stage or if it is indicative of a deliberate communicative strategy, driven by a desire for successful communication through clarity which can be considered a form of reducing uncertainty. Their experimental paradigm was focused on eliciting referential expressions in a narrative retelling of an edited version of the Charlie Chaplin film "Modern Times" using an accessibility theory-based framework to assess if over-explicitness exists in non-native speaker retellings and whether over-explicitness is a developmental stage or a communicative strategy. The study assigned 10 native speakers of English and 10 non-native speakers (L1 Mandarin Chinese) to the role of a speaker retelling the narrative of the edited film to a different native English speaker listener assigned to each participant.

The basic underpinning of the accessibility theory framework employed in this study explains the relationship between the accessibility (ease of recovery) of discourse entities and the choice of noun phrase types used to refer to them, proposing a hierarchy where the more accessible a referent is, the less explicit the referring expression typically is, ranging from pronouns for highly accessible entities to full noun phrases for less accessible ones. Based on this underpinning, it was found that in highly accessible contexts non-native speakers were significantly more explicit, to the point of over-explicitness than native speakers, using full noun phrases where pronouns and anaphora were expected.

Analysis of the retellings of non-native speakers, especially ones where over-explicitness was observed in highly accessible contexts instead of the expected anaphora or zero anaphora and instances of error correction, revealed that non-native speakers employed over-explicitness as a means to reduce uncertainty and avoid communicative breakdowns. This strategic behaviour, evidenced by adjustments in referential expression following instances of communicative breakdown, highlighted a high level of metalinguistic awareness among the non-native speaker participants (Ryan, 2015). The findings of this study therefore suggests that while non-native speakers observed in this study and the previous corpus-based studies (Cheng & Warren, 1999; Önen & İnal, 2019) are opting towards increasing their effort per utterance for the sake of reducing uncertainty for their listener in order to avoid communicative breakdowns.

Although not explicitly noted in Ryan's (2015) study, these results may also reflect an underlying tendency towards an economy of effort that influences L2 communicative

interaction, as avoiding a repair cycle through added effort falls in line with an economy of effort as it represents a reduction in total overall effort and as noted in the study reflects a deeper metalinguistic awareness of their status as a novice speaker making them prone to making errors without the strategic mobilisation of effort. However, and more importantly for the focus of this thesis, is that these previous four studies then reflect an underlying deference in the observed tendency towards an economy of effort between native speakers and non-native speakers when compared to the studies of native speakers. Where an economy of effort for native speakers leads towards a positive correlation between communicative effort and uncertainty, and a negative leaning correlation between communicative effort and uncertainty as a result of a focus on communicative success rather than accurate target language use at least at this novice level.

As previously noted adult learners shift their focus from language form to communicative success (Ramscar & Gitcho, 2007) due to their cognitive and developmental advantages (Arnon & Ramscar, 2012) and that this may be a practical necessity due to the rapid and transient nature of the input adults encounter in these naturalistic contexts creating a now or never bottleneck (Christiansen & Charter, 2016), and when adults make errors in communication these errors result in a request for repair from their listener. While this process of requesting repair is inherently an automatic process, it functions to defer the course of dialogue and initiates a controlled process that serialises information process as the speaker must identify the source of the communicative breakdown before attempting repair (Musslick et al., 2016a, 2016b; Schegloff, 1997). The dichotomy between automatic and serialised controlled processes also plays an important part in how and why adult learners focus on success as part of an economy of effort to avoid the added effort of a repair cycle with implications for the discussion of associative learning. The following subsection covers the differences between automatic and controlled serial processes and highlights the aforementioned implications for associative learning.

3.3 Implication of a focus on success in communication for associative learning

The dichotomy between automatic and controlled processes plays a significant role in how adult learners in naturalistic contexts aim to achieve their goal of communicative success and subsequently how they deal with considerably complex input and therefore the transition from input to intake. Automatic processes are heavily trained, resistant to

interference (Botvinick & Cohen, 2015; Cohen et al., 1990; Shiffrin & Schneider, 1977) and allow for parallel processing (Musslick et al., 2016a, 2016b); while controlled processes require larger commitments of cognitive control and effort in order to transition away from automatic processing as the default behaviour (Botvinick & Cohen, 2015; Cohen et al, 1990; Miller & Cohen, 2001; Shiffrin & Schneider, 1977) and inhibit parallel processing to avoid bottleneck effects in task completion (Musslick et al., 2016a, 2016b).

The experimental conditions within the Bavelas et al. (2000) study provide an example of this dichotomy. As previously noted, distracted listeners were tasked with counting all words starting with [T] produced by the narrators, and they were unable to engage in backchannel responses as a part of the process of calibration. T-word counting can be considered a controlled serial process, compared to normal listening conditions, which can then be considered an automatic process, as normal listening allowed participants to engage in parallel processes such as bimodal gesticulation. This is a significant observation, as a similar distinction can be made between focus on communicative success and a focus on language form when processing language input in a naturalistic setting, as the former allows participants to communicate successfully and calibrate mutual understanding; while the latter is likely to hinder communication as indicated by the performance of distracted listeners in this experiment. This again supports the possibility that adults may track communicative success in naturalistic contexts out of necessity.

The distinction between [T] word counting participants and normal listening participants and their difference in ability to engage in parallel processes such as bimodal gesticulation highlights the differences between normal listening and listening for repair; but also support the notion that adults cannot focus on language form without forgoing the ability to engage in parallel processing, as this results in an insurmountable bottleneck due to the time pressure of the modality of speech. As previously noted, it is a practical necessity for adult language learners to focus on communicative success, especially when their communicative goals far outstrip their linguistic abilities. Studies of adult associative second language learning (review forthcoming) provide clear and distinct evidence of the effects of salience on the automatic processing of language input influencing the order of acquisition; however, their results also indicate that psychological salience is superseded by task-based goals.

If adult learners are processing language input automatically, this allows for parallel processing at the cost of overall accuracy, therefore allowing them to perform as active

listeners that engage with the speaker while parsing through the input to reduce uncertainty and indicate when they understand the speaker. However, when a communicative breakdown occurs a shift from the automatic processing of language input to the controlled serialisation of the same input as the speaker must identify the source of the communicative breakdown before attempting repair. When information is serialised for this controlled process, the usual bottleneck of naturalistic interaction is mitigated by the context of repair, otherwise this controlled processing of input would inhibit parallel processing as was the case for the [T] word counting participants in Bavelas et al. (2000) causing degradation in automatic processing and subsequently task performance. This is due to the transition to controlled processing causing overlaps in the representations used in automatic input processing which necessitates increased cognitive effort (in allocating resources) and compromises parallel processing efficiency. (Musslick et al., 2016a).

Simulations of neural networks trained to use shared representations indicate that shared representations (cognitive or neural patterns reused across tasks to enhance learning efficiency but limit concurrent task performance) promote efficient use of neural network resources, such as needing fewer associative nodes, and allowing for generalisation to occur (Bengio et al., 2013; Saxe et al., 2013); however, even modest amounts of overlap in the representations used by different concurrent tasks causes a large scale bottleneck effect limiting the amount of parallel processing that is possible (Feng et al., 2014; Musslick et al., 2016a). Based on the results of Musslick et al. (2016a, 2016b) overlap in use of shared representations causes a transition from automatic parallel processing to serial controlled processing of incoming information. However, this transition is also dictated by previous feature-based learning of the characteristics of the problem, allowing adults to adapt their strategies not only to recurrent problems, but also to predict which type of process provides the highest expected value of mobilising cognitive control (Lieder & Griffiths, 2015, 2016; Griffiths et al., 2015).

Furthermore, based on the results of Kool et al. (2017) people only transition from automatic to controlled processing when stakes and incentives are high in conjunction with the need for a level of accuracy beyond what automatic processes can yield. Mental effort mediates the degree of cognitive control mobilised in response to task demands along the dimensions of task identity (i.e. what to attend to) and intensity of control (i.e. automatic vs controlled processing) (Shenhav et al., 2017). The model of expected value of control proposes that mental effort weighs the potential value of the reward (desired outcomes) against the amount of accuracy and the cost of control in selecting between automatic and

controlled information processing (Shenhav et al., 2013). The influence of increased incentives on the motivation to mobilise more cognitive control suggests that it is the limited mobilisation of control due to its cost rather than limited cognitive ability per say (Botvinick & Braver, 2015; Shenhav et al., 2017).

Therefore, when participants learn about when and where to mobilise cognitive control for the purpose of engaging in controlled serialisation of language input information for the sake of communicative repair they also learn about the characteristics of that repair. If as a speaker or listener, the adult learner who is in a limited end-state of target language proficiency only experiences communicative breakdowns due to lexical gaps in knowledge (Ellis, 2008b; Klein, 1998; Klein & Perdue, 1997) in spite of the ungrammaticality of this limited end state (Ellis, 2008a); then any effort exerted by the learner in repairing breakdowns usually only results in repairing lexical gaps not morphosyntactic gaps in knowledge, influencing the developmental trajectory of learners' basic variety and causing them to persistently stabilise at that level of proficiency, due to the relative asymptotic state of reduced uncertainty which is achieved through mostly lexical means. Perhaps even this process of only repairing lexical gaps adds to the issues of salience, selective and learned attention, and relative redundancy that affect the acquisition of morphosyntactic cues through associative learning, by making lexical cue dimensions more salient as they become strongly associated with the outcome of repaired communicative breakdowns.

As such the following section focuses on presenting adult learners' focus on success in naturalistic contexts, and the influence of an economy of effort in reducing uncertainty in communication as additional variables that inhibits the morphological development of the limited end-state of adult naturalistic usage-based SLA through associative learning i.e. the basic variety. The following section provides evidence that adults have the means to recover from the blocking and overshadowing of less salient cues similarly to children. However, a focus on communicative success and the resistant nature of automatic input processing that this focus depends on inhibits the development of the associative strengths of morphosyntactic cues. Furthermore, this focus on success facilitates the development of associative strengths of the overall satisficing solutions, shifting the focus of associative learning trials to the less granular level of overall communicative success rather than language form.

4. Review of studies of adult and child associative learning

Based on the review of literature put forth in the previous sections it is clear that adults come to the task of second language acquisition in a significantly different context when compared to children learning their first language. Adults not only come to this task with significant developmental advantages but also come to a significantly different task, as the naturalistic input they encounter is more complex, rapid, demanding, and in a language which they are still novice users of. The aim of this section is to provide evidence that the limited end-state of naturalistic usage-based adult SLA can be explained by the caveat not being satisfied that for associative learning to occur there must be an impetus to trigger changes in the cues associative relationship to the outcome; and that the failure to satisfy this caveat is due a shift in the granularity at which associative learning trials occur. Furthermore, it aims to argue that this shift is the result of adult learners' focus on communicative success rather than language form, and a reliance on automatic parallel processing of naturalistic language input out of necessity allowing them to perform as active listeners that engage with the speaker while parsing through the input to reduce uncertainty and indicate when they understand the speaker.

This section will first cover a detailed review of associative learning theory, the attentional phenomenon it subsumes and the influential Rescorla-Wagner (1972) model and equation that explains how learning from prediction error occurs. This coverage will be followed by an explanation of how a focus on communicative success and a dependence on automatic input processing can result in a shift from language form being the target of associative learning trials to the overall success of an utterance in a communicative context. This is then followed by two sections that review studies of adults' associative learning and children's associative learning. These studies reveal that adults show evidence of the availability of morphosyntactic cues in memory, but these cues are not having their associations strengthened through implicit prediction error due to the influences of a focus on communicative success and a tendency towards an economy of effort. Conversely, children recover from erroneous inferences of overgeneralization when representations of the targeted linguistic cues were available in their memory and that these cues were having the strengths of their associations adjusted through implicit prediction error.

4.1 Overview of associative learning

Associative learning is learning about the predictive relationship between a particular cue and an outcome, and surprisal (i.e. surprisal from discrepancy between expected and actual outcome or prediction error) maximally drives learning (Cintrón & Ellis, 2016; Rescorla, 1988; Rescorla & Wagner, 1972). According to Cintrón and Ellis, (2016) one of the most influential formulas in associative learning theory and learning theory in general is the basic equation of the Rescorla-Wagner (R-W) model (1972). The R-W model (1972) simulates learning as changes in associative strength between a cue and outcome as a result of discrete learning trials (Arnon & Ramscar, 2012). The basic equation of this model is [$\Delta V = \alpha\beta(\lambda - \sum V)$] where ΔV is the new value of the associative or predictive strength of the conditioned stimulus (the predictive cue of the outcome which is λ) after a learning trial has occurred. α is the salience (obviousness, i.e. the degree to which it stands out due to physical or psychological attributes) of the conditioned stimulus, and β is the rate of learning. λ is the outcome (if it is correctly predicted i.e. if A predicts B and B does in fact occur then the value is 1 if not the value is 0). Finally, $\sum V$ is the sum of predictive value of the conditioned stimulus (if it is the first learning trial its value is 0 since no learning has occurred previously) from all learning trials. I.e. $\sum V$ = addition of the results of all previous trials. As Cintrón and Ellis (2016) put it this formula is one of the most influential formulas in learning theory, and salience and surprisal interactively affect the outcome of learning from each trial.

In this formula salience [α] and rate of learning [β] are determiners of what gets learned and how much is learned about it respectively. Salience is a factor of both the inputs physical attributes (e.g. more phonologically pronounced, and contrastive within context) and psychological attributes (e.g. psychologically salient due to previous experience with the target language and/or native language transfer) (Cintrón & Ellis, 2016; Ellis, 2006). Rate of learning is determined by the discrepancy between expected outcome and the actual outcome (Arnon & Ramscar, 2012; Cintrón & Ellis, 2016; Ramscar et al., 2010), that is the greater the degree of discrepancy the more is learned and vice versa the smaller the discrepancy the less is learned until learning reaches an asymptotic state (Arnon & Ramscar, 2012) and the greatest amount of learning is due to surprisal that results from prediction error (Cintrón & Ellis, 2016). As such in each learning trial the values of α and β influences the gains or losses in predictive value which depends on the occurrence of the predicted outcome. However, uncertainty of outcome is finite, if a cue can predict an outcome to the point of asymptotic learning, then there is a finite amount of predictive value

that an outcome can support (Arnon & Ramscar, 2012), as a result cues compete for association with outcomes through the R-W formula, resulting in both positive evidence about co-occurrence between cues and outcomes, and negative evidence about cues and outcomes that did not occur (Ramscar et al., 2010).

Furthermore, the R-W (1972) equation entails a constant and mostly implicit, updating of associative strengths when cue-outcome pairings are encountered (i.e. the learner encountering mean to language form associations as a listener), and it is one where salience and surprisal interactively affect the outcome of learning from these encounters (Cintrón & Ellis, 2016). Overshadowing and blocking are functions of salience and surprisal influencing the results of cue competition where cues are competing to be associated predictively with particular outcomes. When unlearned cues compete to be predictive of an outcome the most salient cue overshadows the others and is associated predictively with the outcome (Ellis, 2008b; Miller et al., 1995); While previously learned cue-outcome associations block (prevent) that particular outcome from being associated with other cues (Ellis, 2008b; Kamin, 1969; Kruschke, 2006; Miller et al., 1995). Furthermore, blocking is a statistical consequence of cue-outcome associations being learned to asymptote (as close as possible to 100% prediction), and outcomes only supporting a finite amount of associative value (Arnon & Ramscar, 2012; Kamin, 1968; Kruschke & Blair, 2000; Mackintosh, 1975).

The R-W (1972) equation presents two caveats for associative learning to occur. First, for a cue to be associatively strengthened or weakened in relation to an outcome it must be available in the memory of the learner. Second, there must be an impetus to trigger changes in the cue's associative relationship to the outcome, in other words a clearly perceived learning trial. As for the first caveat, there are two important components, the learner's memory which is the ability to encode, store and retrieve the encoded information, and the linguistic cues themselves which become memories through encoding and are stored for later retrieval (Divjak, 2019). According to Divjak (2019) encoding is the process of learning where the linguistic cue first enters memory, storage is the phase where the memory itself remains when unused and retrieval refers to the activation of and accessing of these memories of linguistic cues when needed. However, these linguistic cues are not encoded perfectly as memories, and subsequently retrieved verbatim, they are reconstructed during retrieval (Bartlett, 1995; Divjak, 2019). Therefore, if the linguistic cues that make up the parts of the utterance are weak or ambiguous they may not be retrieved as part of the utterance (Tulving & Schacter, 1990).

However, the more often these linguistic forms are retrieved, the more entrenched and strengthened they become. This is demonstrated by studies of child associative error-drive language learning (Dye & Ramscar, 2009; Ramscar & Yarlett, 2007) which found that children recovered from their overgeneralisation of regular plural forms to irregular plural forms through prediction error from encountering irregular plural forms receptively and without production and explicit correction of the irregular plurals. In this case prediction error resulting from language form discrepancies during receptive trials resulted in a U-shaped pattern of learning where under and over predicted cues reached a state of learned equilibrium between irregular and regular plural forms respectively. Therefore, in terms of the R-W (1972) model, children recover from the erroneous overgeneralisations of regular plural forms when the irregular plural forms are present in their memory (observing the first caveat of the model), over the course of multiple instances of reconstruction through retrieval resulting in prediction error as these irregular plural forms were the focus of their task (observing the second caveat of the model).

Conversely, adult learners in naturalistic context do not appear to go through a similar process of implicit updating of their linguistic cues' association with outcomes. However, adults typically stabilize within the limited end-state of a basic variety of the target language (Ellis, 2008a, 2008b; Long, 1990). This limited end-state is characterised by the acquisition of that mostly lexical open-class linguistic cues (nouns, verbs, adjectives), and not closed-class morphosyntactic cues (Ellis, 2008b; Klein, 1998; Klein & Perdue, 1992, 1997). Although the definition of blocking would imply that it is simply the case that morphosyntactic cues are not available in the mind of the learner, and therefore do not appear in their language use; the results of studies of adult associative learning (e.g. Cintrón & Ellis, 2016; Ellis & Sagarra, 2010a, 2010b, 2011) indicate that adults had acquired representations of cues that were meant to be blocked by the design of the study, observing the first of the aforementioned caveats for associative learning. This means that the issue adults are facing is potentially the lack of prediction error occurring when encountering language input, which means that the second caveat that a clear learning trial must be perceived is not being observed may be the cause for adult learners stabilising at a limited end-state of the target language. This could be attributed to the possibility that adult learners are not focusing on language form during naturalistic social interaction. It may in fact be the case that adults are focusing on communicative aspirations (the minimum level of accuracy predicted for success) rather than language form as a means of successful communication and social survival (Arnon & Ramscar, 2012; Birdsong, 2009; Ramscar & Gitcho, 2007). In naturalistic

contexts, the modality of speech presents a formidable challenge for beginner adult L2 learners as it adds a time pressure on learners to process information due to its transient nature (Christiansen, 2019; Cintrón & Ellis, 2016) which causes a now or never bottleneck (Christiansen & Charter, 2016). Furthermore, beginner adult L2 learners face this time pressure when their communicative needs far outstrip their linguistic ability (Slobin, 1993). As suggested by Ryan (2015) in the previous section, evidence of a focus on communicative aspirations by adults comes from their strategic increase in communicative effort to reduce uncertainty for the sake of their listeners. This focus is further evidenced by the unprompted tendency to segment letter strings into word-like units (Isbilen et al., 2017) and to look for vocabulary meaning when encountering a novel language (Andringa & Curcic, 2015). Furthermore, adults misinterpret 93% of implicit morphosyntactic feedback as semantic (Mackey et al., 2000) and require explicit metalinguistic feedback to overcome these misinterpretations (R. Ellis et al., 2006).

This evidence of a focus on communicative success makes it clear that adults are not only coming to the task of language learning with significant differences in cognitive, neurobiological, and linguistic development (Birdsong, 2009; Arnon & Ramscar, 2012). But also, that their focus is not on language form and that learning about language form is incidental and based on the characteristics of the basic variety as a limited end-state is mostly centered around lexical development (Klein, 1998). It further indicates that adults are focusing on a less granular, and more overall utterance level of associative learning, where the general communicative strategy, and its more salient (both physically and psychologically) lexical items are the subject of associative learning trials rather than the more granular language form cues. Essentially, adults are using their attention as a filter to look for meaning (Wu, 2014), prioritising lexical items which are more relevant to the goal of the task (Nobre & Kastner, 2014) for better performance (Mishra, 2015) which in the case of communicative interaction is success through the reduction of uncertainty when listening to or producing an utterance.

Adult learners are therefore likely to not even be making predictions subject to error driven learning regarding language form, perhaps due to being forced to rely on automatic parallel processing of language input to deal with the complexity of language input itself. Meaning that although these language forms may be present in the mind of the learner from an emergentist perspective, they are not part of associative learning trials. It is therefore possible that in the early stages of naturalistic adult usage-based SLA, that psychological and physical salience of utterance and lexical level success draws the learners' attention to

cue dimensions as the starting point for associative error driven learning subsequently overshadowing and blocking morphosyntactic cues not from acquisition but from the development of their associative strengths. This would therefore deprive the morphosyntactic cue dimension of receiving attention from the learner which is a practical necessity for learning in general to occur (Long, 1991; Schmidt, 2001), with this reduction in attention given to this cue dimension likely having the effect of inhibiting the satisfying of one or both caveats of the R-W (1972) model for associative learning to occur.

Based on the previous review of associative learning in which learning is through the strengthening and weakening of associations between cues and outcomes based on predictions, the question is what is the source of prediction error and confirmation that strengthens or weakens the associations between cues and outcomes. The logical problem of language acquisition from a Chomskyan perspective presumes that learners cannot recover erroneous inferences without corrective feedback and that language learning is based on innate abilities (Dye & Ramscar, 2009). Furthermore, there is evidence to suggest that children rarely receive corrective feedback (Brown & Hanlon, 1970) and ignore it when they receive it as evidence by them not repeating their utterances with the given corrections (Marcus, 1993). Similarly for adults, implicit corrective feedback that is intended to repair morphosyntactic errors is rarely recognised as such (Mackey et al., 2000). However, studies of children's associative learning (e.g. Dye & Ramscar, 2009; Ramscar et al., 2010; Ramscar & Yarlett, 2009) has shown that children can and do recover from erroneous inferences through exposure to input and the process of error driven learning and latent learning with the availability of targeted linguistic forms present in memory to be targeted for association. While results from studies of adult associative learning (e.g. Cintrón & Ellis, 2016; Ellis & Sagarra, 2010a, 2010b, 2011) indicate that adults had acquired representations of cues that were meant to be blocked study design indicating that the first caveat for associative learning to occur has been satisfied and that the issue is likely to be a failing in satisfying the second caveat namely the impetus to trigger changes in the associative strength of those language form cues that are available in the memory of the learner. Therefore, the following subsections will first cover the studies of adult associative learning, followed by an exploration of children's associative learning and its implications for adult learners.

4.2 Studies of adults associative learning

This section covers a series of experiments on adult associative learning (Ellis & Sagarra, 2010b, 2011) that all largely follow a similar experimental paradigm with some modifications and variations between them. The focus of these studies was on measuring the effects of short and long-term learned selective attention. Each of these studies conducted two experiments on either native speakers of English, to measure short-term learned selective attention effects; or compared the performance of non-native speakers from different L1 backgrounds to determine the effects of long-term selective attention on the acquisition of a set of temporal reference cues taken from a miniature set of Latin. The experimental paradigm consisted of 4 phases when measuring short-term learned selective attention effects and excluded the pre-training phase when measuring long-term selective attention effects resulting in 3 total phases.

In the pre-training phase experimental group participants are exposed to orthographic representations of either past or present adverbial or verbal inflection cues of temporal reference taken from a miniature set of Latin (Hodi = today, Heri = yesterday / Cogito = I think, Cogitavi I thought). During this phase participants were asked to select the cue that referred to the required tense (i.e. Hodi and Heri are on screen, and the prompt asks the participants to select which of the two is in the present), participants were given feedback in the form of correct or incorrect - [Latin] means [English]. All participants take part in phase two and are required to decode the tense of the sentence, in this phase two new cues of future temporal reference are added (the adverbial Cras, and the verbal inflection Cogitabo). In this phase participants encounter a logical combination (i.e. same tense) of adverbial and verbal inflection cues in counterbalanced order. They are asked to indicate whether the sentence is in the past present or future tense, they are again given feedback identical to phase 1.

Phase 3 is a receptive judgement test, participants now encounter both logical and illogical pairings of adverbial and verbal inflection cues and are asked to rate the tense of these sentences on a scale of 1 (extreme past), 3 (present), 5 (extreme future), participants did not receive feedback during this phase. Phase 4 is the final phase of the experiment where participants are asked to translate logical pairings of adverbial and verbal inflection cues into English, again participants did not receive feedback during this phase. In Ellish and Sagarra (2010b) the first experiment compares the learning of native speakers of English either pre-trained (on adverbial or verbal inflections cues), or under control conditions to measure the effects of short-term learned selective attention from pre-training. While experiment two focused

on native speakers of Chinese without pre-training to measure the effects of long-term learned selective attention.

The results of experiment one showed that pre-trained participants relied almost exclusively on their pre-trained cues to rate tense in phase 3, while the control group was more divided in their reliance on these two cue dimensions. Native speakers of Chinese were found to rely exclusively on adverbial cues similar to their pre-trained counterparts, the authors note that this is due to Chinese being an inflection-free language resulting in the higher psychological salience of adverbial cues. Although the results of this experiment clearly demonstrate the effects of psychological salience from both long and short-term language experience, and that relying on psychological salience is largely a default and automatic behaviour as participants were not instructed to favour one cue dimension over the other but were left to their own devices. This result was also pronounced in participants preference for cues of future temporal reference based on cue dimensions from pre-training, as they were not trained on these cues during phase 1 indicating that their reliance on said cues was not biassed by previous experience with specific cues, i.e. the psychological salience of future reference cues was raised by their similar cue dimensions to previously learned past and present cues.

In terms of the results of phase 4 it was found that for the native speakers of English, the adverb pre-training group provided the correct adverb on every trial, even when not explicitly requested, and tended to provide an idiosyncratic verb form of “cogitavo”. While the verb inflection pre-training group provided the correct inflection where required and when a bare adverb was required, one was provided; however, it was usually an incorrect adverb. Native speakers of Chinese performed similarly to the adverb pre-training group, being better able to produce adverbial cues than inflectional cues, but less able to produce inflectional cues. The authors interpreted these results of these experiments as a confirmation of the clear effects of both short and long-term attentional bias and subsequent blocking of cue acquisition. Essentially their interpretation is that cues that are acquired earlier block the acquisition of later experienced cues, and that by the same token this is what leads to the limited end-state that second language learners find themselves in. Although the appearance of these supposedly blocked linguistic cues in participant productions would indicate that they are present in the mind of the learners, such a finding would mean that the first caveat of the R-W (1972) equation is satisfied.

However, an issue in this study was that both adverbial and verbal inflection cues of temporal reference were artificially balanced (3 of each cue), an influencing factor that was acknowledged by the authors. As this balance was not representative of natural languages;

however, it also caused an economical balance between both cue types. In their 2011 study Ellis and Sagarra increased the complexity of the set of verbal inflection cues by adding second- and third-person singular cues, while adverbial cues remained unchanged. Resulting in 3 adverbial cues and 9 verbal inflection cues total. This increase in the number of verbal inflection cues should reduce their comparative salience, frequency and contingency (Ellis, 2006) as a smaller subset of adverbial cues will ensure that particular cues will occur with more relative frequency with a particular tense. Furthermore, this increase in verbal inflection cues also gives adverbial cues a considerable economical advantage, as participants are required to learn and maintain a smaller vocabulary to correctly judge the tense of training stimuli during phase 2 (Zipf, 1949), subsequently influencing their reliance on cue type during phase 3 in rating tense.

Regression analyses predicting mean temporal interpretation in Phase 3

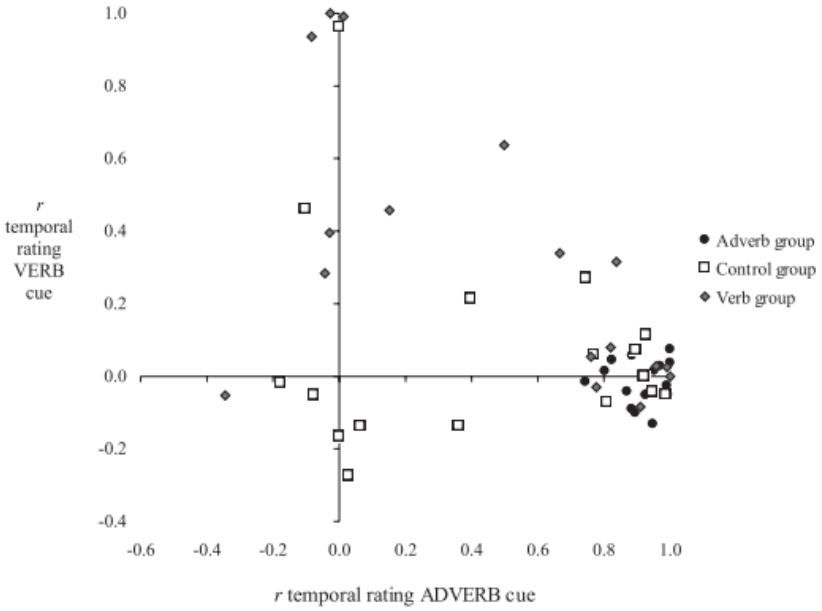
Group	β	95% CI	Adjusted R^2	F
Adverb pretraining				
Adverb	0.99*	[0.96, 1.02]	0.99	2053*
Verb	–0.01	[–0.04, 0.02]		
Verb pretraining				
Adverb	0.76*	[0.69, 0.83]	0.94	413*
Verb	0.60*	[0.54, 0.67]		
Control				
Adverb	0.93*	[0.84, 1.02]	0.88	192*
Verb	0.17*	[0.07, 0.26]		

* $p < .001$.

Table 1: Regression analyses predicting mean temporal interpretation (Ellis & Sagarra, 2011; p599)

As shown by the results in Table 1, this was indeed the case in experiment one focused on pre-trained and control condition participants who were native speakers of English. Adverbial and control group participants were wholly reliant on adverbial cues, while verbal inflection group participants showed a considerable sensitivity to their pre-trained cues, they overall relied more on adverbial cues. This is clear evidence of the influence of both task-based aspirations (the assumed minimum level of accuracy needed for success) and economy of effort on the search for and selection of satisficing solutions. As the trial-and-error learning during phase 2 would have had to have led participants to focusing adverbial cues as a more economical and reliable means of judging tense due to their relative frequency advantages. Due to this cue dimension representing a smaller relative cue set that needed to be learned and maintained for success compared to the larger more varied verbal inflection set of cues. Subsequently

influencing their rating of tense in phase 3. This is highlighted by the results of the Pearson's r correlation which showed that within groups, individuals were highly influenced by adverbial cues in all groups and seven members of the verbal inflection pre-training group performed identically to adverbial pre-training groups as demonstrated by Figure 4.



Sensitivity to adverbial and verbal inflectional cues to temporal reference in each participant.

Figure 4 Sensitivity to adverbial and verbal inflectional cues (Ellis & Sagarra, 2011; p601)

The results of the production test in this experiment parallel the results of Ellis and Sagarra (2010b). Following an identical procedure (translating from English to Latin), the adverb group were able to provide the correct adverb, even on trials where it was not requested, and an idiosyncratic "congitavi" verb throughout their productions. The verb group was generally able to provide a correct verb+inflection and were able to provide an adverb though usually not a correct one. However, while the verb group was superior in their general ability to produce verbs they performed better on adverb production according to phase 4 results, with 7 participants behaving like adverb group members. Again, this was an indication that cues meant to be blocked by the design of the experiment were present in the mind of participants as evidenced by their appearance in their productions, or at the very least that they were becoming aware of that cue dimension.

Experiment 2 extends the findings of experiment 1 to the influence of long-term learned selective attention by comparing the performance of participants from different L1 backgrounds.

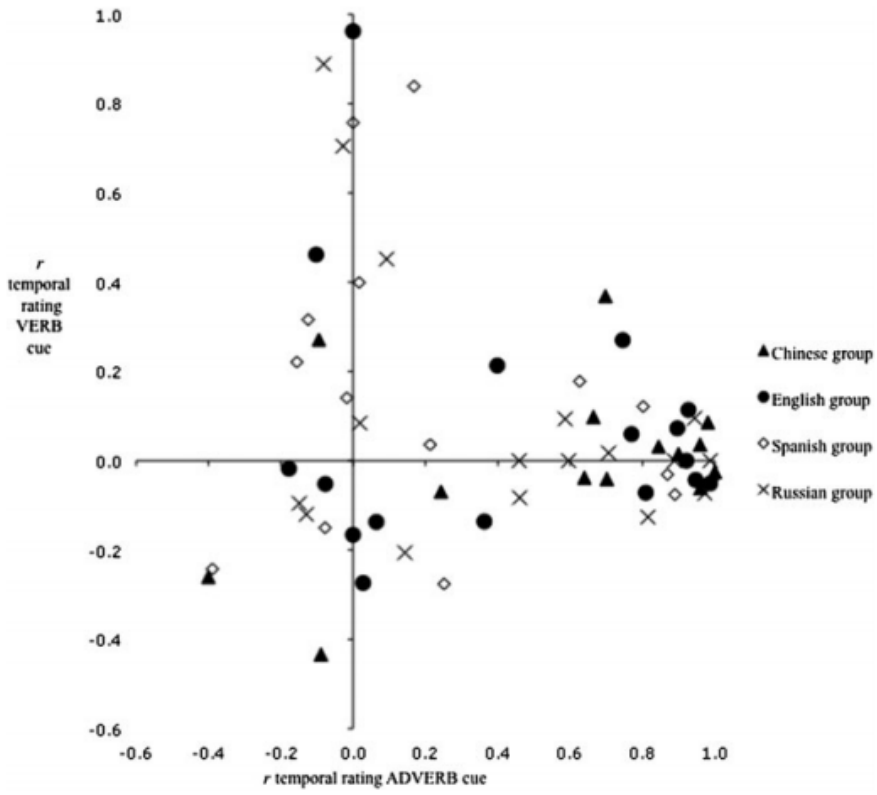
61 participants (inflection-free L1 Chinese [n=12], inflection-light L1 English [n=17], Inflection-rich L1 Spanish and L1 Russian [n=15, and n=17 respectively]) took part in this experiment which excluded the pre-training phase, but with the remaining phases being identical. Based on their L1 backgrounds native speakers of Chinese and English performed as expected by the authors and relied exclusively on adverbial cues in rating tense; while inflection-rich Russian participants also showed a near exclusive reliance on the same adverbial cues, and Spanish participants showed a greater sensitivity to verbal inflection due to the similarity of Latin and Spanish, but nevertheless relied heavily on adverbial cues in rating tense as shown in Table 2.

Regression analyses predicting mean temporal interpretation in Phase 3 of Experiment 2

L1 Group	β	95% CI	Adjusted R^2	F
Chinese				
Adverb	0.95*	[0.86, 1.04]	0.89	221*
Verb	–0.02	[–0.11, 0.07]		
English				
Adverb	0.93*	[0.84, 1.02]	0.88	192*
Verb	0.17*	[0.07, 0.26]		
Russian				
Adverb	0.91*	[0.82, 1.01]	0.88	196*
Verb	0.22*	[0.13, 0.32]		
Spanish				
Adverb	0.75*	[0.60, 0.89]	0.71	66*
Verb	0.41*	[0.26, 0.55]		

* $p < .001$.

Table 2 Regression analyses predicting mean temporal interpretation, experiment 2(Ellis & Sagarra, 2011; p612)



Sensitivity to adverbial and verbal inflectional cues to temporal reference in each participant.

Figure 5 Sensitivity to adverbial and verbal inflection cues (Ellis & Sagarra, 2011; p615)

A Pearson's r correlation again showed that within groups, individuals were highly influenced by adverbial cues in all groups as shown by Figure 5. This is again clear evidence of the influence of both task-based aspirations and the economy of effort on learning to satisfice said aspirations. These findings indicate that the influence of psychological salience depends on its alignment with the solution that represents the path of least probable effort, as determined by the aspiration levels of a task objective. This is evidenced by inflection-rich L1 background participants overcoming their automatic tendency to focus on verbal inflection due to L1 transfer and focusing on adverbial cues instead. Hypothetically speaking, their inflection-free/light counterparts should do the same and overcome their automatic tendency to focus on adverbial cues had verbal inflection cues provided an economical advantage in judging tense in phase 2 over adverbial cues.

4.2.1 Evidence for an economy of effort in associative learning

In Ellis and Sagarra (2011) experiment 2 focused on the effects of different L1 backgrounds on long-term learned attention, and whether the learning of verbal inflections was graded or discrete (i.e. does acquisition gradually change from one L1 background to the other depending on their use in the L1 itself or was there a minimum threshold for learning). The logic was that Chinese participants that come from an inflection-free L1 background would focus on the adverbs alone and each subsequent L1 background (English = inflection-light, Russian and Spanish = inflection-rich) would focus increasingly on verbal inflections. While this was true in the sense that Russian and Spanish participants were better able to make use of verbal inflections; however, the results indicate that as a group both L1 speakers of Russian and Spanish relied more on adverbial cues of temporal reference when rating sentences in the reception test of phase 3. This not only highlights the potential effects of individual cognitive differences, but also points to the possibility of an economization of effort taking place in selecting more salient and easier to attend to cues.

Furthermore, the expected result was for native speakers of inflection-rich languages to make more use of verbal inflection cues. Especially speakers of Spanish due to its close relation to Latin; however, it is possible that during the sentence decoding in phase 2 corrective feedback played a role in diverting participants attention to the more salient adverbial cues that were more reliable as participants found them to be success in sentence decoding more often than not. Essentially, these participants have come into this experimental paradigm with a hypothesis of where to look for temporal information, those participants whose L1 preferred adverbial cues started with them as the expert module and vice versa. The corrective feedback in phase 2 acted as the selective pressure, a measure of the cue's fitness, that demonstrated to them whether this was the correct cue dimension to allocate their attention to or not, and through this process of trial, error, and correction participants were able to find the cue dimension that could satisfice their needs and achieve their goals. Although the findings of these studies do indeed point to the possibility that the limited end-state of adult SLA is founded in the principles of associative learning (Ellis & Sagarra, 2010b), the role of an economy of effort in causing attentional biases must be taken into account.

This is because evidence presented in section 3 (e.g., Ryan 2015) shows that adults are attending to the goal of communicative success, which is a goal with aspiration levels set in the reduction of uncertainty. If an economy of effort influences communication, then under the conditions of bounded rationality (limited cognitive resources, time, and information), the process of satisficing becomes the process that determines the path of least effort for

communicative success by rating potential solutions on their predicted ability to satisfice communicative aspirations or the minimum expected level of reducing uncertainty. This may be why adults are not developing their morphosyntactic cues because they are not part of the associative learning trial if they are blocked due to a focus on success, especially since adult participants in this previous series of experiments have shown evidence of cues in mind from the results of the production test.

When adults perform as active and engaged listeners, these cues that are in memory are failing to develop implicitly even though being an active listener should entail constant instances of prediction error on a language form level for the adult learner. This automatic processing is needed for the parallel reduction of uncertainty and engagement as an active listener that indicates understanding to their interlocutor. But their focus on success and a reliance on automatic parallel input processing inhibits the use of controlled serial processing. Which may be needed to process language input on a granular enough level to experience prediction error on language form. This would mean that adults may be failing to notice input on a language form level due to the necessities of engaging as a listener to indicate understanding to their interlocutor. According to Schmidt (1990), failing to notice this input leads to it failing to become intake, although based on the availability of these cues in memory as previously mentioned, it may be the case that failing to focus on these language form cues leads to a lack of prediction error occurring for these cues leaving them underdeveloped associatively.

However, children, on the other hand, encounter repetitive language input and recover from erroneous inferences through the availability of linguistic forms in memory (Dye & Ramscar, 2009; Ramscar & Yarlett, 2007) when there is evidence that linguistic cues are present in their memories through mostly implicit means and learn to discriminate which outcomes are predicted by which cue without corrective feedback. The following subsection covers these studies to further understand how children recover from these erroneous inferences and what implications they present for naturalistic usage-based adult SLA.

4.3 Studies of children's associative learning

The aim of this subsection is to examine the mechanisms of associative learning in children's language acquisition, focusing on how prediction error and the discrimination of semantic cues and phonological outcomes facilitate learning without explicit feedback. The studies in the forthcoming review demonstrate how children can and do recover from erroneous inferences and overgeneralization of cue-outcome associations and what

implications this holds for understanding the nature of the limited end-state of naturalistic adult usage-based SLA.

In language learning, associative learning occurs as the strengthening or weakening of associations between semantic cues and phonological outcomes and these associations are stored as exemplars of paired semantic cues and phonological outcomes in memory. For example, the plural noun CARS is described by Dye and Ramscar (2009) and Ramscar and Yarlett (2007) as an exemplar in memory of a couplet encoding the semantic cues of car and its plurality and the phonological form /carz/ as a cue and outcome association. When learning which semantic cues or features predicts a phonological outcome, learning becomes the competitive process of discrimination between which features most accurately predict the outcome (Ramscar et al., 2010). In their experiments Ramscar et al. (2010) presented children with images of fictional animals named wugs and nizes, and children had to learn which features of wugs and nizes correctly discriminated them from each other, or in other words what was the difference between the two. When both animals shared the same body shape but different colours children quickly learned to ignore body shape and focus on colour to discriminate between wugs and nizes, and when they could have more than one colour and both animals shared those colours but different bodies children were again able to figure out that the feature of body was the best at discriminating between wugs and nizes, and that body shape helped them correctly name the animal. This process occurred without feedback from the experimenters and depended solely on the children's internal prediction error driven learning. As the features of body and colour were competing for association with the labels wugs and nizes the most competitive and successful cues were those that best discriminated wugs from nizes relative to the other available features, and as such loss in associative strength of one feature is another's gain allowing for association to shift from one feature to another (Ramscar et al., 2010).

As a result, features act as predictive cues for phonological forms that compete for relevance, and this competition is shaped by both positive and negative evidence about which outcomes occurred and did not occur respectively, with a common misconception of the R-W model being that only positive evidence of co-occurrence affects learning (Ramscar et al., 2010). In terms of the learning of plurals both regular and irregular in English, the frequency of regular plurals causes the semantic cues of plurality to often predict the phonological outcome of stem+S leading to overgeneralization in instances of irregular plural nouns such as mouses instead of mice (Dye & Ramscar, 2009; Ramscar et al., 2010; Ramscar & Yarlett, 2009). If learning associations between semantic cues and phonological

outcomes causes a semantic cue to predict a phonological form, and if learning occurs from prediction error. Then over or under generalisation of regular and irregular plural forms is a failure of discrimination that results from shared semantic cues that will self-resolve over the course of error driven learning improving discrimination (Dye & Ramscar, 2009). As learning reaches an asymptote through the strengthening of associations of underpredicted outcomes and the weakening associations of overpredicted outcomes (Ramscar et al., 2010), a state of learned equilibrium can be reached (Ramscar & Yarlett, 2007), and without any external feedback (Dye & Ramscar, 2009; Ramscar & Yarlett, 2007). However, the caveat needed for this process to occur and a state of asymptote and balance to be reached is that the learner must have successfully learned representations of the correct form(s) (i.e. irregular forms have been previously experienced) from previous experience (Ramscar & Yarlett, 2007).

In a series of studies Ramscar and Yarleet (2007) tested these notions. Their first experiment was designed to determine if children that overgeneralize regular plural forms did indeed have representations of the correct irregular plural forms. The results of this experiment were that children who did over generalise regular forms nevertheless had representations of correct irregular plural forms in memory. Furthermore, this experiment found that production of over-regularised forms was a poor predictor of preference in recognition of over-regularized forms or correct irregular forms; while it was found that production of correct forms and zero marked forms (word stems) was a better predictor of preference in recognition of correct plural forms (both regular and irregular). The authors interpreted these results as 1) a dissociation between production knowledge and recognition and 2) as an indication that the increase in zero marked forms was a sign that children were at a stage where they were beginning to master the linguistic aspect of pluralization rather than a sign of poor linguistic knowledge. In addition, there was a correlation between children's age and the production of zero marked forms, leading the authors to compare children by age. The finding of this comparison was that younger children produced less zero marked forms and more comprehension errors while the opposite was the case for older children. These findings are an initial indication that children are expected to get worse at using irregular plural forms before they improve (U shaped learning) as the frequency advantages of regular plurals benefit their error driven learning as they are afforded more learning trials; however, as learning of regular plurals reaches asymptote their frequency advantages diminishes (as those outcomes can only support a finite amount of predictive

value) and improvements in the production of irregular forms can being (Ramscar & Yarlett, 2007).

This assumption was tested in the subsequent experiments of Ramscar and Yarlett (2007). On the basis of having established that over-regularizing children did have correct representations of irregular plural forms, experiment 2 tested the ability of children to self-correct overregularization errors through simple repetition of errors without corrective feedback. In this experiment, children were asked to assist a doll to learn plural nouns of depicted animals on a laptop screen. Children were first presented with singular depictions and named those singular nouns and then they were presented with plural depictions and asked to name those plural depictions. Children performed this task a total of 4 times over a period of 9 days. As speculated, children were found to have significantly improved in their productions of the correct irregular plural forms, while decreasing in their production of over-regularized forms by the end of the experiment. Experiment 3 was conducted using the same methods over the period of 1 day to avoid the possibility of parental interference and incidental learning of irregular plural forms affecting the outcome of the experiment when children went home between trials. The same results were found, as performance by the end of the experiment significantly improved in terms of increased production of correct irregular plural forms and a decrease in the production of over-regularized forms; however, it was also found that children with better initial knowledge of irregular forms were more likely to improve, while children with worse initial knowledge of irregular forms were more likely to perform worse and indication of U shaped learning. Experiment 4 tested 10 more children and pooled their results with those of experiments 2 and 3. Using the criteria of initial knowledge of irregular plural forms the results of the children were split into two groups of better and worse initial knowledge of plural forms. The results of this analysis of pooled data confirmed the assumption that children with better representations of irregular plural forms improved while the performance of children with poor representations of plural form declined. In a subsequent study Dye and Ramscar (2009) tested the hypothesis that the results of Ramscar and Yarlett (2007) could be reproduced using a similar experimental design, while only presenting children with training on regular plural forms. It was found that older children with better initial knowledge of the representations of correct irregular forms improved (i.e. produce more correct irregular forms and less overregularization), while younger children with poor representations of plural form declined in performance.

The implications of these findings is that even in children differences in initial state lead to differences in outcome. As adults and children come to the task of language learning with

different initial states stemming from differences in cognitive, neurobiological and linguistic development (Birdsong, 2009), it is clear from the results of Dye and Ramscar (2009) and Ramscar and Yarlett (2007) that even small differences in initial state at the time of both learning and testing can produce different learning outcomes. Furthermore, based on the results of the two previously reviewed studies, the dissociation between production knowledge and recognition in addition to the positive correlation between zero marked forms and the improvements in productions of correct irregular forms possibly indicates that recovering from erroneous inferences of overgeneralization requires children to reach a state of asymptote and learned equilibrium for linguistic aspects that are over generalised in use. As better associative learning that resulted from those experiments (implicit prediction error over the course of exposure trials) led to better production of correct irregular plural forms, which were previously overtaken by over-regularization of plural forms.

Additionally, the caveat that children must have representations in memory of the correct forms of irregular plurals indicates that some degree of critical mass of items must have been witnessed, for children to learn, improve, and overcome erroneous inferences. Taking the findings of the two previously reviewed studies, it is possible that overcoming erroneous inferences requires development of the ability to recognise regularities in the input, some degree of critical mass of witnessed items, and a critical mass of associative learning through prediction error.

The implication for adult users of a limited end-state of the target language that used unmarked lexical items are close to recovering from their own erroneous inferences about the error free nature of their language use and overgeneralization of unmarked lexical items in contexts where marked lexical items is more appropriate (e.g. worked instead of work in a past tense utterance). Essentially, adults that have stabilised at the limited end-state of a basic variety are likely to meet all the potential requirements to overcome erroneous inferences i.e. having representations in memory of language form cues and witnessing a critical mass of them for prediction error to correct their erroneous inferences about language use. However, the overly effective nature of the limited end-state of naturalistic usage-based adult SLA indicates that adults face the challenge of reaching a state of “learned equilibrium” (Ramscar & Yarlett, 2007). A state which according to the caveats of the R-W (1972) model requires an impetus for changes in the associative strengths of language form cues to be triggered. Adults are hindered in experiencing this impetus due to the effectiveness of the basic variety for everyday communication despite its ungrammaticality, due to the dominance of lexical items in accounting for the majority of

reduction of uncertainty satisficing communicative goals within the boundaries of the construct of economy of effort in both comprehension and production. This is further compounded by the lack of focus on form in implicit learning (Ellis, 2002, 2008b) the fleeting nature of spoken input, and the associative learning phenomenon of selective and learned attention (Cintrón & Ellis, 2016). It is possible that this is the case because their limited end-state adheres to the construct of economy of effort in exerting enough effort to reduce uncertainty for the listener in addition to the effort needed to avoid breakdowns of communication that require relatively more effort to repair than the effort needed to avoid it in the first place, which is reflected in the finding that non-native speakers are regularly over explicit in their utterances for the sake of reducing uncertainty for their listeners (Ryan, 2015).

Furthermore, due to the communicative effectiveness of a limited end-stat, at least at the level of a basic variety of the target language (Ellis, 2008a, 2008b; Klien, 1998; Klein, & Perdue, 1997), it is possible that from the perspective of the learner their language use is error-free (Klein, & Perdue, 1997). With any discrepancies between their model of language and the input encountered from a borderline variety of a fully-fledged language such as English is seen as discrepancies between the learner's acceptable variety of the target language (Klien, 1998) due to the lack of communicative breakdowns adding to the perception of their limited end-state being error free. In addition to the focus on communicative goals rather than form (Ellis, 2002, 2008b; Ramscar & Gitcho 2007), and the phenomenon of selective and learned attention. It quickly becomes apparent how the intake from "I worked two shifts yesterday" can be come "I workØ two shiftØ yesterday" and not cause any problems for the user of the basic variety in comprehension as any uncertainty that arises from the missing morphology can be easily resolved by the reduction of uncertainty provided by the accompanying lexical items two and yesterday. Furthermore, adults tend to interpret recasts containing implicit corrective feedback such as:

NNS: So one man feed for the birds.

NS: So one man's feeding the birds?

NNS: The birds. (Mackey et al., 2000; p 485) (NNS = non-native speaker, NS = native speaker)

As being semantic feedback rather than morphosyntactic, and only recognizing it as morphosyntactic feedback 13% of the time (Mackey et al., 2000). However, the

interpretation of morphosyntactic feedback as could reflect the lack of focus on form in implicit learning (Ellis, 2002, 2008b), selectively attending to the goal of successful communication (Ramscar & Gitcho, 2007) and that language experience has led them to learn that lexical gaps are the cause of uncertainty and reducing uncertainty through lexical cues usually reduces the uncertainty. This would be in line with the previously mentioned criteria for successful communication, being the reduction of uncertainty (Ramscar et al., 2010). If the usefulness of the basic variety in comprehension in addition to the associative learning phenomenon of selective and learned attention, and the focus on communicative goals rather than language form, hinder adult language learners from recognizing and learning from discrepancies between their model of language and the input they are exposed too, and instances of implicit corrections. Then it is clear that for instances of cognitive comparison between errors and corrections to occur, corrective feedback must be explicit and meta linguistically detailed (R. Ellis, et al., 2006).

If children are recovering through mostly implicit means through prediction error then adults encountering abundant input should also be going through a massive amount of prediction error as active listeners, especially with evidence of these forms potentially being in the mind of the learner based on the production results noted in the Ellis and Sagarra studies (2010b, 2011). However, a focus on success may be preventing the impetus to trigger changes in the cue's associative relationship to the outcome which is the second caveat of the R-W (1972) model from manifesting blocking the prediction error of language form cues from developing. Essentially a focus on success is creating a situation where there is no upper limit or boundary for association of overall utterance level success either due to complexity of utterance level associative learning or reaching a pseudo asymptote.

Where the overall utterance is not triggering granular level prediction error because it is succeeding, it is not being seen as an over generalised form that is being predicted and not appearing i.e. receiving some form of negative evidence as was the case for children predicting regular plural forms and encountering irregular plural forms (Dye & Ramscar, 2009; Ramscar & Yarlett, 2007) and allowing a U shaped pattern of associative learning. This is further compounded by a tendency towards an economy of effort that is underlied by the process of satisficing in selecting and maintaining a communicative path of least effort. It may be the case that adults require deep meta linguistic (R. Ellis, et al., 2006) to not only make these cues dimensions more salient but also initiate more repair cycles around this cue dimension. This would have the potential to increase the salience of this cue dimension and sensitivity towards it just as it is possibly the case that lexical items often being the

source of both communicative breakdowns and the means of repair (Ellis, 2008b; Klein, 1998; Klein & Perdue, 1997).

However, the construct of the collaborative tendency of both speaker and listener to economise the effort exerted in reducing uncertainty to achieve communicative goals. This means that the communicative interlocutor engaging with an adult learner is not incentivized to call for repair regarding language form if uncertainty caused by language form issues can be resolved internally with less effort than initiating a call for repair would entail; since it would be more costly in terms of communicative effort (words and turns used to do so) to initiate and engage in a repair cycle. This issue is likely further exacerbated by the observed behaviour that interlocutors prioritise maintaining discourse that is friendly and supportive, over pushing for input that is completely comprehensible through communicative breakdowns (Foster & Ohta, 2005). This again would point to interlocutors preferring to internally resolve uncertainty that is manageable with a reduction of effort potentially being part of the motivation, meaning that this could cause a scarcity of interlocutor scaffolding (Ellis, 2007; Schumann, 1978; Swain, 2005) often mean that a learner is not made aware of their error. According to Long (1985, 1996) this would essentially deprive adult learners of valuable instances to learn from repair, increasing their awareness, bringing their hypotheses of how language works is in tension with the dialectic corrective forces, and this awareness of the error allows for further development (Ellis, 2005), allowing gaps in knowledge to be identified in the form of a cognitive comparison between error and correction (R. Ellis et al., 2006).

Furthermore, listeners may also be relying on automatic parallel input processing to both reduce uncertainty in interpreting the speaker's meaning and function as an engaged listener, inhibiting them from processing the speaker's input at a level granular enough to focus on language form even before their economy of effort would inhibit them from initiating a repair cycle. This is to say that an economy of effort in conjunction with the context of naturalistic usage-based adult SLA is hindering the development of learners beyond a limited end-state; and that an economy of effort, if indeed present in communication, may be one variable that can be manipulated to trigger development beyond the limited end-state of naturalistic usage-based adult SLA that some adults experience. With one potential manipulation being that the listener no longer internally resolves uncertainty raised by issues of language form, and initiating repair cycles to draw attention to this cue dimension. The essential implication here is that the problem of developing beyond the limited end-state for adult learners in a naturalistic usage-based context does not fall squarely on the shoulders

of the learner alone; that in fact it is likely, learning in such a context, is sensitive to what the interlocutor needs for uncertainty to be acceptably reduced also contributes to what an adult learner acquires and which linguistic cue dimensions they focus on during implicit reception as listeners. This leads to the discussion of the role of the interlocutor as a source of feedback for a learners' economy of effort and language use in the following section.

4.4 The role of the interlocutor as a source of influence on what learners acquire

As previously noted in section 3, it is perhaps the case that the process of only repairing lexical gaps adds to the issues of salience, selective and learned attention, and relative redundancy that affect the acquisition of morphosyntactic cues, by making lexical cue dimensions more salient as they become strongly associated with the outcome of repaired communicative breakdowns. If true, this would explain, at least in part, why a basic variety of the target language is observed to continue in its development of its lexical repertoire (Klein, 1998; Klein & Perdue, 1997). However, it would also highlight the influence of the interlocutor on the learner if they are mostly initiating communicative repair due to lexical gaps in knowledge (Ellis, 2008b; Klein, 1998; Klein & Perdue, 1997), indicating to the listener that this gap is the cause for the breakdown. Therefore, interaction with an interlocutor allows the learner to test their hypothetical model of language with their interlocutor and initiating repair can be seen as a form of feedback for the learner, from the interlocutor.

This means that for lexical cues at least, interaction overall facilitated the further development of a specific cue that causes a breakdown and potentially further learning about that cue dimension as a whole. This is essentially the basis of the input hypothesis, that interaction facilitates learning (Long, 1996), through input and output exchanged during interaction and the feedback that a learner receives (Gass & Mackey, 2006, 2007), usually in the form of a repair being initiated highlighting the gap between the L2 speaker's production and their interlocutors. This would make a cognitive comparison between error and correction (R. Ellis et al., 2006) easier to identify due to their proximity, and increasing the awareness of this error needed for L2 development (Schmidt, 1990). However, as previously noted an interlocutor may not necessarily initiate repair for gaps in morphosyntactic knowledge due to social discomfort (Foster & Ohta, 2005), if these gaps can be resolved by the interlocutor through the functional redundancy that lexical items

offer; which act as a fail safe for the misuse or absence of language form, allowing the interlocutor to resolve uncertainty internally at a lower effort cost than initiating repair and the added benefit of avoiding social discomfort.

Therefore, interaction with an interlocutor can help the development of L2 learning and acquisition, but with a dependence on what causes the interlocutor to initiate repair. As such if gaps in morphosyntactic knowledge do not cause the interlocutor to initiate repair, there are no instances of cognitive comparisons between error and correction that help develop this domain of knowledge through interaction. Furthermore, even when these instances of morphosyntactic repairs do occur, they are often misinterpreted (Mackey et al., 2000) and often require to be made in the form of explicit metalinguistic feedback (R. Ellis et al., 2006). Additionally, the feedback that comes as a result of interaction in the form of repair can be seen to influence an economy of effort.

Where the lack of repair being initiated for gaps in morphosyntactic knowledge essentially indicate the fitness and success of solution to a communicative problem selected through the process of satisficing; and therefore, brings about a cessation of the process satisficing through reusing this solution, especially if implicit prediction error from language form is being inhibited due to a focus on communicative success. Conversely, if repair is initiated for gaps in morphosyntactic knowledge, resulting increased effort over multiple instances of repair, this would indicate to an economy of effort that mobilising effort in resuming the search for more accurate solutions that reach the new level of accuracy set by new aspiration levels, will likely result in a reduction of total probable effort. Where this added effort in resuming the process of satisficing is compensated for by the reduction in total probable effort when compared to continually repairing communicative breakdowns that result from reusing solutions that are no longer satisficing communicative needs. Furthermore, these instances of repair allow for the controlled serialised processing of input, as an instance of repair differs the interaction to a repair cycle that removes the now or never bottleneck associated with adult encountered input. This would allow more time for the learner to process input at a more granular level, and perhaps even allowing prediction error based on language form discrepancies to occur.

However, the implication here is that from the perspective of an economy of effort, simply providing modelling of different language behaviour is not enough to cause a learner to make use of that modelling without being enforced through a communicative breakdown and an initiation of repair focused on a specific aspect of language i.e. gaps in morphosyntactic knowledge. Although, a tendency towards an economy of effort would also

suggest that if modelled language demonstrated a clear reduction in total probable work, this model may opportunistically be adopted and used by the learner. Overall, it is clear that the interlocutor in interaction influences what the learner is likely to learn about and acquire. As such, manipulating what causes a communicative breakdown through interlocutor communicative behaviour presents itself as a means of manipulating the learner's economy of effort.

The following section will focus on expanding upon the notion of an economy of effort and its relation to satisficing. Additionally, it will discuss the limitations of the findings of current studies on associative learning and collaborative interaction. This is due to the fact that while they do show evidence of an economy of effort in interaction and associative learning, they are not specifically designed to measure this phenomenon and do not demonstrate the full dimensions and implications of an economy of effort in communication. Finally, the section also discusses the validity of operationalising communicative effort in terms of word count and turns taken.

5. Expanding on the notion of an economy of effort and limitations of adapting previous experimental paradigms

5.1 Elaborating on an economy of effort and its relation to the process of satisficing and associative learning

The aim of this subsection is to elaborate on the relationship between the notion of an economy of effort in communication and its relation to the process of satisficing. As shown by the review of studies in section 3 and 4 adults have a tendency towards focusing on communicative success in both native and non-native communicative interaction. With the criteria for communicative success being the reduction of uncertainty (Ramscar et al., 2010), it is clear that interlocutors are making decisions on how to communicate successfully based on what they assume is an utterance sufficient to reduce uncertainty to the point of achieving communicative success. This points to the process of satisficing under conditions of bounded rationality where actors have limited cognitive resources, time, and information. If an economy of effort influences communication, then under these conditions the process of satisficing becomes the process that determines the path of least effort for communicative success by rating potential solutions on their predicted ability to satisfice communicative aspirations or the minimum expected level of reducing uncertainty.

Looking at the definitions of both an economy of effort and the process of satisficing we can see a marked symmetry that relates these two together as a tendency and process by which this tendency appears in practice. Zipf (1949) defines the notion of an economy of effort as a universal tendency to minimise total probable work in achieving objectives, manifested in selecting and reusing the path of least effort. According to Zipf (1949) this includes the work of searching for and calculating the accuracy of the path of least effort, as a path that requires a longer search and more exhaustive calculation for accuracy is not considered economical, if this added work is not offset by effort saved in selecting said path. While satisficing is a heuristic process of searching for possible solutions to problems, amenable to trial-and-error, under conditions of bounded rationality (Simon, 1957, 1972, 1990). A key characteristic of the process of satisficing is stopping the search for solutions as soon as a probable solution is found, and trivialising future search for solutions by reusing tried and tested solutions previously found

(Simon, 1957, 1972, 1990). Selecting among a set of discovered solutions is further trivialised by the availability of aspiration levels to compare these solutions to (Simon, 1972) and selecting the first solution found that is expected to reach or surpass this aspiration level (Selten, 1999; Simon 1972). According to Simon (1972) the question becomes when to stop allocating limited resources to the search for satisficing solutions rather than which solution to choose. This account of the characteristics of the process of satisficing is indicative of an economy of effort that underlies the process of satisficing.

Based on this symmetry and similarity between both definitions, especially in selecting the first probably successful solution and when to stop the allocation of resources, it can be argued that for an economy of effort to make decisions regarding the selection of a path of least effort; it does so from previous experience or assumptions based on previous experience with the total probable work requirements of the objective. Under conditions of bounded rationality, it is therefore possible and likely that the process of satisficing is at least the process by which a path of least effort is selected through the leveraging of objective aspiration level to select said path, if not a process emergent from a tendency towards least effort in an attempt at rational decision making. Here aspiration levels represent the needed minimum level of performance from a potential solution to a problem (Selten, 1999; Simon, 1972; Weiner, 1995). They function as a stop criteria for the time and cognitive resource consuming search for potentially satisficing solutions, or the process of satisficing (Simon, 1972, 1990).

In terms of communication then the aspiration level is set by the desire to reduce uncertainty for the sake of communicative success, which is the lack of predictability of outcome (Berger & Calabrese, 1974; Kramer, 1999), due to many, equally probable outcomes being possible (Kaan, 2014). Failing to reach this minimum required reduction of uncertainty prevents the listener from correctly ascertaining intended meaning due to their bounded rationality. This subsequent inability to evaluate and analyse all possible interpretations (Simon, 1972), inhibits their ability to find a satisficing interpretation of the speaker's intended meaning, which can potentially result in a repair cycle due to their inability to internally resolve the uncertainty of the message.

This means that the desire to communicate successfully sets the aspiration or objective of reducing uncertainty, while the tendency towards an economy of effort means that the aim is to reduce uncertainty with minimal total probable work, a second aspiration level. Therefore, through satisficing the first probably successful solution would be the first solution that intersects with both aspiration levels, as there is an inherent reduction of effort in selecting the first probably successful solution manifest in the stop to the mobilisation of cognitive resources in the

search process. Furthermore, the reuse of these probably successful solutions after they have been found to be successful is another instance of similarity between both concepts would favour reapplying these solutions over searching for new ones.

Based on this relationship between the two concepts it is therefore likely the tendency towards an economy of effort is catered to by the process of satisficing which provides all the necessary criteria for a path of least effort to be selected. However, as satisficing is amenable to trial-and-error, under conditions of bounded rationality (Simon, 1957, 1972, 1990) the process reuses previously successful solutions for recurring problems or problems with similar aspiration levels, this indicates the characteristic of feature-based learning, or a form of error-driven associative learning Ramscar et al. (2010). Which would indicate that satisficing and therefore a tendency towards an economy of effort are influenced by the process of at least error driven associative learning due to their similarity in the characteristic of reusing solutions for recurrent and similar aspiration levels set objectives that are desired to be achieved. This further highlights the connection between an economy of effort and the associative learning of language and the potential for a focus on communicative success and achieving such success can inhibit the development of associations between language form cues and outcomes in a naturalistic second language acquisition context.

5.2 Expanding upon the notion of an economy of effort and the rationale for its operalisation through word and turn count

The aim of this section is to expand upon the notion of an economy of effort in communicative interaction and the rationale and validity of the use of word and turn count as means of operationalising and measuring the influence of an economy of effort in such a context; as Davies (2007) notes that both the notion and the means of operationalising it need to be better understood before an experimental can be considered to produce valid outcomes that can be interpreted with more certainty as representative of the influence of an economy of effort in communicative interaction.

To reiterate, Zipf (1949) defines the notion of an economy of effort as a universal tendency to minimise total probable work in achieving objectives, manifested in selecting and reusing the path of least effort. This includes the work of searching for and calculating the accuracy of the path of least effort, as a path that requires a longer search and more exhaustive calculation for accuracy is not considered economical, if this added work is not offset by effort saved in selecting said path. Based on the previous review of literature it would appear that an

economy of effort in collaborative interaction would manifest as the collaborative tendency of both speaker and listener to economise the effort they each exert in reducing uncertainty and achieving mutual understanding in social communicative interaction in order to achieve or satisfice communicative goals.

However, this view is not without potential criticism and needs further disambiguation for it to be clearly understood and subsequently operationalized for testing in an experimental paradigm. That is because although adults are collaborative co-participants in the development of dialogue (Schober & Clark, 1989), they are not necessarily driven by a joint economy of effort but rather a bias towards their own individual economies of effort. As Davies (2007) notes that, based on a review of collaborative interaction studies (e.g., Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986; Horton & Gerrig, 2002; Gergle et al., 2004; Bavelas et al., 2002; Brennan & Clark, 1996), it cannot be discounted that individual interlocutors are orientated towards a reduction in individual effort reduction and therefore decisions about effort are made on the basis of this orientation. This would imply then that both economies of effort are at odds in line with the view that economy of effort of the speaker the force of unifying all meanings into a single vocabulary item) is in conflict with that of the listener (force of diversifying all meanings to unique vocabulary items) (Zipf, 1949).

This view however, does not take into account that in conversation the roles of speaker and listener are interchangeable, that just as a speaker would wish to reduce their effort by maintaining and articulating a smaller vocabulary, the same applies to that person when they are the listener (i.e. maintaining a smaller vocabulary to comprehend), in addition to minimising effort in listening to, and comprehending utterances. Therefore, it cannot be discounted that there are two economies of effort at least in a collaborative interaction, and that their decisions regarding the mobilisation of effort in communication are taken with regards to the goal of being understood for the speaker and understanding for the listener, under conditions of bounded rationality (i.e. limited cognitive resources, time, and information). This means that their decisions are made under the assumption that they satisfice the minimum level of accuracy needed to achieve communicative goals. Further meaning that their decisions may be biassed towards their own economies of effort, but that their decisions on the mobilisation of effort are also taken with regards to what their interlocutor's economy of effort would allow in terms of the required reduction of uncertainty needed for communicative success.

Therefore, the observed economisation of effort that occurs between two interlocutors, that interchange in the role of speaker and listener, is not in necessarily the result of a focus on jointly reducing effort by two interlocutors, but rather it is the the emergent property of

individually biassed economies of effort aligning during communicative interaction. Meaning that interlocutors do not collaborate for the sake of a joint economy of effort, but for the sake of their own economies of effort, and that the observed collaboration is the process of building a mutually beneficial source of mutual understanding, that allows the speaker to produce less effortful utterances in articulation, and for the listener to comprehend meaning from a less auditorily effortful utterance to attend to and decode. This can be seen when listeners offer helpful continuers as in listener backchannels (Bavelas et al., 2000) that do not attempt to take a conversational turn but indicate understanding to the speaker (Cameron, 2001, Sacks et al., 1978) or even suggestions of referring terms to establish mutually understanding (Schober & Clark, 1989).

Consequently, the observed reduction in the number of words and conversational turns is not a deliberate product of a joint effort to economise but rather an emergent phenomenon resulting from the interaction of each interlocutor's individually biassed economies of effort. This alignment, albeit not consciously aimed at mutual economisation, underscores a fundamental aspect of communicative interaction: the natural tendency towards minimising effort for both interlocutors. However, this collaboration is susceptible to disruption in the form of communicative breakdowns that occur when the economy of effort of one interlocutor encroaches upon the other's. For the speaker, breakdowns occur when their economisation of effort leads to utterances that fail to sufficiently reduce uncertainty for the sake of the listener, whether due to brevity that exceeds mutual understanding, articulatory errors, or references outside the established mutual understanding. Conversely, for the listener, a failure to actively engage with the speaker in signaling understanding as observed from the [T] word counting participants of Bavelas et al. (2000), can similarly result in a communicative breakdown to be perceived by the speaker on the grounds that they have failed to convey their message.

However, Davies (2007) also calls into question the use of the number of words and turns used in collaborative interaction as a metric for measuring communicative effort as utterance length does not necessarily reflect the cognitive effort that underlies the creation of that utterance. Further adding that word count is not reflective of the quality of the solution itself. On the other hand, if we consider again that the speaker adhering to a reduction in total probable effort would likely prefer to avoid exerting additional effort in repair; then they are likely to produce an utterance that is assumed to be accepted by the listener to avoid this added effort. Furthermore, studies of cognitive effort (e.g. Griffiths et al., 2015; Kool et al., 2017; Lieder & Griffiths, 2015, 2016; Lieder et al., 2014; Shenhav et al., 2013; Shenhav et al., 2017) indicate that people learn to modulate their cognitive effort, manifested in reusing solutions or searching

for new solutions as a response to previous experience with solving problems that have similar task features. According to Ramscar et al. (2010) such feature based learning is a form of error-driven associative learning, meaning that when interlocutors reuse utterances that were previously successful that it is an indication of learning to solve that particular communicative problem and presumably similar communicative problems.

This reuse of communicative solutions implies that once mutual understanding is established, interlocutors are likely to leverage it in subsequent interactions, thus reducing the need for additional collaboration to further refine that same aspect of mutual understanding. Consequently, a decrease in word and turn counts can be seen not merely as a reduction in articulatory effort, but as a reflection of the reduction of cognitive effort and resources, since interlocutors are reusing previously found solutions rather than searching for new ones. From this perspective, while not directly measuring cognitive effort, a reduction in word and turn counts offers a proxy for the reduction of cognitive effort while also representing a reduction in articulatory effort. Therefore, it is plausible that for the purposes of operationalising and measuring an economy of effort in communication the use of the number of words and turns taken can be considered a viable metric.

5.3 Rationale for excluding language learning in experimental paradigm

Although the economy of effort is not a new notion in communicative interaction, as evidenced by studies of collaborative interaction (e.g., Clark & Brennan, 1991; Clark & Schaefer, 1987; Clark & Wilkes-Gibbs, 1986; Horton & Gerrig, 2002; Gergle et al., 2004; Bavelas et al., 2002; Brennan & Clark, 1996), it is, however, underexplored as a variable that has potential subsequent influence on naturalistic usage-based adult second language acquisition after influencing the naturalistic context in which they learn language through usage and experience. In other words, within the domain of naturalistic usage-based second language acquisition, the concept of an economy of effort remains at a preliminary stage, necessitating proof of concept. This proof must demonstrate that an economy of effort not only influences communicative interaction but also can be experimentally operationalised within this context. Additionally, it requires elaboration on the distinctions in the economy of effort between native and non-native speakers. Meaning that at this stage, it is difficult to interpret the results of an experimental paradigm that measures an economy of effort as a variable in naturalistic second language acquisition. These results would be challenging to view as wholly representative of this notion

influencing learning. This is because the influence of this notion on communication itself is not fully clear, nor is it certain how this influence differs for native versus non-native speakers. This is due to the limitations of the observations of an economy of effort influencing communication seen in the studies of collaborative interaction covered in section 3; and how the differences observed between native and non-native speakers in the same section would influence the performance of these two groups of participants in an experimental paradigm designed to measure the influence of an economy of effort. To elaborate the aforementioned studies (Bavelas et al., 2000; Bavelas et al., 2017; Schober & Clark, 1989) present two significant limitations in observations of an economy of effort.

First, the identification of an economy of effort within these studies emerges incidentally, either as a byproduct of observational data in the studies by Bavelas et al. (2000; 2017) or through the specific design of Schober and Clark's (1989) experimental Tangram game. Such incidental findings pose challenges for operationalising the concept of an economy of effort, particularly when its observed effects are limited to scenarios of achieving mutual understanding or addressing perceived communicative breakdowns; without evidence of how an economy of effort influences subsequent interactions, after interlocutors have gone through a repair cycle or more. As after a repair cycle, speakers may prefer more explicit communication methods, thereby increasing articulatory effort in future exchanges. This approach, though demanding more effort in the short term, potentially avoids communicative breakdowns, reducing the overall effort involved in the repair process. This includes the effort of understanding requests for clarification, developing and delivering a refined response, and verifying the success of the communication repair with the listener.

Until such dynamics are observed in an experimental setting designed to operationalise and manipulate the influence of an economy of effort, it cannot be definitively stated that an economy of effort, or a tendency towards it, plays a significant role in shaping communication and communicative interaction. The need for a robust experimental foundation means that embarking on a study to gauge the impact of an economy of effort on second language acquisition represents a considerable leap, predicated on a series of as yet unverified assumptions requiring prior testing and validation. Therefore, the outcomes and their interpretations from such a study risk being tentative and speculative. This is especially relevant considering that non-native speakers exhibit communicative behaviours that might seem at odds with the principles of an economy of effort (Ryan, 2015), further complicating the application of these concepts to the context of second language learning.

This leads to the second limitation, that an economy of effort influences non-native speakers differently to native speakers to the point where it appears to be at odds with the principles of an economy of effort. Where non-native speakers across the studies of their naturalistic interaction regularly exert more effort per utterance; however, as covered in section 3 this was to reduce uncertainty for the sake of avoiding communicative breakdowns that inherently require more overall effort. This means that non-native speakers approach an economy of effort differently or that it manifests in terms of word and turn count differently for them. This would further complicate the interpretation of the influence of an economy of effort on communication and subsequent language learning without first examining how non-native speakers perform in an experimental paradigm and comparing their performance to native speakers. Therefore, Prior knowledge of what is being measured is essential to ensure the accuracy of measurements and interpretations, in addition to the influence of who is being measured and the influence of their background.

Therefore, it becomes imperative to establish an experimental paradigm specifically designed to operationalise, measure, and manipulate the notion of an economy of effort in order to answer the general research question, does an economy of effort influence communicative interaction? Which raises the subsequent question: can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction? To answer this question, study 2 adopts and modifies the experimental design of the Schober and Clark (1989) Tangram task design by adding a pre and posttest to establish baseline performance and changes in that performance post interaction. Furthermore, the training design from Ellis and Sagarra (2010a, 2010b) is used to block the use of either literal or figurative descriptive language used in the study. These experimental methods are specifically adopted as the previous review of literature has shown that they are able to show evidence of an economy of effort in interaction and influence learners' communicative behavior through training and interaction.

This paradigm initially uses native speakers as a baseline, controlling for variables such as language ability, to serve as proof of concept. The aim is to demonstrate that an economy of effort influences communication and that this concept can be operationalised, measured, and manipulated accurately. Study 3 then aims to answer the research question: does an economy of effort influence communicative interaction differently for native and non-native speakers?This is achieved by applying the identical experimental paradigm to non-native speakers to allow for a direct comparison of the economy of effort's influence on both native and non-native speakers. However, this experimental design necessitated the use of bespoke visual stimuli, as

such the following section will first present the design, methods, and results of study 1 that norms these bespoke visual stimuli used in studies 2 and 3. The experimental design, and detailed methodology used for studies 2 and 3 is presented subsequently.

6. Study 1: Image norming

6.1 Visual stimulus norming study and methodology

This section describes the norming of the bespoke visual stimuli created for use in the experimental studies (Study 2 and Study 3) of this thesis. The process of stimulus norming is a process of data gathering regarding the responses and reactions, of a population of participants with similar characteristics to the target population of future studies, to the stimuli slated for use in said studies (Wurm & Cano, 2010). Stimulus norming allows researchers to be confident in the ability of their stimuli to elicit the desired responses from participants and control the variables influencing the results of their studies.

In Studies 2 and 3, participants need to successfully describe the stimuli (see sample in Figure 6) to their artificial interlocutor confederate, in order to complete a matching task. Studies 2 and 3 will have two conditions, in which the participant must either describe the image figuratively or literally (e.g., it looks like... vs there are two squares). As such the stimuli must be able to support both types of descriptions, while not being inhibitively complex or too simple to be useful. However, literal descriptions of these visual stimuli can appear rather artificial in nature to the participants of studies 2 and 3, and this was the rationale for the use of an AI confederate. The AI confederate provides a convincing context in which participants are more likely to accept the use of these literal descriptions, thereby allowing the examination of the influence of an economy of effort in communication in both an L1 and L2 context.

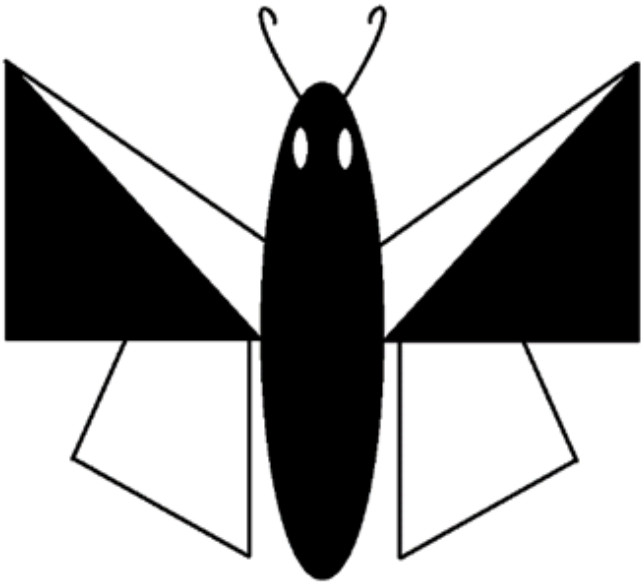


Figure 6: Sample of visual stimuli developed for Studies 2 and 3

The rationale for using figurative vs literal language as a means of describing images such as the one featured in Figure 6 is based on the observation that they offer two means of describing the same image with a potentially distinct difference in communicative effort in terms of word count; which also offer similarities in terms of salience and the degree of difficulty of application to linguistic cue dimensions such as lexical cues and morphological cues. Ellis (2017) notes that lexical and serialisation strategies used to express temporal reference (i.e. temporal adverbials) offer salient, constant, and easy to use means of expressing notions of temporal reference. Similarly, figurative language (i.e. using similes) offers a means of describing an image that is salient, constant, and easy to apply. Where the naming of animals or insects depicted in the image offers access to features of the image that are more distinct ways of discriminating between the images for the director to use when describe the images to the matcher similarly to identifying discriminating features of wugs and nizs in Ramscar et al. (2010) experiments.

This is because literal descriptions are similar to morphological cues in their characteristics, which Ellis (2017) notes these morphological cues as being non-salient they can vary by person or number while also typically having other irregularities associated with their use. Similarly, literal descriptions (e.g. Figure 6 has two black triangles, and a large black oval…) are less salient than the words butterfly or moth when considering that these constituent

parts of image morphology are shared across the images that appear on screen. This results in situations where butterfly is easier to use as a discriminating feature of the image, since other images on screen can have black triangles or a large black oval, and would require a combination of several of these segments of the image to be checked for their ability to discriminate and disambiguate these images from each other, and then combined into one literal description that is able to describe the specific image to the matcher.

This dichotomy between the number of parts that have to be described for literal descriptions to be successful when compared to the number of words needed for figurative descriptions highlights a potential difference in word count as the second rationale for the use of figurative vs literal language in image descriptions in studies 2 and 3. As previously mentioned, distracted [T] word counting participants were observed to increase the word count of speakers that interpreted their distractions as an indication of not understanding (Bavelas et al., 2000). Similarly, in a 2009 study Beukeboom observed that speaker participants that felt understood used more figurative language, while speaker participants that felt misunderstood used more literal language for the purpose of elaboration which appeared to function similarly as an attempt at communicative repair. These means of description appeared to provide two means of describing the same image that are likely to consistently produce differences in word count. Therefore, functioning as two dichotomous ways in terms of communicative effort that can still equally disambiguate an image from the group of images presented together. However, this difference in the number of words elicited by images needs to be established as consistent, and that the images used to elicit these different description types can support these two description types by conducting this image norming study.

The following sections detail the methodology and results for the visual stimulus norming study.

6.2 Summary of visual stimulus norming study and evaluated characteristics

This study is a norming study of the visual stimuli slated for use in studies 2 and 3 included in this thesis. In the present norming study participants are asked to provide descriptions of the visual stimuli they encounter (abstract clip art style images composed of geometric shapes - see Appendix F), as well as providing ratings of both visual complexity (intricacy) and appeal (aesthetics). The visual stimuli are normed to ensure that they are visually complex enough to support literal descriptions, and that naming agreement is high enough to enable Study 2 and 3 participants to use names in the figurative condition. The norming study also aims to confirm

that supplying literal descriptions is more effortful (requires longer descriptions) than figurative descriptions. The normed characteristics are as follows:

- Visual iconicity: Visual iconicity refers to the degree of resemblance between visual stimuli and the real-world objects they are meant to depict (Saryazdi et al., 2018). A high degree of visual iconicity (i.e., resemblance to real world objects) not only elicits good naming agreement from participants but also provides processing benefits for adults in recognizing and identifying objects in images as their real-world counterparts. Both characteristics raise the likelihood that participants will use labels for the images from a range of expected nouns in Studies 2 and 3.
- Productive effort: Productive effort refers to the average number of words participants use to describe each image figuratively or literally. This data is important to collect in order to establish that figurative and literal descriptions of the same image elicit a substantially different average word count for all images used. It is expected that figurative descriptions will have a substantially smaller average word count compared to literal descriptions due to their nature and how much information can be conveyed through a simile as opposed to dissecting an image and describing individual parts. If such a difference is established, this justifies the use of these description types in the subsequent studies to both establish the presence of an economy of effort in communication and to further study its potential influence on SLA.
- Visual complexity: Visual complexity refers to the amount of visual data present in an image such as the number of objects, colours, lines, and structures (Madan et al., 2018). It is important for the purposes of the subsequent studies that the images used are visually complex enough to support literal descriptions, while not being overly complex so as to hinder participants from interacting with those images during the course of upcoming studies. It is also important that images are roughly comparable in terms of visual complexity.
- Visual appeal: Visual appeal refers to a subjective measure of the aesthetic appeal of the image to the participants. This variable functions as a means of triangulation to validate the interpretation of visual complexity ratings, as visual appeal is correlated with visual complexity (e.g. Bauerly & Liu, 2008; Berlyne, 1974; Madan et al., 2018) (where the correlation between appeal and complexity follows an inverted U-shape pattern, as participants' ratings of appeal increase with the increase from low to moderate complexity before decreasing in response to high visual complexity (Geissler et al., 2006; Reinecke et al., 2013). That is to say that visual appeal ratings elaborate on

participants' ratings of complexity, as their interpretation of high or low complexity does not necessarily translate to good, bad, or distracting; and that ratings of appeal allow for a more valid interpretation of participants complexity ratings.

6.3 Materials

The visual stimuli to be normed in this norming study are 19 clipart style images resembling real world animals and insects that are composed of geometric shapes. 18 of these images are slated for use in studies 2 and 3, while the 19th and final image is a simpler image used as a throw away image for examples and instructions. These stimuli will be used in Studies 2 and 3 as the object of descriptions in a dyadic information gap task adopted from the methodology of Schober and Clark (1989). In the task, a directing participant with access to the information (a grid in which the shapes are placed) describes the image and its order on the grid to a matching participant without access to this information.

Studies 2 and 3 will have two conditions, in which the participant must either describe an image figuratively or literally (e.g., it looks like... vs there are two squares). As such these images must be able to support both types of descriptions, while not being inhibitively complex or too simple to be useful. Based on these characteristics a set of images (see Appendix F) were commissioned for the purposes of subsequent studies.

6.4 Summary of measures and procedure

This study was conducted online via the experiment builder Gorilla.sc with participants recruited from the online participant Prolific.com (n = 169). The target population for this study was university students that are native speakers of English, a measure taken to both keep in line with the populations of previous studies, and in order to increase the pool of potential participants due to the higher availability of native speakers of English in online databases. Once recruited, participants were given a link leading to the Gorilla.sc domain where the experiment is hosted. Participants were then asked to read the attached consent form and provide their consent before they are able to proceed with the norming trial. Participants are then presented with the tasks in the following order:

1. Introduction: Participants are greeted and introduced to the purpose and goal of this study as there is no deception involved.
2. Productive stage and instructions: Participants are informed that the aim is to see if these images can support descriptions and are given a completed example to illustrate

what they will be required to do. Then participants are asked to proceed and provide figurative and literal descriptions for each image they encounter.

3. Receptive stage and instructions: Participants are informed that the aim of this stage is to measure the level of visual complexity of each image in addition to its visual appeal. They are informed they will provide their measure of complexity and appeal on separate 5 point scales where 1 = low complexity / appeal and 5 = high complexity / appeal.

For each measure, procedure and scoring is described in turn:

1. Visual iconicity: Visual iconicity is measured on a binary scale (0/1) via the extraction of the figurative name given by participants in their text descriptions. If the name provided by a participant matches the name determined by the researcher or is of a suitable genus (e.g. image of butterfly being named moth) it receives a score of 1, otherwise it is scored as a zero. These scores are then tallied and converted into a percentage to reflect the degree of visual iconicity in the form of naming agreement.
2. Productive effort: Productive effort is measured through the mean number of words produced by participants in their descriptions. Each description type (figurative and literal) for each image has their mean word count calculated and compared in order to ascertain if each image can support both description types and if the mean word count is different enough to both establish the presence of an economy of effort in communication and to further study its potential influence on SLA.
3. Visual complexity: Visual complexity is measured on a 5 point Likert scale, with 1 being low complexity and 5 being high complexity. The mean, median, and standard deviation of these scores are calculated and correlated with scores of visual appeal in order to assess the viability of the level of complexity present in these images for use in subsequent studies.
4. Visual appeal: Visual appeal data is gathered identically to visual complexity.

6.5 Results

The following results represent the findings of the Image Norming Study conducted online using the Gorilla.sc experiment builder and gathering participants online through the Prolific.co platform. Data was gathered from 182 participants of which only the data from 169 participants was used in this analysis as they had completed all tasks with no missing data. Data from each of the dependent variables (visual iconicity, productive effort, visual complexity, visual appeal) is reported in turn, with reports focused on by-item descriptives. Three analyses are also reported: 1) the effect of condition (figurative vs literal) on productive effort is examined via a linear mixed

effects model, 2) the relationship between visual complexity and productive effort is examined via a Pearson's product-moment correlation, and 3) the relationship between visual complexity and visual appeal is examined via a Pearson's product moment correlation.

6.5.1 Visual iconicity

Visual iconicity was measured as the percentage agreement with the canonical name for the [18/19] images that were rated. Participants were asked to provide a figurative description of each image ("It looks like…"). Their responses were evaluated to see if they contained the canonical name of the image, or that of a related species or genus (e.g. butterfly > moth). We first report mean agreement, and then examine individual responses for descriptions that did not match the canonical name.

With 169 participant responses, agreement ranged between a max of 96.30% agreement, and a min of 15% agreement. Agreement proportions for all items are included in Table B1 in Appendix B. Most items (n = 18) did not have complete naming agreement. In order to understand the nature of the disagreement, individual responses for these items are reported in Table C1 in Appendix C. For the item "bird_over" which represents a generic bird from an overhead view, reached 100% naming agreement when accepting the genus related names (e.g., eagle, hawk). Similarly, the item "whale" had a naming agreement of 100% when accepting miss spellings and genus related names (e.g., fish). The items "ant, bird_front, butterfly, camel, duck, elephant, giraffe, hippo, panda, shark, sheep, spider, squid and turtle" had a naming agreement ranging from 78% to 98% when accepting miss spellings and genus related names. The naming agreement results for the aforementioned items indicates that these items do not require any amendments before use in study 2 as participants are both very likely to use these canonical names when describing the items themselves and be able to identify these items when they are described to them through their canonical names. As for the remaining two items "dog" (67% naming agreement) and "fox" (15% naming agreement), they were both most often mistakenly identified as "cat." In the case of the item "dog" 67% naming agreement indicates that the item is usable as is; however, it also indicates that in studies 2 and 3 the item should also be accepted as "cat" and described as cat in further exchanges with the same participant should the randomization measures of studies 2 and 3 make it the case that the participant is describing the item first. In the case of the item "fox" it reaches 80% naming agreement when the term "cat" is accepted rather than "fox" indicating that nature of the disagreement is one between the researcher and the general population of participants when interpreting what the image represents, leading to the conclusion that a simple fix to this

disagreement would be to change the canonical name of the item from “fox” to “cat” for use in study 2. Finally, as there are items that overlap in the real world objects they represent (e.g., bird_over vs bird_front) or can be confused for the same object (e.g., dog vs fox), one measure to avoid confusion and frustration for participants in studies 2 and 3 is to prevent these items from appearing together within the same trial blocks within studies 2 and 3.

6.5.2 Productive effort

Productive effort was measured in each condition (figurative and literal) as the number of words produced to describe each item. We first report the mean number of words elicited per item per condition, and check whether these are statistically different, and then examine the literal condition in more detail, to explore whether any items elicited significantly shorter or longer descriptions.

Word counts of descriptions in the figurative condition (M = 4.82, SD = 1.79) were shorter than descriptions of the same items in the literal condition (M = 37.72, SD = 30.83). We examined whether this difference was statistical by fitting a linear mixed effects model with a fixed effect of condition, random intercepts for items and subjects, and random slopes for description type (R model statement for Analysis 1: wordCount ~ namingType + (namingType|subject) + (namingType|item)). There was a main effect of description type (<U+03C7>^{2}(1) = 66.17; p < .001), confirming that the difference was significant. This finding indicates that eliciting these description types under the conditions of study 2 would allow the experimental design of the study to create situations where the potentially universal tendency towards an economy of effort and its influences can manifest and be investigated.

Next, we visually examined the distribution of word counts within each condition to see whether there are any items that elicited longer or shorter descriptions. Mean word counts in the literal condition are plotted in Figure 7, and figurative mean word counts are plotted in Figure 7. The visual examination of Figure 7 and Figure 8 indicated that there were significant differences in mean word count between items within the literal condition but not the figurative condition. A t.test matrix within each condition (R model statement for Analysis 2: t_test(wordCount ~ item) %>% adjust_pvalue(method = “BH”) %>% add_significance()) revealed that the literal condition had 19 instances of significant differences in mean word count, while the figurative condition did not yield any significant differences.

Table E1 in Appendix E shows the complete findings of the t.test matrix for all significantly different items in the literal condition. However, as discussed in the following section these differences in mean word count cannot be explained through a correlation

between visual complexity and the literal description condition. We also visually examined the random effects from the model (plotted in Appendix D) to examine whether the condition (figurative vs literal) influenced any items or participants more than others.

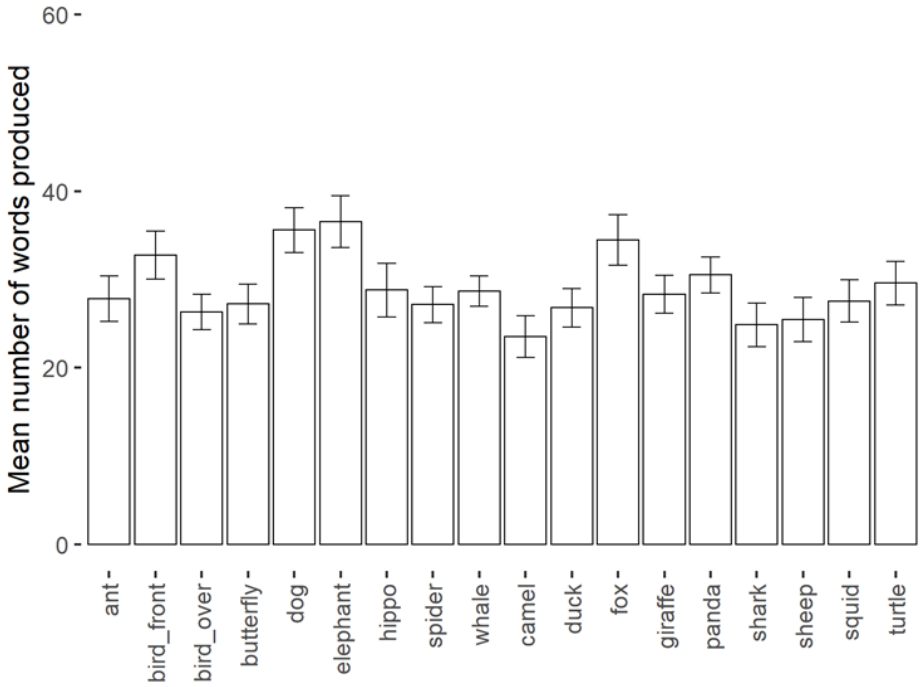


Figure 7. Productive effort measured as mean words produced in the literal condition

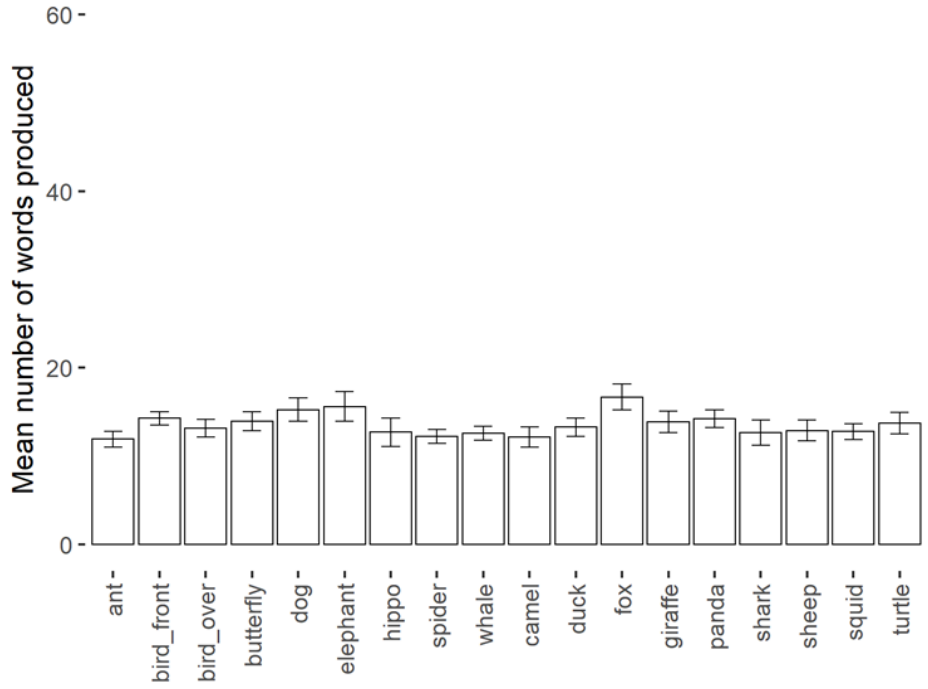


Figure 8. Productive effort measured as mean words produced in the figurative condition

6.5.3 Visual complexity

We asked participants to rate the visual complexity of the images on a 5-point scale from less complex to more complex. Mean visual complexity ranged from a minimum of *M* = 2.02, *SD* = 0.99, for duck, and a maximum of *M* = 3.57, *SD* = 1.08, for elephant. Mean ratings per item are plotted in Figure 9.

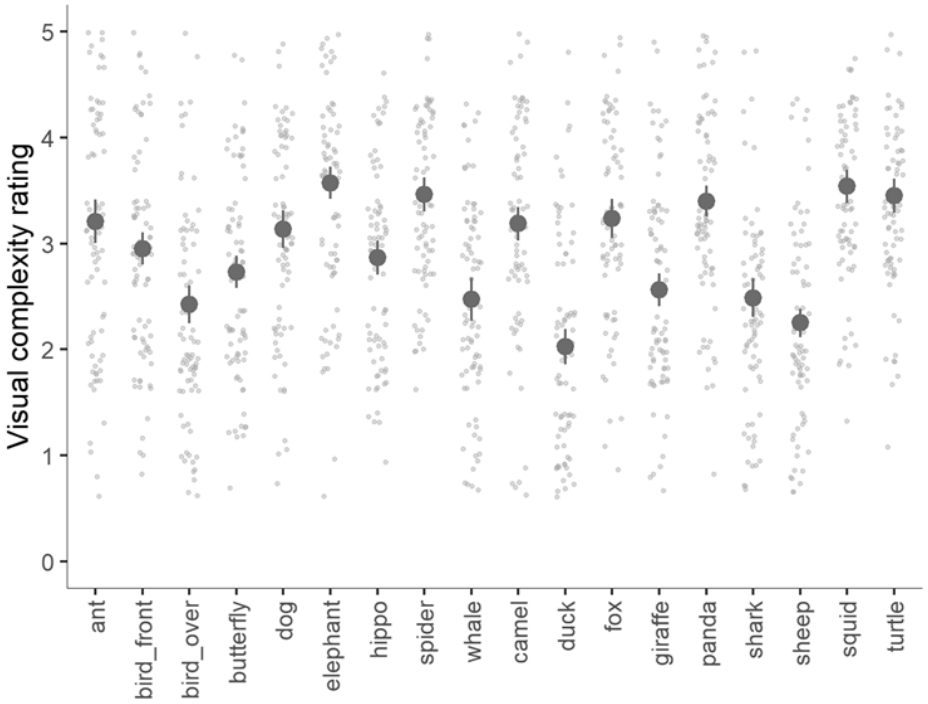


Figure 9. Mean visual complexity rating per item

In order to explore whether more visually complex figures would elicit longer literal descriptions, we correlated the mean visual complexity ratings with literal word length descriptions. A two.sided Pearson's product-moment correlation was not significant $r(16) = 0.43$, $p = 0.08$, indicating that there was no relationship between the self-reported complexity of the image, and the number of words used to describe it literally. However, A t.test matrix within the visual complexity condition revealed 96 instances of significant differences between items in visual complexity with 14 instances of overlap with the 19 instances of significant differences in word count in the literal condition. These overlaps in significant differences of mean word count and mean complexity rating between the literal condition and visual complexity indicate that despite the lack of correlation there might be a relationship between literal word count and visual complexity, but that the sample size maybe too small to reveal this correlation or that a 5 point scale is too small to allow participants to provide more nuanced ratings that distinguish between the complexity of the stimuli to a degree that correlates with literal word count.

6.5.4 Visual appeal

We asked participants to rate the visual appeal of the images on a 5-point scale from less appealing to more appealing. Mean visual appeal ranged from a minimum of $M = 2.06$, $SD =$

1.04, for duck, and a maximum of *M* = 3.36, *SD* = 1.21, for panda. Mean ratings per item are plotted in Figure 10.

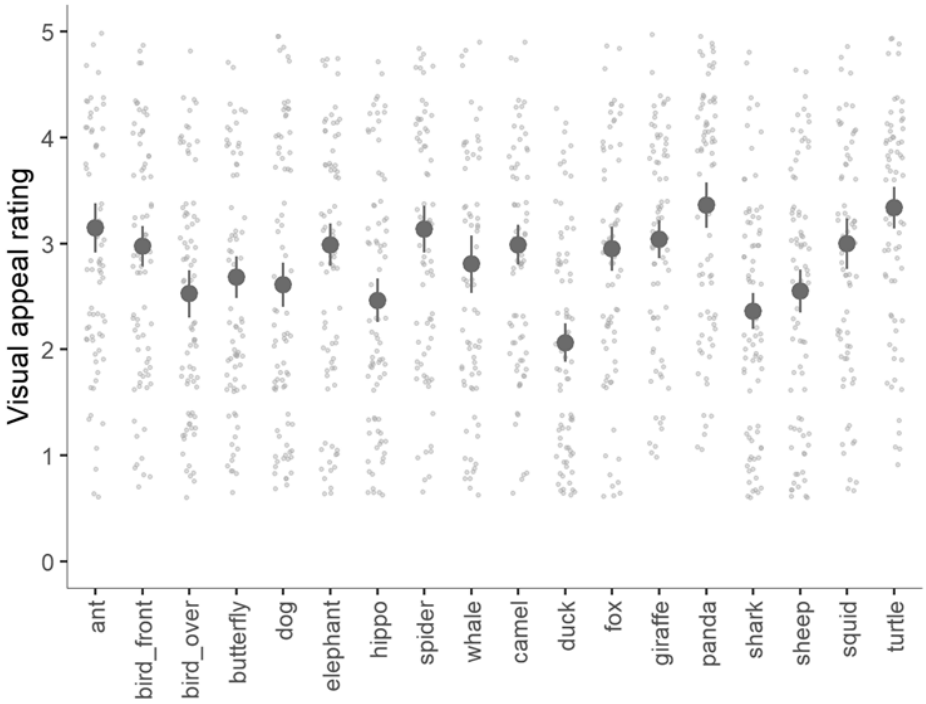


Figure 10. Mean visual appeal rating per item

In order to explore whether the visual appeal of figures was related to their visual complexity, we correlated the mean visual appeal ratings with visual complexity ratings. A two.sided Pearson's product-moment correlation was not significant $r(16) = 0.78$, $p = 0$, indicating that in the case of this study the ratings of visual appeal did not have a bearing on the interpretations of ratings of visual complexity . However, again it may be possible that a 5 point rating scale may not be providing the needed nuance to highlight a correlation between visual appeal and complexity with such a sample size. As previous studies (e.g. Bauerly & Liu, 2008; Berlyne, 1974; Madan et al., 2018) indicate that visual appeal and complexity do correlate, following an inverted U-shape pattern where participants' ratings of appeal increase with the increase from low to moderate complexity before decreasing in response to high visual complexity (Geissler et al., 2006; Reinecke et al., 2013).

6.6 Discussion

The purpose of this study was to first test if the visual stimuli created for use in studies 2 and 3 was suitable for the purpose of eliciting literal and figurative descriptions that were significantly different in mean word count between the two description types, and second to test if in the figurative conditions the visual stimuli were recognizable as their intended real-world counterparts (i.e., looking like the animal they were designed to represent). Based on the results reported in sections 6.5.1 Visual iconicity and 6.5.2 Productive effort, the images were mostly recognizable as their real-world counterparts (with the exception of the item fox) and elicited significantly different mean word counts in each description type (figurative = (*M* = 4.82, *SD* = 1.79), literal = (*M* = 37.72, *SD* = 30.83)). Based on the general results of these two sections it appears that the visual stimuli normed for studies 2 and 3 present no problematic characteristics that need any modifications or retooling before their use in studies 2 and 3 nor does it appear that there is a need to run an amended version of the study for more data gathering. The following sections discuss each variable individually.

6.6.1 Visual iconicity

The measure of visual iconicity was used to see if most of the figurative descriptions elicited by the visual stimuli matched the canonical names given to them to avoid problems in figurative descriptions during studies 2 and 3 that maybe caused by a mismatch in canonical names and what the participants perceive the images to be depicting. Based on the results reported in Table B1 (Appendix B) The majority of the images were not problematic in this measure and were suitable for use in studies 2 and 3 in terms of their visual iconicity. In the case of the problematic items “fox” and “dog,” “cat” will become the canonical name for the item “fox” and will be accepted for the item “dog” if that figurative description is used by a participant. The reason for this solution is based on the finding that “fox” reached 80% naming agreement when “cat” was accepted instead of “fox” and that “dog” reached 67% naming agreement with only “cat” being the incorrect description provided by participants. Furthermore, these two items as in the case for “bird_front, bird_overhead, and duck” will not appear in the same trial together during studies 2 and 3 as their overlap in potential names may cause unintended confusion for the participants.

6.6.2 Productive effort

The results for productive effort measured as mean word count elicited by each description type (literal vs figurative) showed that there was a significant difference in mean

word between description types. This finding indicates that each visual stimulus elicited significantly longer literal descriptions than figurative descriptions meaning that for the purposes of studies 2 and 3 these images allow for the opportunity to identify if and how an economy of effort may influence communicative behaviour. Additionally, significant differences in mean word count were identified between items within the literal condition; however, these differences were not explained through visual complexity as there was no correlation identified between mean word count and complexity rating.

6.6.3 Visual complexity

Visual complexity was added to the study as an explanatory variable for instances of a visual stimulus being problematic for participants to describe. (i.e., an image that consistently fails to elicit a literal or figurative description). None of the visual stimuli normed in this study presented as problematic in eliciting descriptions and there was no correlation between visual complexity and mean literal word count, as such in the case of the data collected it appears that visual complexity has little to no bearing on the interpretation of the analysis results of the current data collected. However, as previously noted it may be a case of small sample size, or that a 5-point scale was not nuanced enough to reveal a correlation.

6.6.4 Visual appeal

As noted in section 6.4.5 Visual appeal, the data gathered under the current conditions of the study showed no correlation between visual appeal and complexity, meaning that it is not possible with the current data to use ratings of appeal as a resource in interpreting ratings of visual complexity.

The following sections detail the experimental design and methodology for studies 2 and 3.

7. Study 2 methods

7.1 Introduction

The aim of this study is to both investigate the potential influence of an economy of effort in communication in a setting with more limited influencing variables (i.e. using participants L1 rather than a miniature artificial language) and demonstrate that this paradigm is able to operationalise and manipulate the influence of an economy of effort. This study adopts and modifies the Schober and Clark (1989) Tangram study design, where the core design of the study focuses on the cooperative matching of abstract images by two participants. The first participant the “Director” has access to the information needed to sort the images into the correct order and must relay this information to the second participant “Matcher” who is tasked with sorting the images into the same order as presented to the director using the director’s descriptions. The modifications made to this design is the use of a confederate that plays the role of an artificial interlocutor rather than a second non-confederate participant. The decision to use an AI confederate was made to accommodate the use of the artificial sounding literal descriptions, as these descriptions would not likely occur in natural interaction between human interlocutors but facilitate the testing of the influence of an economy of effort in interaction. In this case the confederate artificial interlocutor prefers the use of either figurative descriptions (e.g. it looks like a butterfly) or literal descriptions (e.g. there is a large black oval in the center with three triangular shapes on either side) for an image such as Figure 11.

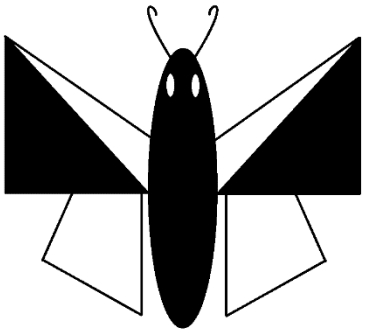


Figure 11: Butterfly.png

A further modification of the design of the study is the addition of pre and post testing stages and a 2AFC training stage adopted from Ellis and Sagara (2010a, 2010b, 2011) which is based on the Kruschke (2006) blocking paradigm designed to prevent a certain type of cue from being used as a tool for outcome prediction. In this study however the same design is used to discourage the use of one description type over the other while the pre and post-testing stages are used to measure changes in which description types participants use before and after completing the study.

The interaction of these description preferences that guide the "artificial interlocutor" with the training conditions of the participants create an environment suitable for the testing of the hypotheses which can answer the research questions of study 2 which summarily asks does an economy of effort influence communication? Table 3 below summarises participant training conditions and artificial interlocutor preference in a factorial design format. In the following sections the research questions and hypotheses are covered in more detail, along with the operationalization of the key variables that will allow these hypotheses to be tested. Furthermore, a detailed description of the stages of the study will be provided along with details of the interaction protocols between participant and confederate during the interactive Tangram task stage.

Table 3: IV interaction in factorial format

Independent Variables	Artificial Interlocutor Preference			
Training Conditions		Literal Preference	Figurative Preference	Semi-Preference
	Control group	1. Likely incongruent	4. Likely congruent	7. Semi congruent
	Literal description pre-training	2. Congruent	5. Incongruent	8. Semi congruent
	Figurative description pre-training	3. Incongruent	6. Congruent	9. Semi congruent

For Table 3 congruence or incongruence refers to the likelihood of a breakdown occurring during interaction due to the conflict (assumed or by design) between a participants description type training and the description preference of the AI; where a figurative training participant is likely to face a communicative breakdown when interacting with the literal preference AI which will not accept figurative descriptions. While a control group participant is assumed to fare similarly to a figurative training participant and the literal training participant is likely to not encounter breakdowns due to the similar description type preference.

7.2 Research questions and hypotheses

- Research question1: Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?
- Research question 2: Does an economy of effort influence communicative interaction differently for native and non-native speakers?
- Hypothesis 1: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that in the pre-test that without training participants should favour the use of figurative language-based descriptions.
- Hypothesis 2: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that untrained participants should show a bias towards figurative descriptions and perform similarly to the figurative description pre-training group throughout the experiment.
- Hypothesis 3: Without an increase in communicative effort due to breakdowns, participants will maintain their trained or untrained language type used for describing images (in the case of control group participants) unless participants come across less effortful means of achieving the goals of the communicative task.
    - For example, figurative and control group participants do not switch to literal descriptions in the semi condition since there are no breakdowns by design, while the literal group will switch to figurative descriptions in the semi condition even when there are no breakdowns due to an economy of effort.
- Hypothesis 4: Participants will switch to the description type of the AI that causes breakdowns when the AI refuses to understand the participants' descriptions.

- For example, figurative and control participants switching to literal descriptions and literal participants switching to figurative.

7.3 Operationalisation of key variables

As it is the aim of this study to operationalise, measure, and manipulate an economy of effort in communication, this section presents a number of key variables that must be operationalized in order for them to be validly tested. These variables are operationalised as follows:

- Communicative aspiration: operationalised as the objectives that a communicative interaction aims to achieve and sets the minimum level of accuracy needed from a satisficing solution.
    - For example, in this study the communicative aspiration is to describe an image to the matcher, while the minimum level of accuracy is the description type the matcher will accept (i.e. figurative or literal).
- Communicative effort: operationalised as the number of words and conversational turns used to satisfice communicative aspirations (i.e. complete the task).
- Economy of effort in communication: operationalised as a tendency towards a reduction in total probable communicative effort manifest as a reduction in the total number of words and conversational turns used to satisfice communicative aspirations, that can be collaborative.

7.4 Participants

This study was conducted with a total of 90 native speakers of English. Participants in this group were aged between 18 and 40 years old, and gender was not accounted for. Participants were recruited from Canada, Ireland, the United Kingdom, and the United States. Recruitment was limited to these countries due to a higher concentration of native speakers of English and concerns regarding connectivity issues that would cause the experiment to terminate. Participants were recruited through the Prolific.com platform, and were paid using the platform's hourly rate system, set at 15 GBP per hour. Prolific.com calculated payments based on participants' average time spent on the experiment. Participants were excluded from the study if they encountered technical issues preventing completion or if they exhibited non-compliance with the study format, such as attempting to skip to the end for payment purposes.

7.5 Materials

The materials used in this study were the same 19 images used in study 1 for image norming. Both the images and their descriptions (literal and figurative) are presented in Appendix F and G respectively. Finally, the researcher plays the role of the confederate artificial interlocutor which uses the literal and figurative descriptions to interact with participants in addition to a set of prewritten phrases to aid in interaction such as affirmations, calls for clarification, and sorting that are presented in appendix H.

7.6 Measures and Procedures

The study is composed of 4 potential stages for participants to complete which are detailed below, Table 4 displays each participant group and which stages they complete. Table 4 Experimental stages that participant groups take part in. Each stage is described in detail below, the type of data collected, and the dependent variables of each stage. The details for the procedure of analysis are presented in section 9, and the summary of the analysis strategy are presented in section 10.

Table 4: Experimental stages that participant groups take part in

	Control group	Figurative training group	Literal training group
Stage One Pre-Testing	✓	✓	✓
Stage Two Training	**X**	✓	✓
Stage Three Communicative Task	✓	✓	✓
Stage Four Post-Testing	✓	✓	✓

7.6.1 Stage 1 Pre-testing

In this stage participants are tasked with describing 5 random images from the pool of 19 images used in this study. Participants are instructed with the following prompt “For the following task please describe each of the 5 images so that another person would be able to

identify it from a group of images.” Each image is presented individually with a text entry field to enter their description of the image as shown in Figure 12.

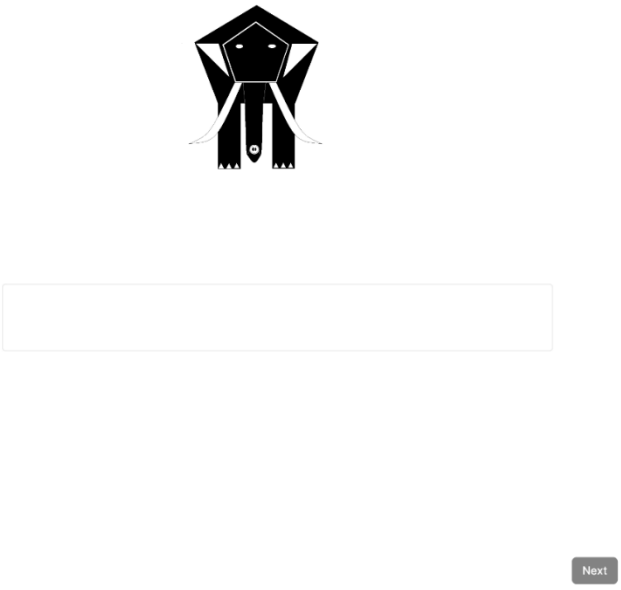


Figure 12: Sample Pre-testing Task.

The goal of this stage is to collect data on participants' typically favoured descriptive language (figurative vs literal) and their average word count used to describe the images. This helps indicate if there was any effect for training and changes in word (increase or decrease) during the communicative task in stage three.

- Data collected: Written description of each visual stimulus.
- Dependent variables: word count and language type used.

7.6.2 Stage 2 Training

Participants engage in a 2AFC task where they must choose 1 of 2 descriptions of an image. Both descriptions are correct; however, one is a literal description, and one is a figurative description. Participants are given the following instructions “You are required to click on the description you think can be best understood by an A.I. personal assistant.” before completing the task. There are a total of 12 items for participants to choose a description for and they are given corrective feedback that guides them towards the description type that is designated for their group (literal group gets an error message if they select a figurative description and vice versa). After a participant selects one of the two forced alternatives they are transitioned to an

intervening screen where they are given their feedback in the form of “Correct!” or “Incorrect, the correct answer was…” and must click on the “next” button to progress to the next training item, Figures 13 and 14 provide examples of a training item and corrective feedback respectively.

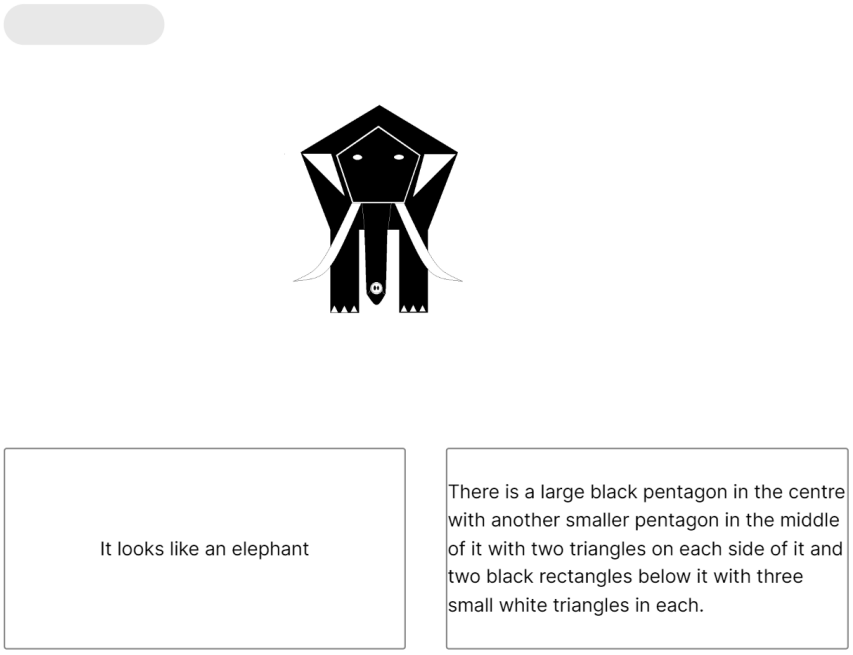


Figure 13: Training Item

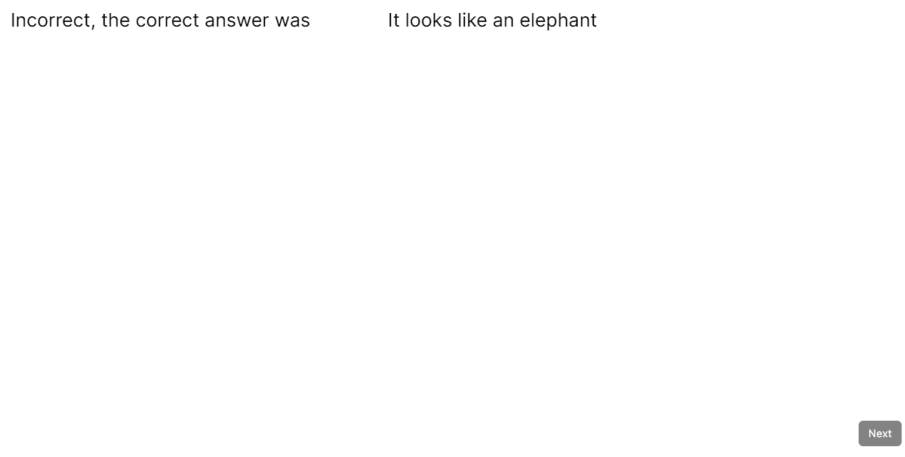


Figure 14: Corrective feedback during training

- Data collected: Click responses to visual stimuli scored as correct or incorrect based on training group.

- Dependent variables: First and second half performance accuracy.

7.6.3 Stage 3 Tangram task

Description of stage: Participants engage in a cooperative matching task with an assumed artificial interlocutor (researcher confederate). When participants are playing the role of the director, they are tasked with describing images from a grid for the artificial interlocutor to identify and place them in the correct zone; when participants are playing the role of the matcher they are tasked with identifying and placing the target image in the correct zone using the artificial interlocutor description.

Procedure of the stage: Participants engage with the “artificial interlocutor” in a cooperative matching task where they alternate between the role of the Director and Matcher. The Director is tasked with describing the images and which order they are placed in a grid for the matcher; while the matcher is tasked with identifying and placing the described images in the correct “drop zone” on the grid.

Participants always start as the director in order for the study to be able to capture the effect of the interaction as during piloting, participants that encountered an artificial interlocutor with preferences that differed from their training would adopt the artificial interlocutor’s preferences upon their first director turn. Making it difficult to ascertain if their training was effective, and they only matched the artificial interlocutors' preferences due to interaction or if their descriptions were unaffected by the training. The stage consists of four grids of six images and participants alternate roles with the artificial interlocutor meaning they play the role of director twice and matcher twice.

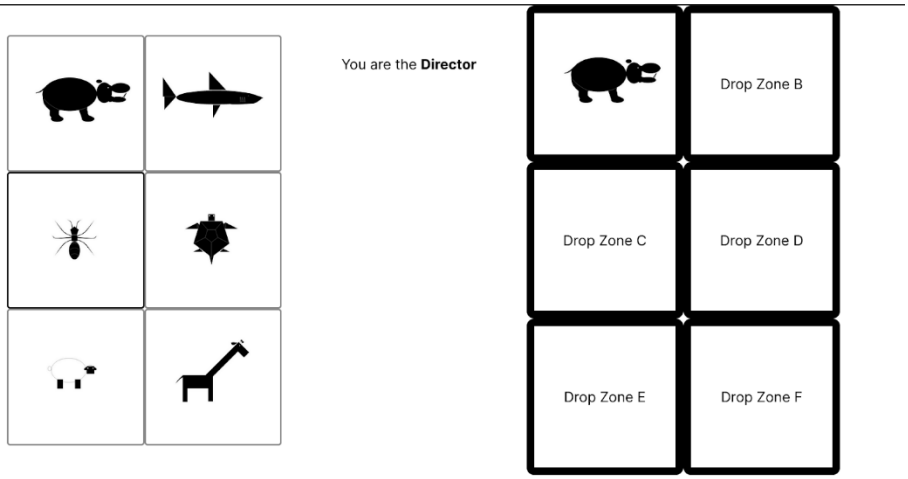
The procedure for the artificial interlocutor is fully scripted and the full list of phrases is available in appendix H. Participants are greeted with “hello” and the confederate “artificial interlocutor” waits for the response. Participants are then informed that [they] “You are the Director, and you will be describing the images for me to match. Are you ready?”. Once participants indicate their readiness for being the task they are asked to describe the first image, if the description given matches the artificial interlocutor’s preferences in the condition the image is dragged and dropped into the designated drop zone and an affirmation question is sent e.g. “is that the right one” which once confirmed moves the focus on to the next image.

However, if the description used does not match the artificial interlocutor’s preference in that condition, then participants receive a call for clarification e.g. “Hmmm, I’m not quite sure which image you are referring to.”. Participants will receive a second call for clarification if they use the same description type even if reformulated e.g. “I’m sorry I don’t understand, could you

try again please.”. Finally, if participants again use the same description type the protocol is to indicate that their description was not understood “I can’t find a match for that description.” and to provide a potential alternative to their description with the phrase “Did you mean the one that looks like a…” for a figurative description preference artificial interlocutor and “Did you mean the one that has…” for the literal preference artificial interlocutor.

At the end of the first director turn for the participant they are informed that the screen will transition and that they and the artificial interlocutor will switch roles. Once the screen transitions, participants are asked if they are ready, and once indicated they are informed that the image that will be described is for a specific drop zone “I will now describe the image that goes into Drop Zone A-F”. Once the participant drags and drops the image into the correct drop zone, they are met with an affirmation e.g. (Great job!, Awesome!, Fantastic!) or a correction “Sorry” if the image chosen is incorrect or if the drop zone is incorrect. Once all four grids have been completed participants are informed that the screen will again transition to a different task, that is the final task of the study. Figure 15 provides an example of a partially completed grid and Figure 16 provides an of a chat exchange.

Figure 15: Task grid



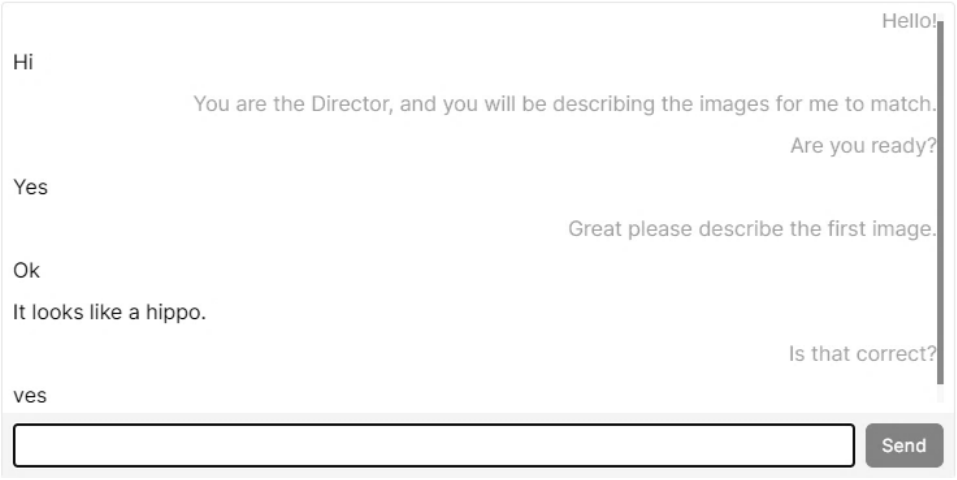


Figure 16: Sample chat exchange

Confederate concerns: As the role of the artificial interlocutor is played by the researcher who is aware of the goals of the study this raises the reasonable concern that this knowledge may indeed influence the results of the study (Mills et al., 2013). However, in the case of this study and study 3, the confederate is guided by a strict script for interaction with participants in a near binary fashion that would have been automated had the researcher been able to develop a suitable and reliable tool to do so. This will be clear from the data collected that participants were not in fact primed to adapt their responses in specific ways by the confederate.

- Data collected: Chat log sorted into turns between participant and AI.
- Dependent variables: total turns taken to complete stage, total breakdowns during stage, and switching language type used.

7.6.4 Stage 4 Post-testing

Description of stage: Participants are asked to describe 5 more images after they have completed the Tangram task. They are simply instructed to describe a final set of images with no other priming instructions.

- Data collected: Written description of each visual stimulus.
- Dependent variables: word count and language type used.

The following section extends this experimental design to non-native speakers.

8. Study 3 methods

Study 3 aims to extend the findings of Study 2 by examining whether the observed effects would be consistent among a population of non-native English speakers. To this end, Study 3 was designed as a direct replication of Study 2, employing identical methods and procedures to ensure comparability between the two populations.

8.1 Research question and hypothesis

- Research question 1: Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?
- Research question 2: Does an economy of effort influence communicative interaction differently for native and non-native speakers?
- Hypothesis 1: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that in the pre-test that without training participants should favour the use of figurative language-based descriptions.
- Hypothesis 2: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that untrained participants should show a bias towards figurative descriptions and perform similarly to the figurative description pre-training group throughout the experiment.
- Hypothesis 3: Without an increase in communicative effort due to breakdowns, participants will maintain their trained or untrained language type used for describing images (in the case of control group participants) unless participants come across less effortful means of achieving the goals of the communicative task.
    - For example, figurative and control group participants do not switch to literal descriptions in the semi condition since there are no breakdowns by design, while the literal group will switch to figurative descriptions in the semi condition even when there are no breakdowns due to an economy of effort.
- Hypothesis 4: Participants will switch to the description type of the AI that causes breakdowns when the AI refuses to understand the participants' descriptions.
    - For example, figurative and control participants switching to literal descriptions and literal participants switching to figurative.

8.2 Operationalisation of key variables

The key variables are operationalised identically to Study 2 and are presented again below:

- Communicative aspiration: operationalised as the objectives that a communicative interaction aims to achieve and sets the minimum level of accuracy needed from a satisficing solution.
    - For example, in this study the communicative aspiration is to describe an image to the matcher, while the minimum level of accuracy is the description type the matcher will accept (i.e. figurative or literal).
- Communicative effort: operationalised as the number of words and conversational turns used to satisfice communicative aspirations (i.e. complete the task).
- Economy of effort in communication: operationalised as a tendency towards a reduction in total probable communicative effort manifest as a reduction in the total number of words and conversational turns used to satisfice communicative aspirations, that can be collaborative.

8.3 Participants

This study was conducted with a total of 90 non-native speakers of English. Participants in this group were aged between 18 and 40 years old, and gender was not accounted for. Participants were recruited from Austria, Belgium, Denmark, Canada, Finland, France, Germany, Iceland, Italy, Ireland, Luxembourg, Netherlands, Norway, Poland, Portugal, Spain, Sweden, Switzerland, the United Kingdom, and the United States. Recruitment was limited to these countries due to concerns regarding connectivity issues that would cause the experiment to terminate. Participants were recruited through the Prolific.com platform, and were paid using the platform's hourly rate system, set at 15 GBP per hour. Prolific.com calculated payments based on participants' average time spent on the experiment. Participants were excluded from the study if they encountered technical issues preventing completion or if they exhibited non-compliance with the study format, such as attempting to skip to the end for payment purposes. In terms of proficiency level, the Prolific platform did not offer further screening options at the time, as such it was unknown. However, throughout the course of interaction L2 participants showed a good command of the target language and would be difficult to distinguish from native speakers within the chatroom format used in studies 2 and 3.

8.4 Materials

The materials used in this study are identical to the materials used in Study 2. Both the images and their descriptions (literal and figurative) are presented in Appendix F and G respectively, with the confederate researcher's script for prewritten phrases available in appendix H.

8.5 Measures and procedures

This study is again composed of the same 4 experimental stages and measure in this study are again identical to Study 2 with participants going through the 4 stages listed below in brief:

- Stage 1 pre-testing stage.
    - Data collected: Written description of each visual stimulus.
    - Dependent variables: word count and language type used.
- Stage 2 training stage.
    - Data collected: Click responses to visual stimuli scored as correct or incorrect based on training group.
    - Dependent variables: First and second half performance accuracy.
- Stage 3 Tangram task.
    - Data collected: Chat log sorted into turns between participant and AI.
    - Dependent variables: total turns taken to complete stage, total breakdowns during stage, and switching language type used.
- Stage 4 post-testing.
    - Data collected: Written description of each visual stimulus.
    - Dependent variables: word count and language type used.

The details for the procedure of analysis are presented in section 9, and the summary of the analysis strategy are presented in section 10.

9. Procedure of analysis for studies 2 and 3

This section describes the producer of data processing and analysis that was carried out on the data collected for the two general participant groups (i.e. native and non-native speakers of English). The study included four stages (stage 1 pre-testing, stage 2 training, stage 3 Tangram

task, stage 4 post-testing), processing and analysis are in turn described per stage and per dependent variable. This study aims to answer the following research questions by testing the hypotheses that follow them for both the population of native speakers (L1 English) and non-native speakers (L2 English).

- Research question 1: Can the notion of an economy of effort be experimentally operationalised in the context of communicative interaction?
- Research question 2: Does an economy of effort influence communicative interaction differently for native and non-native speakers?
- Hypothesis 1: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that in the pre-test that without training participants should favour the use of figurative language-based descriptions.
- Hypothesis 2: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that untrained control group participants should show a bias towards figurative descriptions and perform similarly to the figurative description pre-training group throughout the experiment.
- Hypothesis 3: Without an increase in communicative effort due to breakdowns, participants will maintain their trained or untrained language type used for describing images (in the case of control group participants) unless participants come across less effortful means of achieving the goals of the communicative task.
    - For example, figurative and control group participants do not switch to literal descriptions in the semi condition since there are no breakdowns by design, while the literal group will switch to figurative descriptions in the semi condition even when there are no breakdowns due to an economy of effort.
- Hypothesis 4: Participants will switch to the description type of the AI that causes breakdowns when the AI refuses to understand the participants' descriptions.
    - For example, figurative and control participants switching to literal descriptions and literal participants switching to figurative.

9.1 Participants characteristics descriptives

This study was conducted with a total of 90 native English speakers and 90 non-native English speakers recruited through the Prolific.com platform. Participants in both groups were aged between 18 and 40 years. Native speaker participants were recruited from Canada,

Ireland, the United Kingdom, and the United States, while non-native speakers were recruited from these countries as well as Austria, Belgium, Denmark, Finland, France, Germany, Iceland, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Spain, Sweden, and Switzerland. These countries were chosen based on connection stability for the online experiment. Participants were excluded from the study if they encountered technical issues preventing completion or if they exhibited non-compliance with the study format, such as attempting to skip to the end for payment purposes. Payment was made using Prolific.com's hourly rate system, set at 15 GBP per hour. Prolific.com calculated payments based on participants' average time spent on the experiment.

9.2 Pre-processing and scoring

This section describes how the data were processed for each of the four stages of the study and how the data were scored. All data, including that of both native and non-native speakers, was processed using R via the RStudio integrated development environment. Data processing and analysis were identical for both native and non-native speakers; however, it's important to note that the data for native and non-native speakers were processed independently of each other.

9.2.1 Stage 1 Pre-test

This stage of the study collected image description data as baseline data for pre-intervention tendencies and for comparison with post-test data for the dependent variables language type use, and word count. Participants provided five total descriptions one for each randomly presented image during the pre-test stage for a total of 450 entries.

9.2.1.1 Pre-test word count

The individual data frames for all participants were consolidated into one data frame. This step involved the removal of rows with missing values (NA) and extraneous columns, and the addition of indicator columns for starting groups and the AI language type preference that participants will interact with. To quantify the number of words used by participants in their descriptions, a custom function was employed to calculate the word count for each response, utilising the str_count function from the stringr package; this process involved identifying and tallying word boundaries within each text entry. The resulting word count for each response was then appended to the data frame as a new column, facilitating the analysis of word count as a dependent variable.

9.2.1.2 Pre-test language type use

To categorise the descriptive language type used by participants as either figurative or literal, a bespoke analytical method was devised. This involved the development of two functions to search text responses for predefined terms indicative of figurative or literal language, respectively. The grepl function, a base R function for pattern matching within strings, served as the core of this methodology. Lists of terms were curated to represent figurative language (e.g., animal names) and literal language (e.g., geometric shapes and sizes), with iterative refinements to accommodate common misspellings and variations.
Iteration continued until every response was successfully categorised. Each response was evaluated against these lists, and binary columns were added to the data frame to indicate the presence of figurative, literal, or mixed language types (e.g. binary value for both figurative and literal indicator columns was equal to 1), based on the occurrence of terms from the respective lists.

9.2.2 Stage 2 Training

This stage of the experiment was a training stage that only included the figurative and literal training groups and collected accuracy data for participant performance on a two answer forced choice training paradigm. Participants engaged in a total of 12 trials for a total of 1080 entries, and accuracy data on performance was gathered based on participants' selection of the description appropriate for their training group. E.g. a figurative group participant selecting a figurative description is coded as correct and selecting a literal description is coded as incorrect and vice versa for the literal training group participant.

9.2.2.1 Training accuracy

The individual data frames for all participants were consolidated into one data frame. This step involved the removal of rows with missing values (NA) and extraneous columns, and the addition of indicator columns for the starting group. To analyse performance accuracy, the data was first subset by starting group resulting in two groups (figurative and literal training) for overall accuracy. The data frame was then further subset by the intersect of starting group and first and second half (Block A and Block B respectively) of trials to compare performance accuracy between both halves to assess the effectiveness of the training trials and identify learning phenomena evidenced by changes in accuracy rates from Block A to Block B. Due to the nature of the data collected in this stage, only descriptive analysis was carried out. Using these data subsets, the summarise function was used to calculate mean, standard deviation,

and median accuracy for both the figurative and literal groups overall performance across 12 trials and first and second half accuracy performance across Block A and Block B.

9.2.3 Stage 3 Tangram Task

This stage of the experiment was an interactive stage that included all participant groups (Control, Figurative, and Literal) n = 90, and these groups intersected with the 3 AI language type preferences (Figurative preference, Literal preference, and Semi preference) resulting in 9 total groups with 10 participants in each group. This stage collected data in the form of chat logs capturing the communicative interaction between participants and the AI interlocutor. This data was parsed to focus on the main dependent variables of total turns taken to complete the interactive Tangram task, the total communicative breakdowns that may have occurred, and the binary observation in language type switching that may have occurred. The aim was to aggregate the findings of the analysis of these dependent variables to assess the influence of AI interlocutor interaction and subsequently the influence of a tendency towards an economy of effort in communicative interaction.

The individual data frames for all participants were consolidated into one data frame. This step involved the removal of rows with missing values (NA) and extraneous columns, and the addition of indicator columns for starting group, the Language type preference of the AI interlocutor each participant interacted with, and an indicator column for the intersection of two indicator columns. Chat data from interactions between participants and the AI interlocutor were parsed into individual utterances organised by turn. This parsing utilised the “fromJSON” function from the jsonlite package to accurately separate and structure the conversational data. Custom functions were developed and applied to parse the chat data further.

9.2.3.1 Tangram task total turns taken

The custom function for total turns taken identified turn boundaries was based on the script used by the confederate researcher playing the role of the artificial interlocutor available in appendix (H). Where indicator phrases like “please describe the first/next image” or calls for clarification such as “I’m sorry I don’t understand, could you try again please.” inherently indicates that a turn is completed or needs to be redone (i.e. another turn for the same image). Each turn was then counted and tallied for the total turns taken dependent variable per participant.

9.2.3.2 Tangram task total breakdowns

Similarly, the custom function for total breakdowns identified breakdowns was based on the script used by the confederate researcher playing the role of the artificial interlocutor available in appendix (H). Again, indicator phrases for breakdowns such as “I can’t find a match for that description.” inherently provided the boundaries for a communicative breakdown. Each instance of these phrases was counted as a communicative breakdown and was tallied for the total breakdown dependent variable per participant.

9.2.3.3 Tangram task language type switch

For the language type switch dependent variable, the same grepl function and list of terms developed for the identification of language type in the pre-test was used to identify the language type used in each turn by the participant. This generated the same binary columns for figurative, literal, and mixed language type as in the pre-test stage with 1 indicating the presence of that language type and 0 indicating the lack of that language type in each turn. A binary switch in language type was indicated when the binary indicator for the first language type identified switch from 1 to 0, and 0 to 1 in the subsequently identified language type in the following turn. This resulted in a binary switch indicator for language type switch with 0 indicating no switch was observed and 1 indicating a switch was observed.

9.2.4 Stage 4 Post-test

The post-test stage mirrored the pre-test in data collection, focusing on the dependent variables language type use and word count for comparative analysis with pre-intervention tendencies. Each participant provided descriptions for five images, resulting in an identical total of 450 entries. Data from all participants was consolidated into a single data frame, with the removal of missing values (NA) and irrelevant columns. Indicator columns for starting group and AI language preference were included. Word counts were again calculated using the str_count function from the stringr package, and the language type used in descriptions was categorised as figurative, literal, or mixed through the same bespoke functions employing grepl for pattern matching. This process identified and categorised each response, appending the results to the data frame for further analysis.

10. Summary of analysis strategy for each stage of the study

This section provides a summary of the analysis strategy for each stage of the study, explaining how and why the data was subset in the analysis for each stage of the study. Furthermore, this section describes the specific type of analysis carried out for each dependent variable. Again, all data, including that of both native and non-native speakers, was analysed using R via the RStudio integrated development environment. Analysis was identical for both native and non-native speakers; however, it's important to note that the analysis for native and non-native speakers were carried out independently of each other. Their results will be reported and discussed separately before a final joint discussion comparing the outcomes of both participant groups.

10.1 Stage 1 Pre-test

For the analysis of the data collected from the pre-test stage the data was subset as one group since all participants are in a pre-intervention state and the data collected at this stage is used as a baseline average for the dependent variables word count and typical language type use.

10.1.1 Pre-test word count analysis

For this dependent variable, the data was summarised for the descriptive statistics of mean word count and standard deviation. Furthermore, a Bayesian Regression Model (BRM) was used to estimate the mean number of words used to describe an image in the pre-test. The Brms model formula is `Word_count ~ (1|ParticipantID) + (1|Image), family = gaussian`. 95%CI are Credible Intervals, taking into account only the random effects of participant and image.

10.1.2 Pre-test language type analysis

For this dependent variable, each description collected from participants (n=450) was labelled using the previously mentioned grepl function for pattern matching, using either Figurative, Literal, or Mixed to label language types. The frequencies of each of these description types were expressed as a tally and as percentages of the total number of descriptions. Furthermore, a BRM was used to estimate the likelihood of using either Figurative, Literal, or Mixed language to describe an image in the pre-test. The Brms model formula is `Language_Type ~ (1|ParticipantID) + (1|Image), family = categorical(link = "logit")`. 95%CI are Credible Intervals, taking into account only the random effects of participant and image.

10.2 Stage 2 Training

For the analysis of the accuracy of performance on a two-answer forced choice training paradigm collected from the training stage, the data was subset to reflect the two groups taking part in this stage (i.e., figurative and literal training groups), and was further subset by the intersect of starting group and first and second half (Block A and Block B respectively) of trials to assess the effectiveness of the training trials and identify learning phenomena evidenced by changes in accuracy rates from Block A to Block B. the summarise function was used to calculate mean and standard deviation of accuracy for both all participants and the figurative and literal groups first and second half accuracy performance across Block A and Block B.

10.3 Stage 3 Tangram task

For the analysis of chat data collected from Tangram task, the data was subset into the three starting groups (i.e. Control, Figurative training, and Literal training) resulting in 3 groups with 30 participants in each group. Each group was then analysed by AI language type preference (i.e. Figurative, Literal, and Semi language type preference) resulting in a total of 9 groups with 10 participants in each group. This subset strategy was used because it allows the investigation of the influence of interacting with the different AI language type preferences within the same starting group on the dependent variables total turns taken, total breakdowns, and language type switch.

10.3.1 Tangram task total turns taken analysis

For this dependent variable the data was summarised for the descriptive statistics of mean total turns taken and standard deviation for each overall group and for each group analysed by AI language type preference. Furthermore, a BRM was used to estimate the total number of turns taken in the Tangram task on the log scale. The Brms model formula is `Final_Participant_Turn ~ AI_Preference_Code, family = poisson()` with no random effects taken into account as the data is aggregated.

10.3.2 Tangram task total breakdowns analysis

Similarly for this dependent variable the data was summarised for the descriptive statistics of mean total breakdowns and standard deviation for each overall group and for each group analysed by AI language type preference. Again, a BRM was used to estimate the total breakdowns that occurred in the Tangram task on the log scale. The Brms model formula is `Total_Breakdowns ~ AI_Preference_Code, family = poisson()` with no random effects taken into account as the data is aggregated.

10.3.3 Tangram task language type switch analysis

For this dependent variable the data was again summarised for the descriptive statistics of mean number of language type switches observed and standard deviation for each overall group and for each starting group analysed by AI language type preference. Similarly, a BRM was used to estimate the likelihood for a switch in language type to be observed during the Tangram task where the estimates represent the log-odds of observing a switch. The Brms model formula is `Switch_Binary ~ AI_Preference, family = bernoulli()` with no random effects taken into account as the data is aggregated.

10.4 Stage 4 Post-test

For the analysis of description data collected in this stage the data was subset identically to the Tangram task subsets for analysis, first by starting group then analysed by AI language type preference. Again, this strategy was used because it allows the investigation of the influence of interacting with the different AI language type preferences within the same starting group on the dependent variables word count and language type use.

10.4.1 Post-test word count analysis

For this dependent variable the data was summarised for the descriptive statistics of mean word count and standard deviation for each starting group analysed by AI language type preference. Furthermore, a BRM was used to estimate the mean number of words used to describe an image in the post-test. The Brms model formula is `post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian()`, taking into account the random effects of participant and image.

10.4.2 Post-test language type analysis

For this dependent variable each description collected from participants (n=450) was labelled using the previously mentioned grepl function for pattern matching, using either Figurative, Literal, or Mixed to label language types. The frequencies of each of these description types were expressed as a tally and as percentages of the total number of descriptions for each starting group analysed by AI language type preference. Similarly, a BRM was used to estimate the likelihood of using either Figurative, Literal, or Mixed language to describe an image in the post-test. The Brms model formula is `post_test_language_type ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = categorical(link = "logit")`, taking into account the random effects of participant and image.

11. Native speaker results

This section presents the results of the study for native speakers of English (n=90) in order of each stage of the study (pre-test, training, Tangram task, post-test) covering the descriptives and Bayesian Regression Modeling results for each dependent variable.

11.1 Pre-test results

For the pre-test stage, there are two dependent variables: word count and language type use. Word count is first presented in terms of descriptive statistics for all participants overall, accompanied by the results of Bayesian Regression Modeling, which provides an estimate of the word count used in the pre-test. Language type use is subsequently presented through the analysis of the proportion of language type use across all 450 descriptions provided by participants, complemented by Bayesian Regression Modeling results that indicate the likelihood of a language type being used.

11.1.1. Pre-test word count results

For all participants (n = 90), the mean number of words used to describe an image in the Pre-Test was 13.58, with a standard deviation of 12.82. The distribution of words per image can be seen in the histogram in Figure 17.

Figure 17. Words per image frequency histogram - native speaker

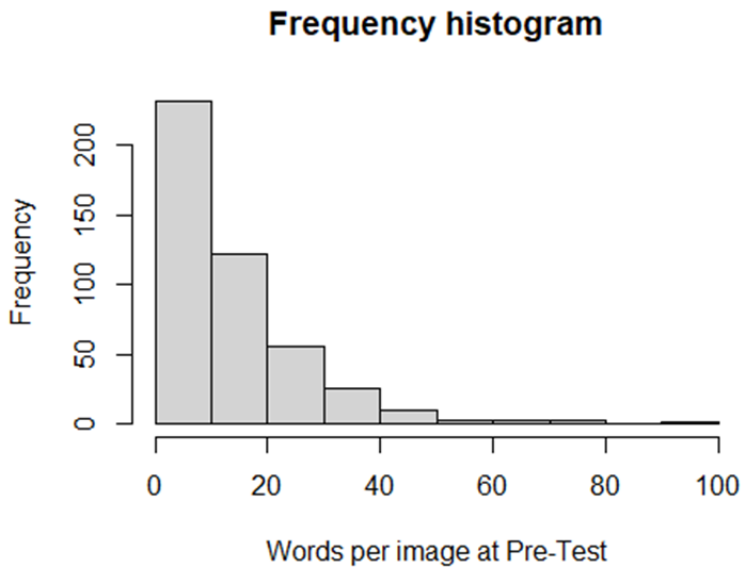


Table 5: Estimated word count (Pre-Test) - all native speaker participants

		estimate	std.error	95%CI
	Intercept	13.50	1.25	[11.00, 15.93]
Image	sd	1.73	0.63	[0.49, 3.08]
Participant	sd	10.94	0.89	[9.32, 12.85]
Residual	sd	6.83	0.26	[6.35, 7.37]

Note: Brms model formula is `Word_count ~ (1|ParticipantID) + (1|Image), family = gaussian`. 95%CI are Credible Intervals.

We estimated the mean number of words used to describe an image in the Pre-Test, taking only the random effects of participant and image into account. As can be seen in the model output in Table 5, random effect estimates of the standard deviations for images and participants confirm that there was relatively little variability in word count attributable to different images, but participants varied more widely.

11.1.2 Pre-test language type use results

Each description in the Pre-Test was categorised based on the language type used: Figurative, Literal, or Mixed. With all participants describing 5 images each, the total amounted to 450 descriptions. Among these, Figurative language was used in 233 descriptions (52%), Literal in 20 descriptions (4%), and Mixed in 197 descriptions (44%). Table 6 showcases the estimated likelihoods of using each language type (Figurative, Literal, or Mixed) for image descriptions by all participants. As noted below, the model confirms that without any training or instructions, people are less likely to use literal language than figurative language, but that they are equally likely to use mixed language and figurative language.

Table 6: Estimated language types (Pre-Test) - all native speaker participants

		estimate	std.error	95%CI
	Intercept (Literal language)	-4.34	0.99	[-6.56, -2.80]
	Intercept (Mixed language)	-0.38	0.35	[-1.10, 0.27]
Image	sd (Literal Language)	1.38	0.70	[0.23, 2.98]
Participant	sd (Literal Language)	2.18	0.71	[1.00, 3.84]
Image	sd (Mixed Language)	0.70	0.25	[0.24, 1.25]
Participant	sd (Mixed Language)	2.54	0.39	[1.87, 3.39]

Note: Brms model formula is `Language_Type ~ (1|ParticipantID) + (1|Image), family = categorical(link = "logit")`. 95%CI are Credible Intervals.

Table 6 contains the output for a multinomial model which estimated the likelihood of using either Figurative, Literal, or Mixed language to describe an image in the Pre-Test, taking only

the random effects of participant and image into account. There were three possible outcomes (Figurative language use, Literal language use, and Mixed language use) for each description. In our model, Figurative language use was used as the reference level. Therefore, we obtained two sets of coefficients, one for Literal language use compared with Figurative language use Intercept(Literal Language), and the other for Mixed Language Use compared with Figurative language use Intercept(Mixed Language). The model confirms that without any training or instructions, people are less likely to use literal language than figurative language, but that they are equally likely to use mixed language and figurative language. Random effect estimates of the standard deviations for individual images and participants confirm that there was relatively little variability attributable to different images, but slightly more to participants.

11.2 Training results

For the training stage, descriptive statistics of accuracy are presented in Table 7. The mean accuracy for all participants was high (M = 0.96, SD = 0.20), indicating that, on average, participants performed accurately during the training stage. Notably, both the figurative and literal training groups achieved perfect mean accuracy (M = 1.00, SD = 0.00) in the second half of the training trials. This suggests that participants effectively learned the correct description type for their respective training group.

Table 7: Training accuracy descriptives for native speakers - by group

Category	Mean Accuracy	SD
All groups	0.96	0.20
Figurative Group Block A Accuracy	0.91	0.29
Figurative Group Block B Accuracy	1.00	0.00
Literal Group Block A Accuracy	0.92	0.27
Literal Group Block B Accuracy	1.00	0.00

Note: This table displays the mean and standard deviation accuracy by group for the training assessment.

11.3 Tangram task results

For the Tangram stage task there are three dependent variables which are total turns taken, total breakdowns, and Language type switch; they are also presented in that order. For all dependent variables in this stage descriptive statistics are presented and accompanied by Bayesian Regression Modeling that provides an estimate of the number of turns and breakdowns on a log scale and the likelihood of a switch being observed as log-odds. The descriptive statistics for these variables are presented in Table 8 below.

Table 8: Tangram dependent variables descriptives for native speakers - by group

Group	Mean Switch	SD Switch	Mean Turns Taken	SD Turns Taken	Mean Breakdowns	SD Breakdowns
Control - Figurative Preference	0.7	0.48	27.2	3.39	1.5	2.59
Control - Literal Preference	1.0	0.00	29.3	1.25	4.9	1.91
Control - Semi Preference	0.7	0.48	25.3	1.70	0.1	0.32

Figurative - Figurative Preference	0.2	0.42	24.2	4.29	0.1	0.32
Figurative - Literal Preference	1.0	0.00	32.5	3.78	6.9	1.79
Figurative - Semi Preference	0.1	0.32	24.6	4.79	0.0	0.00
Literal - Figurative Preference	1.0	0.00	27.1	3.98	3.1	1.52
Literal - Literal Preference	0.2	0.42	25.7	3.40	0.1	0.32
Literal - Semi Preference	1.0	0.00	26.0	2.54	0.0	0.00

Note: This table displays the mean and standard deviation for the dependent variables Total turns taken, Total breakdowns, and Language type switch within the Tangram task.

Tables 9 and 10 below present the supplementary summary data for the variables Language type switch and total breakdowns.

Table 9: Number of participants that switched and experienced breakdowns for native speakers - by group

Group Indicator	Number of participants that switched language type	Number of participants that experienced a breakdown
Control - Figurative Preference	7	4
Control - Literal Preference	10	10

Control - Semi Preference	7	1
Figurative - Figurative Preference	2	1
Figurative - Literal Preference	10	10
Figurative - Semi Preference	1	0
Literal - Figurative Preference	10	10
Literal - Literal Preference	2	1
Literal - Semi Preference	10	0

Note: This table displays the number of participants that switched language type and the number of participants that experienced a breakdown by group

Table 10: Direction of language type switches for native speakers - by group

Group	Figurative > Literal	Figurative > Mixed	Literal > Figurative	Mixed > Figurative	Literal > Mixed
Control - Figurative Preference	1	1	2	3	0
Control - Literal Preference	9	0	1	0	0
Control - Semi Preference	4	0	1	2	0

Figurative - Figurative Preference	2	0	0	0	0
Figurative - Literal Preference	10	0	0	0	0
Figurative - Semi Preference	0	1	0	0	0
Literal - Figurative Preference	0	0	10	0	0
Literal - Literal Preference	0	0	0	0	2
Literal - Semi Preference	0	0	9	0	1

Note: This table displays the number of participants that switched language type in specific directions by group.

11.3.1 Tangram task total turns taken

For mean total turns taken the descriptive statistics show that within the Control group, participants interacting with the Literal preference AI demonstrated the highest mean number of turns taken to complete the Tangram task (M = 29.3, SD = 1.25). However, the BRM results for the Control group presented in Table 11 do not show a credible difference in the number of turns taken for the Literal preference AI compared to the Figurative preference AI (Estimate = 0.07, 95% CI = [-0.09, 0.24]).

Table 11: Estimated number of turns in tangram - native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	3.30	0.06	[3.18, 3.42]
Semi-Literal preference AI	-0.07	0.09	[-0.24, 0.11]
Literal AI	0.07	0.08	[-0.09, 0.24]

Note: Estimates are on the log scale. Brms model formula is `Final_Participant_Turn ~ AI_Preference_Code, family = poisson()`. `preference AI ` is treatment-coded with Figurative preference, the baseline to which other preference AI s are compared. Semi-Literal preference AI represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, participants engaging with the Literal preference AI recorded the highest mean number of turns taken (M = 32.5, SD = 3.78). The BRM results for the Figurative training group presented in Table 12 indicate a credible increase in the number of turns taken only when interacting with Literal preference AI (Estimate = 0.30, 95% CI = [0.13, 0.46]), compared to the Figurative preference AI.

Table 12: Estimated number of turns in tangram - native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	3.19	0.06	[3.06, 3.31]
Semi-Literal preference AI	0.02	0.09	[-0.16, 0.20]
Literal AI	0.30	0.08	[0.13, 0.46]

Note: Estimates are on the log scale. Brms model formula is `Final_Participant_Turn ~ AI_Preference_Code, family = poisson()`. `preference AI ` is treatment-coded with Figurative preference, the baseline to which other preference AI s are compared. Semi-Literal preference AI represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

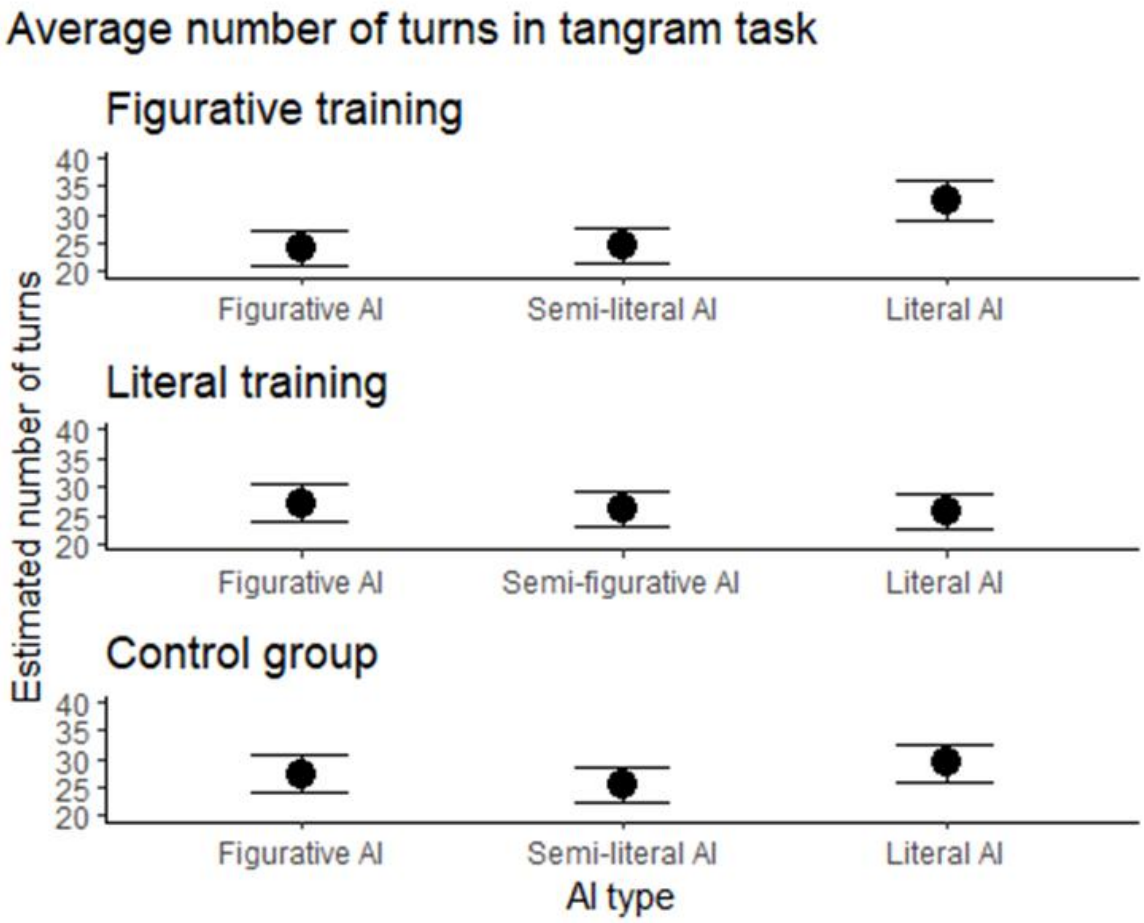
Within the Literal training group, the descriptive statistics show that participants interacting with the Literal preference AI had a slightly lower mean number of turns taken (M = 25.7, SD = 3.40) compared to interaction with the Figurative preference AI (M = 27.1, SD = 3.98). The BRM results for the Literal training group presented in Table 13 do not indicate a credible difference in the number of turns taken for the Literal preference AI as opposed to the Figurative preference AI. The estimates for both Semi-Figurative preference AI (estimate = -0.04, 95% CI = [-0.22, 0.13]) and Literal preference AI (estimate = -0.05, 95% CI = [-0.23, 0.11]) are close to zero and their 95% credible intervals include 0. This suggests that, within a Bayesian analysis framework, the observed differences in mean turns taken across AI preferences are inconclusive..

Table 13: Estimated number of turns in tangram - native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	3.30	0.06	[3.18, 3.42]
Semi-Figurative preference AI	-0.04	0.09	[-0.22, 0.13]
Literal AI	-0.05	0.09	[-0.23, 0.11]

Note: Estimates are on the log scale. Brms model formula is `Final_Participant_Turn ~ AI_Preference_Code, family = poisson()`. `preference AI ` is treatment-coded with Figurative preference, the baseline to which other preference AI s are compared. Semi-Figurative preference AI represents an AI who communicates using figurative language, but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

Figure 18. Estimated number of turns taken - native speaker



Note: Plots represent conditional effects and 95% Credible Intervals from Brms model formula Final_Participant_Turn ~ AI_Preference_Code, family = poisson(). Plot A runs on the subset of participants who underwent figurative training, Plot B on participants with literal training, and Plot C on the control participants. Semi-Literal or semi-Figurative AI preference represents an AI who communicates using the named language type, but accepts either language type from its interlocutor.

11.3.2 Tangram task total breakdowns

For total breakdowns within the Control group, interaction with the Literal preference AI resulted in the highest mean number of breakdowns (M = 4.9, SD = 1.91). The number of participants that experienced breakdowns within the Control group is also noticeably higher for

those that interacted with the Literal preference AI than other AI preferences with all 10 participants experiencing breakdowns (Table 9). The BRM results presented in Table 14 reveal that interaction with the Literal preference AI credibly increases the likelihood of experiencing breakdowns during the Tangram task (Estimate = 1.21, 95%CI [0.65, 1.82]).

Table 14: Estimated number of breakdowns in tangram - native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	0.37	0.26	[-0.17, 0.84]
Semi-Literal AI preference	-3.26	1.25	[-6.19, -1.34]
Literal AI	1.21	0.30	[0.65, 1.82]

Note: Estimates are on the log scale. Brms model formula is `Total_Breakdowns ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, again interaction with the Literal preference AI resulted in the highest mean number of breakdowns (M = 6.9, SD = 1.79). The number of participants that experienced breakdowns within the Figurative training group is also noticeably higher for those that interacted with the Literal preference AI than other AI preferences with all 10 participants experiencing breakdowns (Table9). The BRM presented in Table 15 reveals that interaction with the Literal preference AI credibly increases the likelihood of experiencing breakdowns during the Tangram task (Estimate = 4.66, 95%CI [2.82, 7.60]).

Table 15: Estimated number of breakdowns in tangram - native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	-2.74	1.19	[-5.64, -0.92]
Semi-Literal AI preference	-9.88	10.70	[-35.48, 0.96]
Literal AI	4.66	1.19	[2.82, 7.60]

Note: Estimates are on the log scale. Brms model formula is `Total_Breakdowns ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

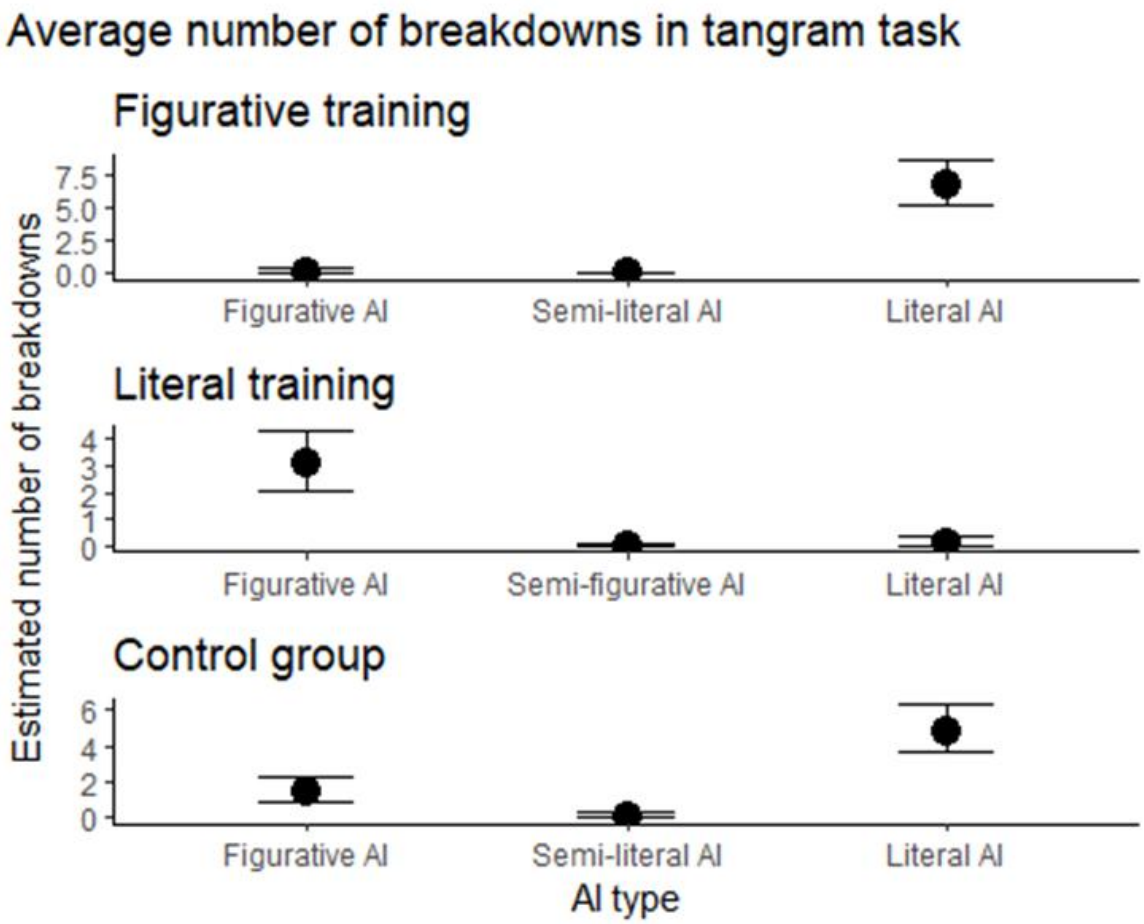
Within the Literal training group, it was found that interaction with the Figurative preference AI resulted in the highest mean number of breakdowns (M = 3.1, SD = 1.52). The number of participants that experienced breakdowns within the Literal training group is also noticeably higher for those that interacted with the Figurative preference AI than other AI preferences with all 10 participants experiencing breakdowns (Table 9). The BRM presented in Table 16 reveals that interaction with the Figurative preference AI credibly increases the likelihood of experiencing breakdowns during the Tangram task with a negative estimate for interaction with the Literal preference AI compared to interaction with the Figurative preference AI (Estimate = -3.82, 95%CI [-6.81, -2.04]).

Table 16: Estimated number of breakdowns in tangram - native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	1.11	0.18	[0.73, 1.45]
Semi-Figurative AI preference	-12.52	7.66	[-33.12, -3.77]
Literal AI	-3.82	1.23	[-6.81, -2.04]

Note: Estimates are on the log scale. Brms model formula is `Total_Breakdowns ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

Figure 19. Estimated number of breakdowns in Tangram - native speaker



11.3.3 Tangram task language type switch

For language type switch within the Control group interaction with the Literal preference AI resulted in the highest mean language type switch (M = 1, SD = 0), with a language type switch being observed for all 10 participants (Table 9) that interacted with the Literal preference AI. The main switch direction that resulted from interacting with the Literal preference AI was from figurative to literal language type for 9 of 10 participants (Table10). The BRM presented in Table 17 indicates a credible estimated increase in the log-odds of switching language type

during the Tangram task after interacting with the Literal preference AI (Estimate = 9.94, 95%CI [0.97, 36.85]). Meanwhile, interaction with the Semi-Literal preference AI, performance was notcredibly different than the baseline interaction with the Figurative preference AI.

Table 17: Likelihood of switching language type during Tangram task - native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	0.87	0.72	[-0.45, 2.42]
Semi-literal AI	0.02	1.03	[-1.95, 2.14]
Literal AI	9.94	9.86	[0.97, 36.85]

Note: Estimates are log-odds. Brms model formula is `Switch_Binary ~ AI_Preference, family = bernoulli()`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, again interaction with the Literal preference AI resulted in the highest mean language type switch (M = 1, SD = 0). Furthermore, interaction with both the Figurative and Semi-Literal preference AIs resulted in noticeably smaller means for language type switch (M = 0.2, SD = 0.42) and (M = 0.1, SD = 0.32) respectively. The main switch direction that resulted from interacting with the Literal preference AI was from figurative to literal language type for all 10 participants (Table 10). The BRM presented in Table 18 indicates a credible estimated increase in the log-odds of switching language type during the Tangram task after interacting with the Literal preference AI (Estimate = 2.18, 95%CI [0.79, 3.54]). While interaction with the Semi-Literal preference AI did not result in performance that was credibly different than the baseline interaction with the Figurative preference AI.

Table 18: Likelihood of switching language type during Tangram task - native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	-0.72	0.53	[-1.79, 0.27]
Semi-literal AI	-0.84	0.72	[-2.30, 0.51]
Literal AI	2.18	0.71	[0.79, 3.54]

Note: Estimates are log-odds. Brms model formula is `Switch_Binary ~ AI_Preference, family = bernoulli()`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

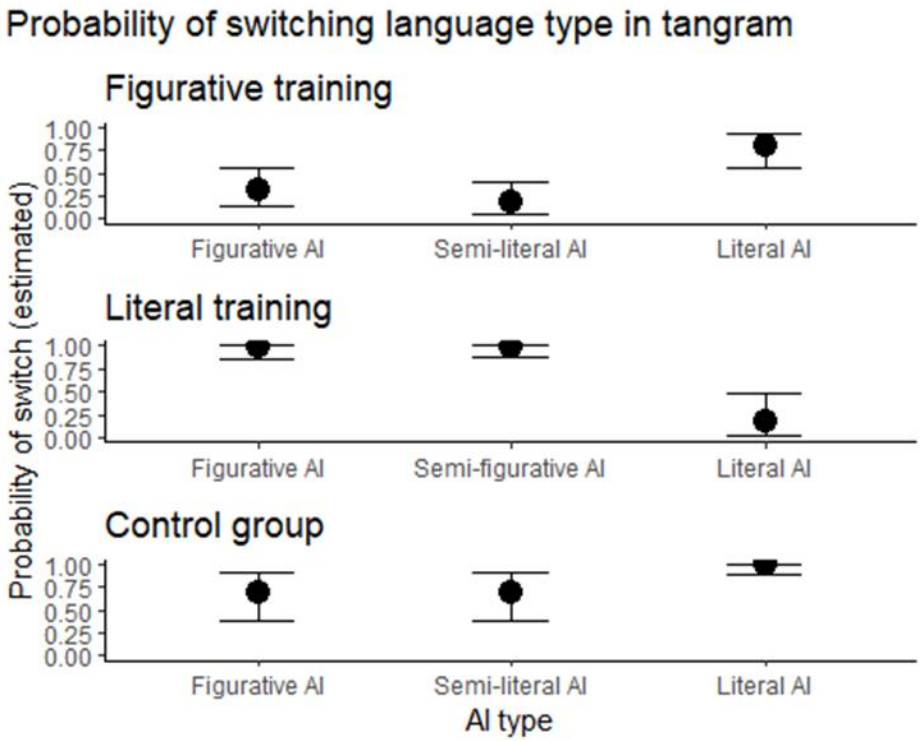
Within the Literal training group, interaction with both the Figurative and Semi-Figurative preference AIs resulted in identical means for language type switch (M = 1, SD = 0), and interaction with the Literal preference AI resulted in a noticeably small mean of (M = 0.2, SD = 0.42). The main switch direction that resulted from interacting with both the Figurative and Semi-Figurative preference AIs was from literal to figurative with all 10 participants that interacted with the Figurative preference AI adopting this language type and 9 of 10 participants that interacted with the Semi-Figurative preference AI (Table10). The BRM presented in Table 19 indicates a credible estimated increase in the log-odds of switching language type during the Tangram task after interacting with the Figurative AI preference (Estimate = 7.19, 95%CI [1.82, 18.35]). While interaction with the Semi-Figurative preference AI resulted in an increase in the log-odds of switching language type (Estimate = 1.83, 95%CI [-8.82, 14.82]) but that this increase was not credible due the inclusion of 0 in the 95% credible interval, indicating that interacting with the Semi-Figurative preference AI resulted in similar log-odds of language type switch to the Figurative preference AI.

Table 19: Likelihood of switching language type during Tangram task - native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	7.19	4.32	[1.82, 18.35]
Semi-figurative AI	1.83	5.82	[-8.82, 14.82]
Literal AI	-8.82	4.36	[-19.84, -3.08]

Note: Estimates are log-odds. Brms model formula is `Switch_Binary ~ AI_Preference, family = bernoulli()`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

Figure 20. Estimated probability of switching language types in Tangram - native speaker



11.4 Post-test results

For the post-test there are two dependent variables: word count and language type use. Word count is first presented in terms of descriptive statistics for each training group analysed by AI language type preference. The results of Bayesian Regression Modeling provide an estimate of the word count used by each training group analysed by AI language type preference at this stage of the study. Language type use is subsequently presented through the analysis of the proportion of language type use, with Bayesian Regression Modeling results indicating the likelihood of a language type being used.

11.4.1 Post-test word count results

Tables 20, 21, and 22 present the descriptive statistics for mean word count, analysed by AI language type preference within each training group (Control group, Figurative training, and Literal training respectively).

Table 20: Post-Test Word Count Descriptives - native speaker Control Training

preference AI	Mean	SD
Figurative AI	4.90	4.75
Literal AI	22.96	13.1
Semi AI	19.12	19.54

Note: This table displays the mean and standard deviation of word counts in the post-test the Control group.

Table 21: Post-Test Word Count Descriptives - native speaker Figurative Training

preference AI	Mean	SD
Figurative AI	4.22	1.25
Literal AI	12.04	9.2
Semi AI	3.82	5.61

Note: This table displays the mean and standard deviation of word counts in the post-test the Figurative training group.

Table 22: Post-Test Word Count Descriptives - native speaker Literal Training

preference AI	Mean	SD
Figurative AI	6.42	5.47
Literal AI	35.12	18.58
Semi AI	10.52	12.21

Note: This table displays the mean and standard deviation of word counts in the post-test the Literal training group.

For the dependent variable word count within the Control group, interaction with the Literal preference AI resulted in the highest mean word count (M = 22.96, SD = 13.1). The BRM presented in Table 23 indicates that interaction with the Literal preference AI is associated with a credible increase in estimated word count (Estimate = 17.57, 95% CI [5.48, 29.54]) compared to the Figurative AI baseline. Similarly, the model indicates that interaction with the Semi-Literal preference AI is associated with a credible increase in estimated word count (Estimate = 14.07, 95% CI [2.99, 25.65]) compared to the Figurative AI baseline.

Table 23: Estimated word count (post-test) - native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	4.53	4.20	[-3.83, 12.92]
Literal AI	17.57	6.08	[5.48, 29.54]
Semi-Literal AI preference	14.07	5.84	[2.99, 25.65]

Note: Brms model formula is `post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, interaction with the Literal preference AI resulted in the highest mean word count (M = 12.04, SD = 9.2), and interaction with the Figurative and Semi-Literal preference AIs resulted in similar mean word counts (M = 4.22, SD = 1.25) and (M = 3.82, SD = 5.61) respectively. The BRM presented in Table 24 indicates that interaction with the Literal preference AI is associated with a credible increase in estimated word count (Estimate = 28.62, 95% CI [17.07, 39.87]) compared to the Figurative AI baseline. However, there was no credible difference observed in the estimated word count for interaction with the Semi-Literal preference AI (Estimate = 4.18, 95% CI [-7.10, 15.44]) compared to the Figurative AI baseline.

Table 24: Estimated word count (post-test) - native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	5.45	4.08	[-2.61, 13.71]
Literal AI	28.62	5.80	[17.07, 39.87]
Semi-Literal AI preference	4.18	5.73	[-7.10, 15.44]

Note: Brms model formula is `post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Literal training group, interaction with the Literal preference AI resulted in the highest mean word count (M = 35.12, SD = 18.58), while interaction with the Figurative preference AI resulted in noticeably lower mean word count (M = 6.42, SD = 5.47). The BRM presented in Table 25 indicates that interaction with the Literal preference AI is associated with a credible increase estimated word count (Estimate = 7.82, 95% CI [3.44, 12.39]) compared to the Figurative AI baseline. While interaction with the Semi-Figurative preference AI did not result in performance that was credibly different compared to the Figurative baseline.

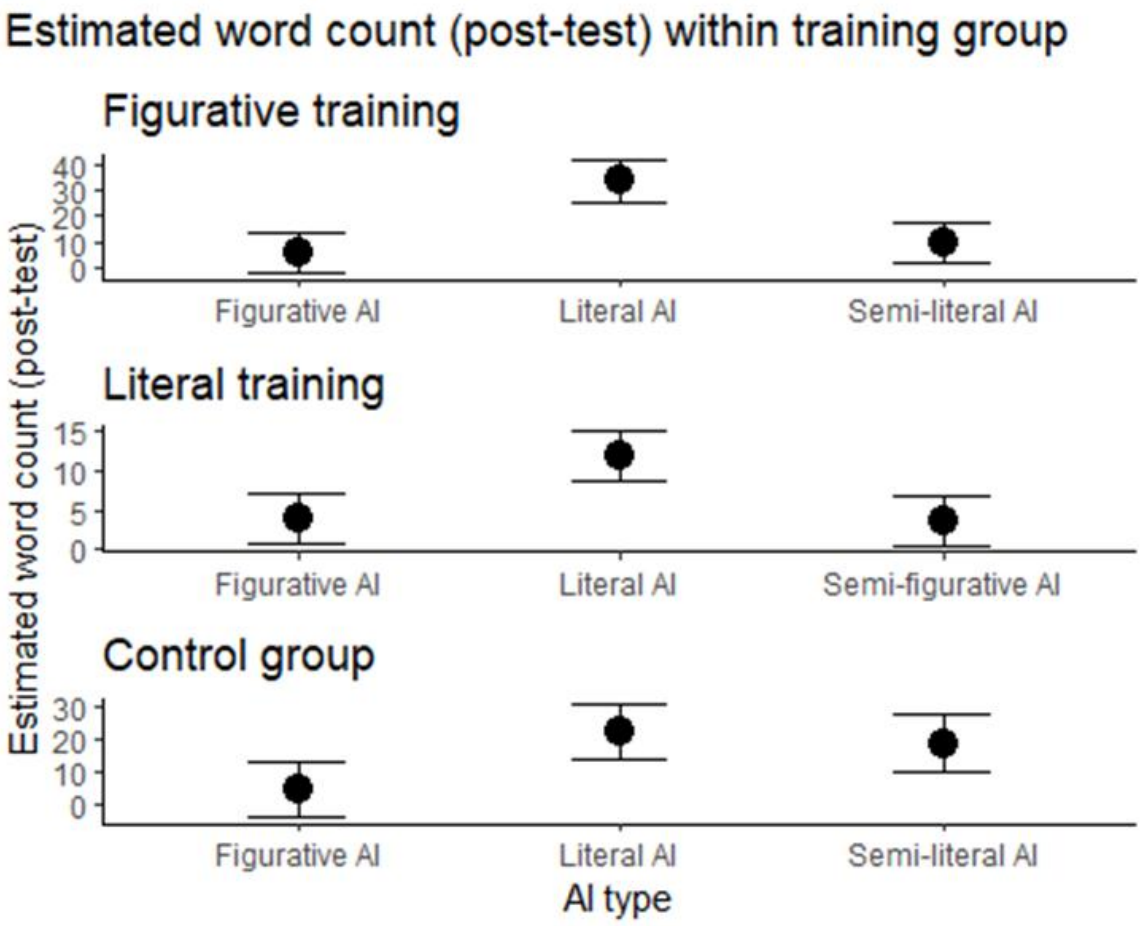
Table 25: Estimated word count (post-test) - native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	3.97	1.60	[0.73, 7.08]
Literal AI	7.82	2.29	[3.44, 12.39]
Semi-Figurative AI preference	-0.40	2.28	[-4.86, 4.22]

Note: Brms model formula is `post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

The following plots estimate how word counts in the post-test (following initial training and interacting with an AI during the Tangram game) differ among participants who interacted with different types of AIs but had the same training (or no training in the case of the Control group). For participants with a given type of language experience (i.e. training condition), what is the effect of interacting with an AI who is either i/ congruent (same language preference as training), ii/ incongruent but cooperative (uses different language from training, but accepts either type from its interlocutor), or iii) incongruent and uncooperative (uses a different language type from training, and only accepts that language type from its interlocutor).

Figure 21. Estimated word count in the post-test - native speaker plots



Note: Plots represent conditional effects and 95% Credible Intervals from Brms model formula post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian(). Plot A runs on the subset of participants who underwent figurative training, Plot B on participants with literal training, and Plot C on the control participants. Semi-Literal or semi-Figurative AI preference represents an AI who communicates using the named language type, but accepts either language type from its interlocutor.

11.4.2 Post-test language type use results

Tables 26, 27, and 28 present the proportion of language type use, analysed by AI language type preference within each training group (Figurative training, Control group, and Literal training respectively).

Table 26: Proportion of language type use (post-test) - native speaker Figurative training

	Figurative language use	Literal language use	Mixed language use
Figurative AI	98%	0%	2%
Literal AI	30%	66%	4%
Semi AI	98%	2%	0%

Note: The table displays the proportion of language type use (post-test)

Table 27: Proportion of language type use (post-test) - native speaker Control group

	Figurative language use	Literal language use	Mixed language use
Figurative AI	88%	0%	12%
Literal AI	4%	76%	20%
Semi AI	52%	30%	18%

Note: The table displays the proportion of language type use (post-test)

Table 28: Proportion of language type use (post-test) - native speaker Literal training

	Figurative language use	Literal language use	Mixed language use
Figurative AI	84%	0%	16%
Literal AI	0%	84%	16%
Semi AI	70%	10%	20%

Note: The table displays the proportion of language type use (post-test)

For the dependent variable language type use within the Figurative training group AI preferences resulted in noticeably different proportions of language type use with the Figurative and Literal preference AIs resulting in high proportion use of their preferred language type (proportion = 98%, figurative) and (proportion = 66%, literal); while interaction with the Semi-Literal preference AI did not influence language type use with participants using

a higher proportion of figurative language (proportion = 98%). The BRM presented in Table 29 indicates that, compared to the baseline interaction with Figurative AI, interacting with Literal preference AI is associated with a credible increase in the estimated log-odds of using literal language post-test (Estimate = 113.55, 95% CI [35.41, 338.81]). Similarly, interaction with the Literal preference AI is also associated with a credible increase in the estimated log-odds of using mixed language post-test (Estimate = 62.05, 95% CI [15.16, 184.12]). The model also indicates that interaction with the Semi-Literal preference AI results in higher log-odds for both literal language and mixed language use (Estimate = 37.60, 95% CI [-5.75, 158.52]) and (Estimate = 4.54, 95% CI [-19.48, 35.83]) respectively but that these increases were not credible as both 95% credible intervals include 0. Therefore, interaction with the Semi-Literal preference AI resulted in similar log-odds of literal language use as the Figurative preference AI baseline.

Table 29: Estimated language type use (post-test) - native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Literal language)	-51.68	43.09	[-172.06, -12.34]
Intercept (Mixed language)	-21.13	19.16	[-70.32, -4.08]
Literal language:Literal AI	113.55	85.11	[35.41, 338.81]
Literal language:Semi-literal AI	37.60	42.96	[-5.75, 158.52]
Mixed language:Literal AI	62.05	49.60	[15.16, 184.12]
Mixed language:Semi-literal AI	4.54	15.19	[-19.48, 35.83]

Note: Brms model formula is `post_test_language_type ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = categorical(link = "logit")`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Interpreting the model

Table 29 contains the output for a multinomial model in which the dependent or response variable is the type of language used in the post-test by the subset of participants who underwent Figurative training. There are three possible outcomes (Figurative language use, Literal language use, and Mixed language use). In our model, Figurative language use was used as the reference level. Therefore, we obtained two sets of coefficients, one for Literal

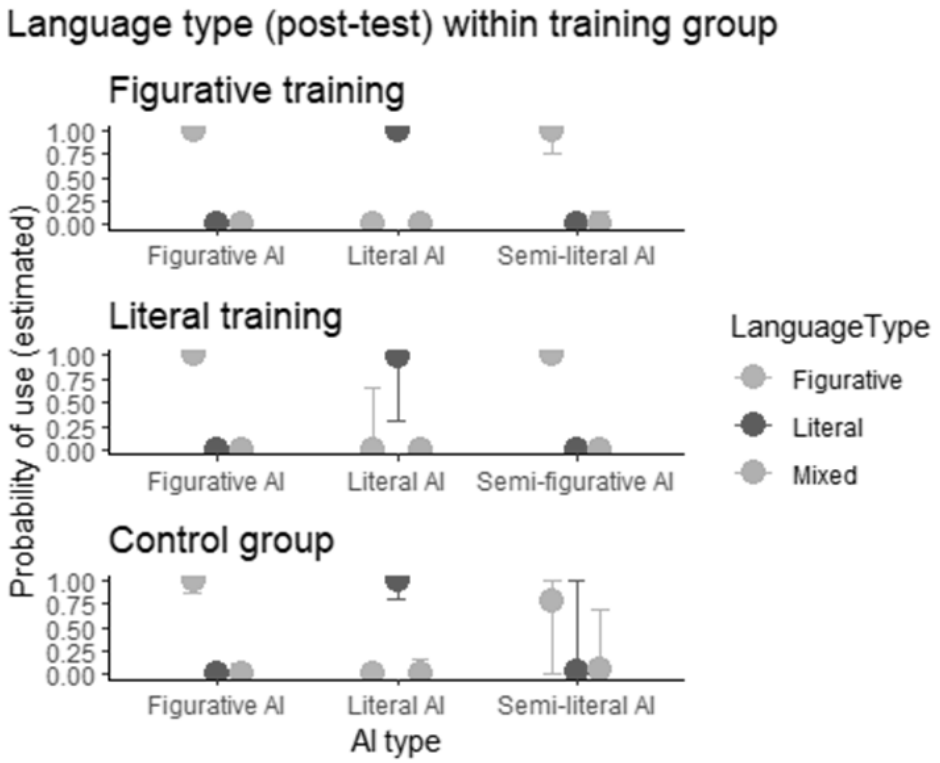
language use compared with Figurative language use, and the other for Mixed Language Use compared with Figurative language use.

The model estimates the influence of interacting with a particular type of AI. Thus the independent variable, AI type, also has three levels: Figurative AI, Literal AI and Semi-literal AI. In our model, AI type is treatment-coded with Figurative AI as the baseline. The model baseline intercept (which does not appear in the table) therefore represents the log-odds of a participant using Figurative language after having interacted with a Figurative AI. The intercept(Lit Language) term represents an odds-ratio: the difference between this baseline estimate and the log-odds of a participant using Literal language (after having interacted with a Figurative AI since this is the baseline of the independent variable). Similarly the intercept(mixed language) term represents the difference between the baseline estimate and the log-odds of a participant using Mixed language (after having interacted with a Figurative AI since this is the baseline of the independent variable).

The next term Literal Language:Literal AI represents a difference of differences, namely the difference between the estimates for Literal language use relative to Figurative language use following interaction with a Literal AI compared to the same difference following interaction with a Figurative AI (the baseline comparison). The next term Literal language:Semi-literal AI similarly represents the difference between the estimates for Literal language use relative to Figurative language use following interacting with a Semi-Literal AI compared to the same difference following the baseline interaction with a Figurative AI. The final two terms represent the differences between the estimates for Mixed language use relative to Figurative language following respective interactions with the Literal and Semi-literal AIs, compared to the baseline difference between those terms following interaction with a Figurative AI.

Given the difficulty in interpreting the coefficients in a categorical model directly, a common approach is to use the inverse logit function to interpret the outcomes on the probability scale. Estimated probabilities for each type of language use following interaction with each of the AIs can be seen in Figure 22.

Figure 22. Estimated language type use in post-test - native speaker plots



Within the Control group, interaction with the Figurative preference AI resulted in the highest proportion of figurative language use (proportion = 88%), while interaction with the Literal preference AI resulted in the highest proportion of literal language use (proportion = 76%). The BRM presented in Table 30 indicates that, compared to the baseline interaction with Figurative AI, interacting with Literal AI is associated with a credibly higher log-odds of using literal language post-test (Estimate = 47.23, 95%CI [16.84, 127.10]). The interaction with Semi-Literal AI is associated with a credibly higher log-odds of using literal language compared to the baseline interaction with Figurative AI (Estimate = 25.11, 95% CI [2.24, 81.37]).

Table 30: Estimated language type use (post-test) - native speaker Control group

	estimate	std.error	95%CI
Intercept (Literal language)	-28.95	19.40	[-84.21, -8.82]
Intercept (Mixed language)	-6.13	2.72	[-12.86, -1.96]
Literal language:Literal AI	47.23	29.18	[16.84, 127.10]
Literal language:Semi-literal AI	25.11	20.55	[2.24, 81.37]
Mixed language:Literal AI	9.93	4.60	[3.14, 20.88]
Mixed language:Semi-literal AI	3.41	3.29	[-2.05, 10.95]

Note: Brms model formula is `post_test_language_type ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = categorical (link = "logit")`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

The output for Table 30 (Estimated language type use (post-test) - Control group) can be interpreted in the same fashion as the previous output which estimated the same terms but on the post-test data from participants who had undergone Figurative training.

Within the Literal training group, AI preferences resulted in noticeably different proportions of language type use, with the Figurative and Literal preference AIs resulting in a high proportion use of their preferred language type (proportion = 84%, figurative) and (proportion = 84%, literal); while interaction with the Semi-Figurative AI resulted in a noticeably high proportion of figurative language use (proportion = 70%).The BRM presented in Table 31 indicates that, within the Literal training group, compared to the baseline interaction with Figurative AI, interacting with Literal AI is associated with a credibly higher log-odds of using literal language post-test (Estimate = 32.37, 95% CI [9.57, 99.84]). While interaction with the Semi-Figurative preference AI did not result in statistically different log-odds of literal or mixed language use compared to the baseline interaction with the Figurative preference AI. Therefore, interaction with the Semi-Literal preference AI resulted in similar log-odds of literal language use as the Figurative preference AI baseline.

Table 31: Estimated language type use (post-test) - native speaker Literal training

	estimate	std.error	95%CI
Intercept (Literal language)	-26.01	21.97	[-86.59, -6.90]
Intercept (Mixed language)	-9.24	4.64	[-21.20, -3.62]
Literal language:Literal AI	32.37	24.47	[9.57, 99.84]
Literal language:Semi-figurative AI	12.62	20.49	[-8.00, 67.04]
Mixed language:Literal AI	4.07	4.15	[-3.12, 13.54]
Mixed language:Semi-figurative AI	-12.97	19.47	[-63.63, 3.65]

Note: Brms model formula is `post_test_language_type ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = categorical(link = "logit")`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language, but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

The output for Table 31 (Estimated language type use (post-test) - Literal training) can be interpreted in the same fashion as the previous output which estimated the same terms but on the post-test data from participants who had undergone Figurative training.

11.5 Summary of Study 2: Native speaker results

This section presents a summary of the results of study 2 for native speakers of English (n=90). This section covers the following research question and hypotheses.

- Research question 1: Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?
- Hypothesis 1: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that in the pre-test that without training participants should favour the use of figurative language-based descriptions.
- Hypothesis 2: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that untrained participants should show a bias towards figurative descriptions and perform similarly to the figurative description pre-training group throughout the experiment.

- Hypothesis 3: Without an increase in communicative effort due to breakdowns, participants will maintain their trained or untrained language type used for describing images (in the case of control group participants) unless participants come across less effortful means of achieving the goals of the communicative task.
    - For example, figurative and control group participants do not switch to literal descriptions in the semi condition since there are no breakdowns by design, while the literal group will switch to figurative descriptions in the semi condition even when there are no breakdowns due to an economy of effort.
- Hypothesis 4: Participants will switch to the description type of the AI that causes breakdowns when the AI refuses to understand the participants' descriptions.
    - For example, figurative and control participants switching to literal descriptions and literal participants switching to figurative.

Key results are summarised in Table 33 for ease of reference. These include the hypothesis tests for all dependent measures in the Tangram task, and descriptives for language type use in the post-test. Credible differences between conditions are indicated by use of bold.

Table 33: Summary of dependent measure in Tangram and Post-test - native speakers

	Figurative training	Literal Training	Control group
Tangram: number of turns	**Literal AI > Fig AI** = Semi AI	Lit AI = Fig AI = Semi AI	Lit AI = Fig AI = Semi AI
Tangram: breakdowns	**Literal AI > Fig AI** = Semi AI	**Lit AI = Semi AI < Fig AI**	**Lit AI > Fig AI** = Semi AI
Tangram: Switches	**Literal AI > Fig AI** = Semi AI	**Lit AI < Fig AI** = Semi AI	**Literal AI > Fig AI** = Semi AI
Post-test: word count	**Literal AI > Figurative AI** = Semi AI	**Lit AI < Fig AI** = Semi AI	**Literal AI > Semi AI > Figurative AI**
Post-test: language type	**Literal AI = 66% Literal language use Figurative AI = 98% Figurative language use Semi AI = 98% Figurative language use**	**Literal AI = 84% Literal language use Figurative AI = 84% Figurative language use Semi AI = 70% Figurative language use**	**Literal AI = 76% Literal language use Figurative AI = 88% Figurative language use Semi AI = 52% Figurative language use**

11.5.1 Hypothesis 1

In terms of the first hypothesis, the pre-test results show that native speaker participants used a mean word count of 13.58, with a standard deviation of 12.82. Figurative language was used in 233 descriptions (52%) and Mixed in 197 descriptions (44%). The BRM results confirm that without training participants are less likely to use literal language compared to figurative language, but that figurative and mixed language are equally likely to be used. This result confirms hypothesis 1 as purely figurative language was the most used language type in the pre-test.

11.5.2 Hypothesis 2

In terms of hypothesis 2, The overall similarity in performance between both groups, discussed below, generally confirms the hypothesis, with the possibility that differences in performance observed between Control and Figurative training group participants are possibly attributed to the lack of training for Control group participants leaving them more susceptible to the influence of AI preference.

The dependent variables for the Tangram task stage were total turns taken, total breakdowns, and language type switch. The BRM results for total turns taken indicated that the estimates for number of turns taken for baseline interaction with the Figurative preference AI were very similar; and that interaction with the Semi-Literal preference AI also resulted in similar estimates that were not credible meaning that interaction with this AI preference did not result in different estimates for total turns taken compared to the baseline interaction. Interaction with the Literal preference AI resulted again in very similar estimates for total turns taken for both groups. However, the 95% credible interval for the Control group included zero, indicating weak evidence or a lack of credible difference. This contrasts with the Figurative training group, where the credible interval excluded zero, suggesting a small difference in the influence of the Literal preference AI on total turns taken between the two groups. The model results for total breakdowns reveal credile increases in the estimate for total breakdowns when interacting with the Literal AI preference across both groups. Interaction with the baseline Figurative AI preference resulted in a credible decrease in the estimated total number of breakdowns for the Figurative training group. This credible effect was not observed in the Control group, as indicated by the inclusion of 0 within the 95% credible interval for the baseline estimate. The small positive magnitude of the estimate for baseline interaction with the Figurative AI preference in the Control group, alongside its 95% credible interval that narrowly includes zero, may be attributed to the group's small sample size (n=10). Additionally, the lack of explicit training or preparation for interacting with the AI could have influenced participants' language choices, leading to slightly more breakdowns in this group and thereby affecting the baseline estimate.In terms of language type switch the BRM results for both groups reveal very similar estimates for the log-odds of language type switch for both groups. Where interaction with the baseline Figurative preference AI resulted in a 95% credible interval that includes 0 for both groups, indicating that there was no clear tendency to switch language type when interacting with the Figurative preference AI.

Interaction with the Semi-Literal preference AI resulted in a 95% credible interval that included zero, indicating weak evidence or a lack of credible difference compared to baseline

interaction with the Figurative preference AI. This suggests no clear tendency to switch language type. Together with these results it can be inferred that participants in both groups used and maintained the use of figurative language throughout interaction with both AI preferences, because the Figurative preference AI rejects the use of literal language and would have indicated a language type switch if participants had used literal language. Furthermore, this is consistent with the BRM results for total breakdowns for interaction with the Figurative preference AI which indicated a credile decrease in the estimated number of breakdowns for Figurative training group participants i.e. there was a decrease in breakdowns because figurative language was used.

Post-test results also indicate similar performance between Control group and Figurative training participants in terms of word count, where the baseline interaction with the Figurative preference AI resulted in very similar estimates for word count. However, interaction with the Semi-literal preference AI resulted in a credible increase in the estimate in word count for Control group participants that was not similarly revealed for Figurative training participants as the 95% credible interval included zero indicating weak evidence or a lack of credible difference from the baseline interaction with the Figurative preference AI. For both groups interaction with the Literal preference AI resulted in a credible increase in the estimate for word count indicating that this AI preference influenced both groups similarly.

Model results for language type use in the post-test reveal a similar pattern to word count. Where baseline interaction with the Figurative preference AI resulted in a credible decrease in the log-odds of literal language use for both groups, interaction with the Literal preference AI however increased the log-odds of literal language use. In terms of the Semi-Literal preference AI, the Control group was more susceptible to AI influence without breakdowns and showed a credible increase in the log-odds of literal language use after interaction. Unlike the Figurative training group, where interaction with the Semi-Literal preference AI did not result in a credible difference in the use of literal language compared to the baseline, thereby indicating similar performance to interaction with the Figurative preference AI. Overall, these results reflect several areas of similarity between both groups, indicating that without training Control group participants perform similarly to Figurative group participants, with the main difference being that the lack of training likely increases the influence of AI preferences on untrained participants.

11.5.3 Hypothesis 3

In terms of hypothesis 3, the overall results discussed below confirm the hypothesis and indicate that an economy of effort influences communication. Furthermore, these results demonstrate that merely modelling different language behaviour isn't sufficient for a speaker to make use of that language if it is a higher effort strategy (e.g. literal language).

The dependent variables for the Tangram task stage were total turns taken, total breakdowns, and language type switch. The BRM results for interaction with the Semi-preference AI for both Figurative and Literal training participants indicated that interaction did not result in a credible difference in the estimated number of turns taken when compared to the baseline interaction with the Figurative preference AI. In terms of breakdowns the Semi-preference AI by design avoids breakdowns by accepting the participants trained language use in descriptions but uses the opposite description type, where the AI responds with literal descriptions to Figurative training participants that use figurative descriptions and vice versa. Therefore, the dependent variable language type switch reflects language type switch behaviour that is not influenced by breakdowns but rather a tendency towards an economy of effort. For both groups interaction with the Semi-preference AI did not result in credible differences in the log-odds of language type switch when compared to the baseline interaction with the Figurative preference AI. For Figurative training participants this means that the Semi-literal preference AI did not cause participants to switch to literal language use since the baseline interaction indicated that the Figurative preference AI did not show a tendency towards language type switch. However, the inverse is true for Literal training participants, where the baseline interaction resulted in a credible increase in the log-odds of observing a language type switch for this group. Meaning that when interaction with the Semi-Figurative preference AI for Literal training participants caused participants to switch language type and use figurative language since the Semi-Figurative preference AI did not result in performance that was credibly different compared to the baseline interaction i.e. the Semi-Figurative preference AI resulted in log-odds of language type switch similar to the baseline.

Post-test results also confirm these similarities where the Semi-preference AI for both groups did not result in an estimate for word count that was credibly different from the baseline estimate for interaction with the Figurative AI. This pattern was also observed for language type use where interaction with the Semi-preference AI resulted in both groups having a credible decrease in the log-odds of literal language use, since the Semi-preference AI did not result in credibly difference log-odds of literal language type use compared to the baseline interaction

with the Figurative preference AI. Where the Figurative preference AI showed a credible decrease in the log-odds for literal language use.

Overall, these results confirm hypothesis 3 and show that without added in the form of communicative breakdowns participants are unlikely to switch to higher effort (word count) communicative behaviour; while these results also confirm that the inverse is true where participants that were trained on the use of higher effort literal descriptions opportunistically adopted the use of lower effort figurative descriptions without added effort in the form of communicative breakdowns. These results therefore indicate a tendency towards an economy of effort in communication since it was not an issue with the ability of either strategy to satisfice the communicative needs of the task as breakdowns were eliminated by design. Meaning that participants likely only switched when the opportunity to reduce total communicative effort arose when the AI provided a less effortful model of language that was likely to satisfice the communicative needs of the interaction as indicated by the AI's use of figurative language for Literal training participants.

11.5.4 Hypothesis 4

In terms of hypothesis 4, the discussion of results below indicates that breakdowns from interacting with AI language type preferences that conflicted with trained language type use resulted in increased effort in communication leading to language type switching. This was observed even for the Figurative training group with switching to literal language descriptions that involve more effort per description but less total effort overall by avoiding the need for repair, especially when considered in conjunction with the discussion of hypothesis 3 above.

For the Tangram task the BRM results for total turns taken indicated that interaction with the different preference AIs for the Literal training group did not result in credible differences in the estimated number of turns taken. While interaction with the Literal preference AI did result in a credible increase in the estimated number of turns taken compared to the baseline interaction with the Figurative preference AI for Figurative training participants. In terms of breakdowns, for the Literal training group the baseline interaction with the Figurative preference AI resulted in the highest estimate for breakdowns which was a credible increase compared to interaction with the Literal preference AI. Conversely, interaction with the Literal preference AI resulted in a credible increase in the estimated number of breakdowns for Figurative training group participants when compared to the estimate for baseline interaction with the Figurative preference AI. In terms of language type switch a similar pattern was observed where interaction with the Literal preference AI resulted in a credible increase in the log-odds of

language type switch for Figurative training participants, and interaction with the Figurative preference AI resulted in a credible increase in the log-odds of language type switch for Literal training participants.

Taken in conjunction with the discussion of hypothesis 3, the pattern of increased language type switch after increased breakdowns when interacting with the Literal preference AI for Figurative training participants, indicates that communicative breakdowns resulted in an increased effort in repair to the point that adopting literal language use was likely in an attempt to reduce overall effort in spite of the increased effort per description in terms of word count. However, Figurative training participants also showed more resistance to switching to literal language use than did Literal training participants to switching to figurative language use. This was indicated by the higher mean number of breakdowns for Figurative training participants indicating participants attempted the use of figurative language descriptions more than their Literal training counterparts did with their attempts to use literal language descriptions. Furthermore, Literal training participants showed a much higher estimate for language type switch to figurative language based on the BRM results for language type switch when compared to Figurative training participants. Indicating less resistance to switching to a less effortful figurative language descriptions.

Post-test results also revealed a similar pattern where interaction with the Literal preference AI resulted in a credible increase in the estimate for word count and a credible increase in the log-odds of literal language use for Figurative training participants; while interaction with the Figurative preference AI resulted in a credible decrease in estimated word count and a credible decrease in the log-odds of both literal and mixed language use thereby indicating the use of figurative language.

Overall, the comparison of groups indicates that participants that interacted with an AI preference that matched their training were observed to have less breakdowns and were less likely to switch language type use; however, when they interacted with an AI preference that did not match their training they were observed to have a credible increase in breakdowns and a credible increase in the likelyhood of switching language type use. These observations are in line with the expected influence of a tendency towards an economy of effort and indicates that the experimental paradigm is able to operationalise, manipulate, and measure the influence of an economy of effort in communication.

11.6 Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?

Based on the overall results for native speakers, it appears that the experimental paradigm used in this study shows good potential for use in experimentally operationalising and manipulating an economy of effort in communication as indicated by the discussion of hypothesis 3 and 4 above. The experimental paradigm showed switches in language type use without instances of breakdowns for literal training participants that interacted with Semi-Figurative preference AI indicating and essentially demonstrating the influence of an economy of effort. In contrast, participants that interacted with AI preferences that did not match their language type training experienced breakdowns and subsequently switched language type use, with figurative training participants showing signs of more resistance to switching to literal language.

Overall, the results of native speakers thus far indicate that the experimental paradigm and the dependent variables used to operationalise and measure the influence of an economy of effort are effective for the purpose of testing the influence of an economy of effort in communication, and potentially be subsequently applied to test the influence of this notion on usage based naturalistic adult second language acquisition. However, these results also clearly indicate that both the experimental paradigm and dependent variables require further replication with larger sample sizes for more robust interpretations to be made with regards to the data collected.

12. Non-native speaker results

This section presents the results of the study for non-native speakers of English (n=90) in order of each stage of the study (pre-test, training, Tangram task, post-test) covering the descriptives and Bayesian Regression Modeling results for each dependent variable.

12.1 Pre-test results

For the pre-test stage, there are two dependent variables: word count and language type use. Word count is first presented in terms of descriptive statistics for all participants overall, accompanied by the results of Bayesian Regression Modeling, which provides an estimate of the word count used in the pre-test. Language type use is subsequently presented through the analysis of the proportion of language type use across all 450 descriptions provided by participants, complemented by Bayesian Regression Modeling results that indicate the likelihood of a language type being used.

12.1.1 Pre-test word count results

For all participants (n = 90), the mean number of words used to describe an image in the Pre-Test was 12.28, with a standard deviation of 13.94. The distribution of words per image can be seen in the histogram in Figure 23.

Figure 23 Words per image frequency histogram - non-native speaker

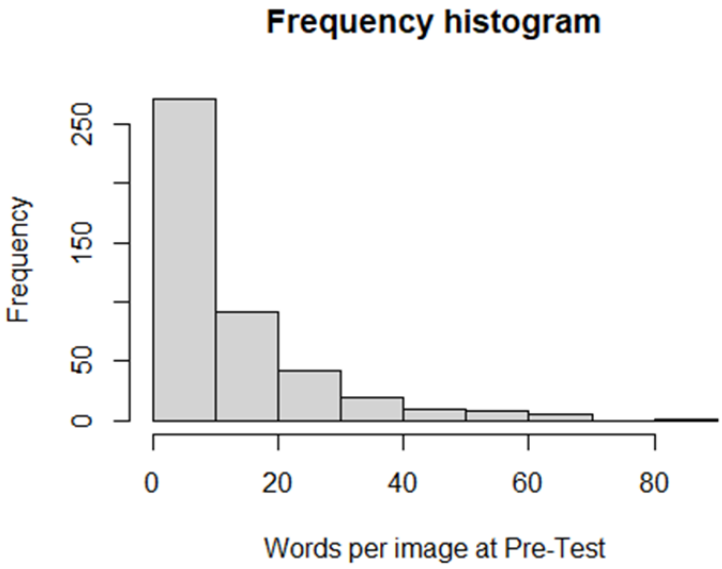


Table 33: Estimated word count (Pre-Test) - all non-native speaker participants

		estimate	std.error	95%CI
	Intercept	11.96	1.42	[9.17, 14.75]
Image	sd	1.77	0.49	[0.94, 2.87]
Participant	sd	12.76	1.01	[11.04, 14.91]
Residual	sd	5.53	0.21	[5.14, 5.96]

Note: Brms model formula is `Word_count ~ (1|ParticipantID) + (1|Image), family = gaussian`. 95%CI are Credible Intervals.

We estimated the mean number of words used to describe an image in the Pre-Test, taking only the random effects of participant and image into account. As can be seen in the model output in Table 33, random effect estimates of the standard deviations for images and participants confirm that there was relatively little variability in word count attributable to different images, but participants varied more widely.

12.1.2 Pre-test language type use results

Each description in the Pre-Test was labelled as using either Figurative, Literal, or Mixed language types. All participants described 5 images, resulting in a total of 450 descriptions. Of these, 237 (53%) were Figurative, 15 (3%) were Literal, and 198 (44%) were Mixed. Table 34

showcases the estimated likelihoods of using each language type (Figurative, Literal, or Mixed) for image descriptions by all participants. As noted below, the model confirms that without any training or instructions, people are less likely to use literal language than figurative language, but that they are equally likely to use mixed language and figurative language.

Table 34: Estimated language types (Pre-Test) - all non-native speaker participants

		estimate	std.error	95%CI
	Intercept (Literal language)	-6.47	1.79	[-10.92, -3.82]
	Intercept (Mixed language)	-0.32	0.36	[-1.03, 0.37]
Image	sd (Literal Language)	1.38	0.86	[0.09, 3.28]
Participant	sd (Literal Language)	3.74	1.21	[1.92, 6.66]
Image	sd (Mixed Language)	0.51	0.26	[0.05, 1.07]
Participant	sd (Mixed Language)	2.75	0.42	[2.04, 3.69]

Note: Brms model formula is `Language_Type ~ (1|ParticipantID) + (1|Image), family = categorical(link = "logit")`. 95%CI are Credible Intervals.

Table 34 contains the output for a multinomial model which estimated the likelihood of using either Figurative, Literal, or Mixed language to describe an image in the Pre-Test, taking only the random effects of participant and image into account. There were three possible outcomes (Figurative language use, Literal language use, and Mixed language use) for each description. In our model, Figurative language use was used as the reference level. Therefore, we obtained

two sets of coefficients, one for Literal language use compared with Figurative language use Intercept (Literal Language), and the other for Mixed Language Use compared with Figurative language use Intercept(Mixed Language). The model confirms that without any training or instructions, people are less likely to use literal language than figurative language, but that they are equally likely to use mixed language and figurative language. Random effect estimates of the standard deviations for individual images and participants confirm that there was relatively little variability attributable to different images, but slightly more to participants.

12.2 Training results

For the training stage, descriptive statistics of accuracy are presented in Table 35. The mean accuracy for all participants was high (M = 0.95, SD = 0.22), indicating that, on average, participants performed accurately during the training stage. Notably, both the figurative and literal training groups achieved near-perfect mean accuracy with a mean score (M = 0.99, SD = 0.07) and (M = 0.98, SD = 0.13) respectively in the second half of the training trials. This suggests that participants effectively learned the correct description type for their respective training group, albeit not to the extent of achieving flawless performance.

Table 35: Training accuracy descriptives for non-native speakers - by group

Category	Mean Accuracy	SD
All groups	0.95	0.22
Figurative Group Block A Accuracy	0.91	0.29
Figurative Group Block B Accuracy	0.99	0.07
Literal Group Block A Accuracy	0.91	0.29
Literal Group Block B Accuracy	0.98	0.13

Note: This table displays the mean and standard deviation accuracy by group for the training assessment.

12.3 Tangram task results

For the Tangram stage task there are three dependent variables which are total turns taken, total breakdowns, and Language type switch; they are also presented in that order. For all dependent variables in this stage descriptive statistics are presented and accompanied by

Bayesian Regression Modeling that provides an estimate of the number of turns and breakdowns on a log scale and the likelihood of a switch being observed as log-odds. The descriptive statistics for these variables are presented in Table 36 below.

Table 36: Tangram dependent variables descriptives for non-native speakers - by group

Group	Mean Switch	SD Switch	Mean Turns Taken	SD Turns Taken	Mean Breakdowns	SD Breakdowns
Control - Figurative Preference	0.8	0.42	27.7	6.80	0.3	0.48
Control - Literal Preference	1.0	0.00	29.6	2.50	3.8	2.15
Control - Semi Preference	0.4	0.52	25.5	1.08	0.0	0.00
Figurative - Figurative Preference	0.1	0.32	25.3	1.95	0.1	0.32
Figurative - Literal Preference	1.0	0.00	30.9	2.60	5.7	3.65
Figurative - Semi Preference	0.4	0.52	26.0	6.99	0.0	0.00

Literal - Figurative Preference	1.0	0.00	27.0	3.33	2.7	1.83
Literal - Literal Preference	0.5	0.53	27.3	2.58	0.7	1.57
Literal - Semi Preference	0.8	0.42	25.1	2.77	0.4	0.52

Note: This table displays the mean and standard deviation for the dependent variables Total turns taken, Total breakdowns, and Binary switch.

Tables 37 and 38 below present the supplementary summary data for the variables Language type switch and total breakdowns.

Table 37: Number of participants that switched and experienced breakdowns for non-native speakers - by group

Group Indicator	Number of participants that switched language type	Number of participants that experienced a breakdown
Control - Figurative Preference	8	3
Control - Literal Preference	10	9
Control - Semi Preference	4	0
Figurative - Figurative Preference	1	1
Figurative - Literal Preference	10	10

Figurative - Semi Preference	4	0
Literal - Figurative Preference	10	8
Literal - Literal Preference	5	3
Literal - Semi Preference	8	4

Note: This table displays the number of participants that switched language type and the number of participants that experienced a breakdown by group

Table 38: Direction of language type switches by group for non-native speakers

Group	Figurative > Literal	Figurative > Mixed	Literal > Figurative	Mixed > Figurative	Mixed > Literal	Literal > Mixed
Control - Figurative Preference	4	1	2	1	0	0
Control - Literal Preference	8	0	0	0	2	0
Control - Semi Preference	0	1	1	2	0	0
Figurative - Figurative Preference	0	0	1	0	0	0
Figurative - Literal Preference	9	1	0	0	0	0

Figurative - Semi Preference	1	2	0	0	0	0
Literal - Figurative Preference	0	0	10	0	0	0
Literal - Literal Preference	3	0	0	0	1	1
Literal - Semi Preference	0	0	8	0	0	0

Note: This table displays the number of participants that switched language type in specific directions by group.

12.3.1 Tangram task total turns taken

For mean total turns taken the descriptive statistics show that within the Control group, participants interacting with the Literal preference AI demonstrated the highest mean number of turns taken to complete the Tangram task (M = 29.6, SD = 2.50). However, the BRM results for the Control group presented in Table 39 do not show a credible difference in the number of turns taken for the Literal preference AI compared to the Figurative preference AI (Estimate = 0.07, 95% CI = [-0.10, 0.23]).

Table 39: Estimated number of turns in tangram - non-native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	3.32	0.06	[3.20, 3.44]
Semi-Literal AI preference	-0.08	0.09	[-0.24, 0.08]
Literal AI	0.07	0.08	[-0.10, 0.23]

Note: Estimates are on the log scale. Brms model formula is `Final_Participant_Turn ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, participants engaging with the Literal preference AI recorded the highest mean number of turns taken (M = 30.9, SD = 2.60). The BRM for the Figurative training group presented in Table 40 indicates a credible increase in the number of turns taken when interacting with Literal preference AI (Estimate = 0.20, 95% CI = [0.04, 0.37]), compared to the Figurative preference AI.

Table 40: Estimated number of turns in tangram - non-native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	3.23	0.06	[3.11, 3.35]
Semi-Literal AI preference	0.03	0.09	[-0.14, 0.21]
Literal AI	0.20	0.08	[0.04, 0.37]

Note: Estimates are on the log scale. Brms model formula is `Final_Participant_Turn ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Literal training group, the descriptive statistics show that participants interacting with the Literal preference AI had a similar mean number of turns taken (M = 27.3, SD = 2.58) to the mean number of turns taken when interacting with the Figurative preference AI (M = 27.0, SD = 3.33). The BRM results for the Literal training group presented in Table 41 do not exhibit a credible difference in the number of turns taken for the Literal preference AI as opposed to the Figurative preference AI. The estimates for both Semi-Figurative preference AI (Estimate = -0.07, 95% CI = [-0.24, 0.10]) and Literal preference AI (Estimate = 0.01, 95% CI = [-0.15, 0.18]) are close to zero and their 95% credible intervals include 0. This suggests that,

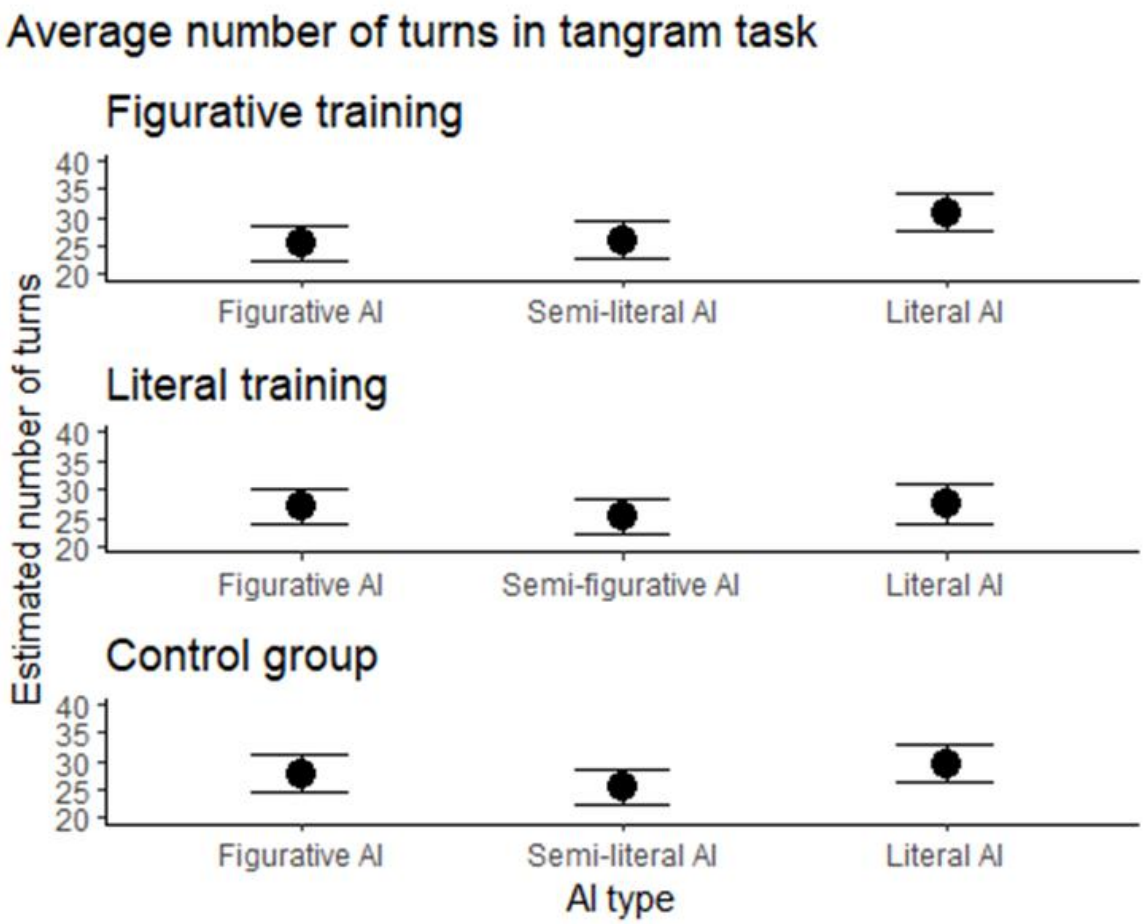
within a Bayesian analysis framework, the observed differences in mean turns taken across AI preferences were not credible.

Table 41: Estimated number of turns in tangram - non-native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	3.29	0.06	[3.18, 3.41]
Semi-Figurative AI preference	-0.07	0.09	[-0.24, 0.10]
Literal AI	0.01	0.09	[-0.15, 0.18]

Note: Estimates are on the log scale. Brms model formula is `Final_Participant_Turn ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

Figure 24 Estimated number of turns taken - non-native speaker



Note: Plots represent conditional effects and 95% Credible Intervals from Brms model formula Final_Participant_Turn ~ AI_Preference_Code, family = poisson(). Plot A runs on the subset of participants who underwent figurative training, Plot B on participants with literal training, and Plot C on the control participants. Semi-Literal or semi-Figurative AI preference represents an AI who communicates using the named language type but accepts either language type from its interlocutor.

12.3.2 Tangram task total breakdowns

For total breakdowns within the Control group, interaction with the Literal preference AI resulted in the highest mean number of breakdowns (M = 3.8, SD = 2.15). The number of participants that experienced breakdowns within the Control group is also noticeably higher for those that interacted with the Literal preference AI than other AI preferences with 9 of 10 participants experiencing breakdowns (Table 37). The BRM presented in Table 42 reveals that interaction with the Literal preference AI credibly increases the likelihood of experiencing breakdowns during the Tangram task (Estimate = 2.67, 95%CI [1.66, 3.91]).

Table 42: Estimated number of breakdowns in tangram - non-native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	-1.37	0.55	[-2.60, -0.41]
Semi-Literal AI preference	-2.29	1.35	[-5.19, 0.08]
Literal AI	2.67	0.57	[1.66, 3.91]

Note: Estimates are on the log scale. Brms model formula is `Total_Breakdowns ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, interaction with the Literal preference AI resulted in the highest mean number of breakdowns (M = 5.7, SD = 3.65). The number of participants that experienced breakdowns within the Figurative training group is also noticeably higher for those that interacted with the Literal preference AI than other AI preferences with all 10 participants experiencing breakdowns (Table 37). The BRM presented in Table 43 reveals that interaction with the Literal preference AI credibly increases the likelihood of experiencing breakdowns during the Tangram task (Estimate = 4.52, 95%CI [2.70, 7.64]).

Table 43: Estimated number of breakdowns in tangram - non-native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	-2.78	1.26	[-5.90, -0.98]
Semi-Literal AI preference	-9.12	11.15	[-32.37, 1.05]
Literal AI	4.52	1.26	[2.70, 7.64]

Note: Estimates are on the log scale. Brms model formula is `Total_Breakdowns ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

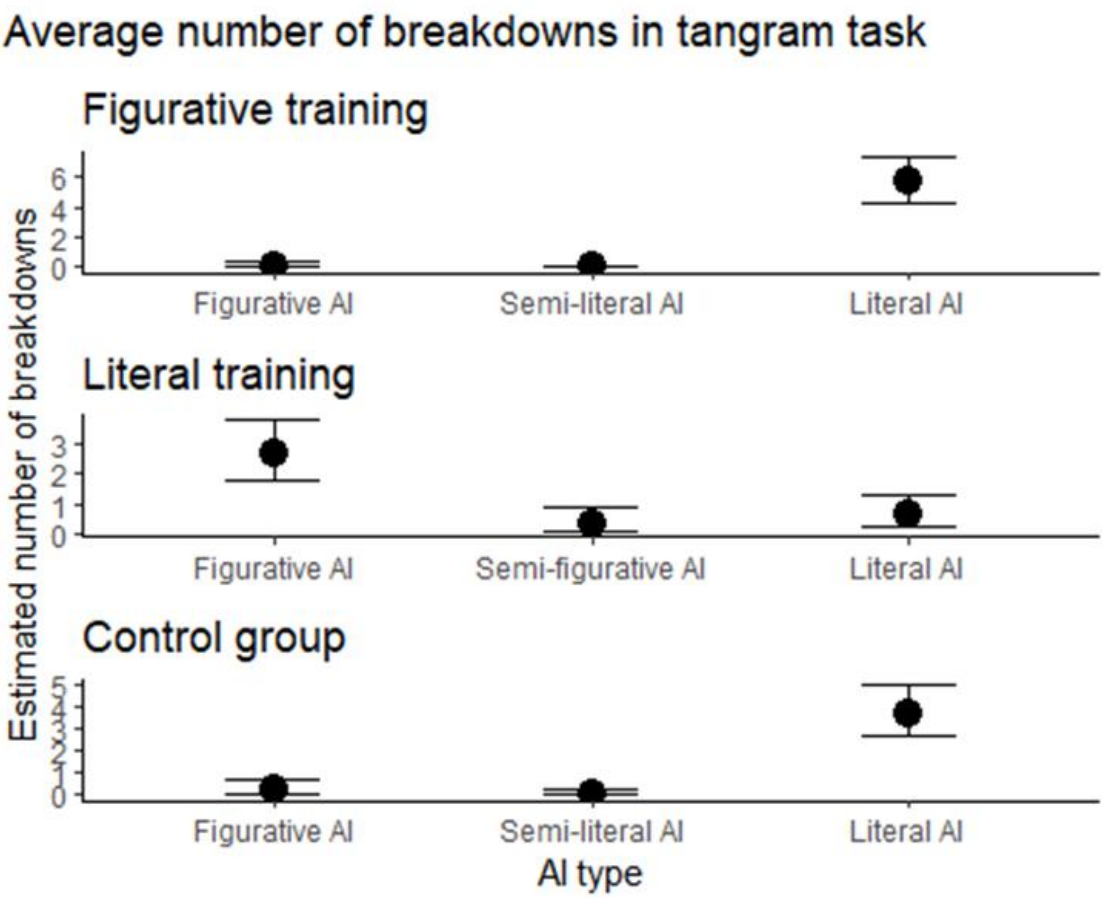
Within the Literal training group, it was found that interaction with the Figurative preference AI resulted in the highest mean number of breakdowns (M = 2.7, SD = 1.83). The number of participants that experienced breakdowns within the Literal training group is also noticeably higher for those that interacted with the Figurative preference AI than other AI preferences with 8 of 10 participants experiencing breakdowns (Table 37). The BRM presented in Table 44 reveals that interaction with the Figurative preference AI credibly increases the likelihood of experiencing breakdowns during the Tangram task (Estimate = 0.97, 95%CI [0.59, 1.33]), compared to the credible negative estimate for interaction with the Literal preference AI (Estimate = -1.43, 95%CI [-2.36, -0.65]).

Table 44: Estimated number of breakdowns in tangram - non-native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	0.97	0.19	[0.59, 1.33]
Semi-Figurative AI preference	-2.04	0.58	[-3.30, -1.02]
Literal AI	-1.43	0.44	[-2.36, -0.65]

Note: Estimates are on the log scale. Brms model formula is `Total_Breakdowns ~ AI_Preference_Code, family = poisson()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

Figure 25 Estimated number of breakdowns in Tangram - non-native speaker



12.3.3 Tangram task language type switch

For language type switch within the Control group interaction with the Literal preference AI resulted in the highest mean language type switch (M = 1, SD = 0), with a language type switch being observed for all 10 participants (Table 37) that interacted with the Literal preference AI. The main switch direction that resulted from interacting with the Literal preference AI was from figurative to literal language type for 8 of 10 participants (Table 38). The BRM presented in Table 45 indicates a credible estimated increase in the log-odds of switching language type during the Tangram task after interacting with the Literal preference AI (Estimate = 8.86, 95%CI [0.19, 32.66]). While interaction with the Semi-Literal preference AI did not result in performance that was credibly different than the baseline interaction with the Figurative preference AI.

Table 45: Likelihood of switching language type during Tangram task - non-native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	1.47	0.85	[-0.03, 3.30]
Semi-literal AI	-1.97	1.13	[-4.31, 0.17]
Literal AI	8.86	8.47	[0.19, 32.66]

Note: Estimates are log-odds. Brms model formula is `Switch_Binary ~ AI_Preference, family = bernoulli()`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, again interaction with the Literal preference AI resulted in the highest mean language type switch (M = 1, SD = 0). Furthermore, interaction with both the Figurative and Semi-Literal preference AIs resulted in noticeably smaller means for language type switch (M = 0.1, SD = 0.32) and (M = 0.4, SD = 0.52) respectively. The main switch direction that resulted from interacting with the Literal preference AI was from figurative to literal language type for 9 of 10 participants (Table 38). The BRM presented in Table 46 indicates a credible estimated increase in the log-odds of switching language type during the Tangram task after interacting with the Literal preference AI (Estimate = 2.14, 95%CI [0.86, 3.58]). While interaction with the Semi-Literal preference AI did not result in performance that was credibly different than the baseline interaction with the Figurative preference AI.

Table 46: Likelihood of switching language type during Tangram task - non-native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	-0.73	0.52	[-1.79, 0.26]
Semi-literal AI	0.19	0.66	[-1.10, 1.52]
Literal AI	2.14	0.71	[0.86, 3.58]

Note: Estimates are log-odds. Brms model formula is `Switch_Binary ~ AI_Preference, family = bernoulli()`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

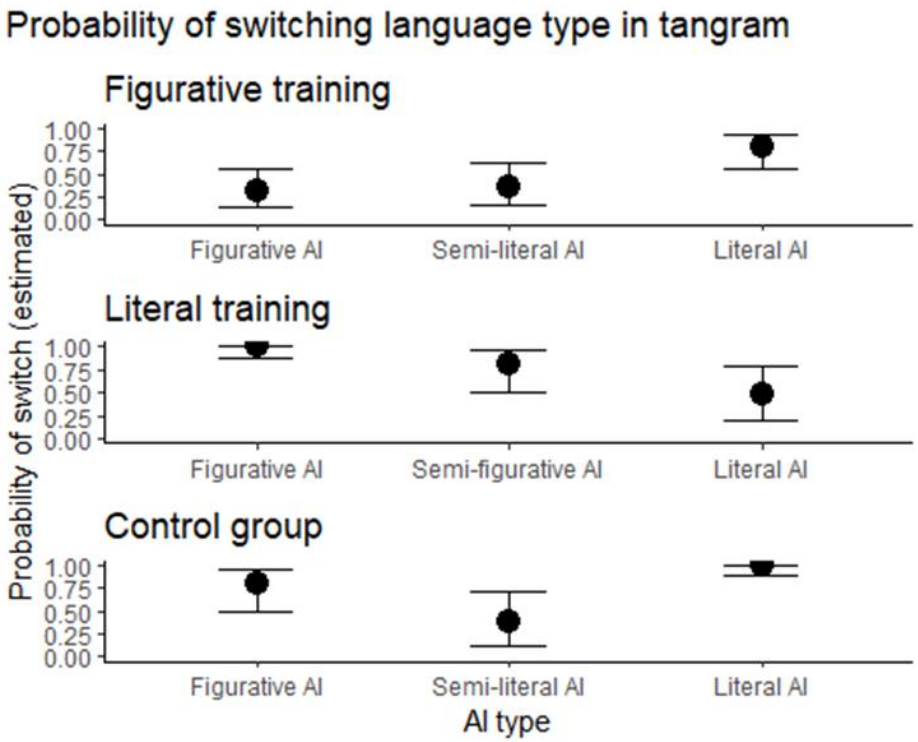
Within the Literal training group, interaction with both the Figurative and Semi-Figurative preference AIs resulted identical means for language type switch (M = 1, SD = 0), and interaction with the Literal preference AI resulted in a noticeably small mean of (M = 0.2, SD = 0.42). The main switch direction that resulted from interacting with both the Figurative and Semi-Figurative preference AIs was from literal to figurative with all 10 participants that interacted with the Figurative preference AI adopting this language type and 8 of 10 participants that interacted with the Semi-Figurative preference AI (Table 38). The BRM presented in Table 47 indicates a credible estimated increase in the log-odds of switching language type during the Tangram task after interacting with the Figurative AI preference (Estimate = 10.81, 95%CI [1.96, 36.50]), while interaction with the Semi-Figurative preference AI resulted in a decrease in the log-odds of switching language type (Estimate = -9.32, 95%CI [-35.30, -0.18]) but that this decrease was marginally credible as the 95% credible interval just excludes 0. However, taken in conjunction with the summary results mentioned above it can be inferred that participants did switch to figurative language use after interaction with the Semi-Figurative preference AI, but that due to a small sample size, two participants not switching may have skewed the results.

Table 47: Likelihood of switching language type during Tangram task - non-native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	10.81	9.29	[1.96, 36.50]
Semi-figurative AI	-9.32	9.33	[-35.30, -0.18]
Literal AI	-10.86	9.33	[-36.38, -1.67]

Note: Estimates are log-odds. Brms model formula is `Switch_Binary ~ AI_Preference, family = bernoulli()`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

Figure 26 Estimated probability of switching language types in Tangram - non-native speaker



12.4 Post-test results

For the post-test there are two dependent variables: word count and language type use. Word count is first presented in terms of descriptive statistics for each training group analysed by AI language type preference. The results of Bayesian Regression Modeling provide an estimate of the word count used by each training group analysed by AI language type preference at this stage of the study. Language type use is subsequently presented through the analysis of the proportion of language type use, with Bayesian Regression Modeling results indicating the likelihood of a language type being used.

12.4.1 Post-test word count results

Tables 48, 49, and 50 present the descriptive statistics for mean word count, analysed by AI language type preference within each training group (Control group, Figurative training, and Literal training respectively).

Table 48: Post-Test Word Count Descriptives - non-native speaker Control Training

AI Preference	Mean	SD
Figurative AI	6.12	6.47
Literal AI	21.32	15.74
Semi AI	12.92	10.58

Note: This table displays the mean and standard deviation of word counts in the post-test the Control group.

Table 49: Post-Test Word Count Descriptives - non-native speaker Figurative Training

AI Preference	Mean	SD
Figurative AI	5.50	4.68
Literal AI	14.68	11.27
Semi AI	6.50	7.83

Note: This table displays the mean and standard deviation of word counts in the post-test the Figurative training group.

Table 50: Post-Test Word Count Descriptives - non-native speaker Literal Training

AI Preference	Mean	SD
Figurative AI	11.62	17.73
Literal AI	24.60	13.26

Semi AI	15.82	22.22

Note: This table displays the mean and standard deviation of word counts in the post-test the Literal training group.

For the dependent variable word count within the Control group, interaction with the Literal preference AI resulted in the highest mean word count (M = 21.32, SD = 15.74). The BRM presented in Table 51 indicates that interaction with the Literal preference AI is associated with a credibly higher estimated word count (Estimate = 15.28, 95% CI [4.89, 25.83]) compared to the Figurative AI baseline. However, interaction with the Semi-Literal preference AI indicates a slight increase in estimated word count (Estimate = 6.58, 95%CI [-4.39, 17.47]) but that this increase was not credible due to the inclusion of 0 in the credible interval. Therefore, indicating that interaction with the Semi-Literal AI preference did not cause a credible difference in estimated word count compared to the baseline interaction with the Figurative preference AI.

Table 51: Estimated word count (post-test) - non-native speaker Control group

	estimate	std.error	95%CI
Intercept (Figurative AI)	6.02	3.77	[-1.67, 13.67]
Literal AI	15.28	5.27	[4.89, 25.83]
Semi-Literal AI preference	6.58	5.35	[-4.39, 17.47]

Note: Brms model formula is `post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group, interaction with the Literal preference AI resulted in the highest mean word count (M = 14.68, SD = 11.27), and interaction with the Figurative and Semi-Literal preference AIs resulted in similar mean word counts (M = 5.50, SD = 4.68) and (M = 6.50, SD = 7.83) respectively. The BRM results presented in Table 52 indicates that interaction with both the Literal preference AI and Semi-Literal preference AI are associated with a higher estimated word count (Estimate = 13.17, 95% CI [-2.52, 29.04]) and (Estimate = 3.98, 95%CI [-12.07, 19.54]) respectively when compared to the Figurative AI baseline. However,

these increases were not credible due to the inclusion of 0 in both credible intervals. Therefore, indicating that interaction with both AI preferences did not cause a credible difference in estimated word count compared to the baseline interaction with the Figurative preference AI.

Table 52: Estimated word count (post-test) - non-native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Figurative AI)	10.78	5.75	[-0.77, 21.69]
Literal AI	13.17	8.18	[-2.52, 29.04]
Semi-Literal AI preference	3.98	8.05	[-12.07, 19.54]

Note: Brms model formula is `post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Literal AI preference represents an AI who communicates using literal language but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Literal training group, interaction with the Literal preference AI resulted in the highest mean word count (M = 24.60, SD = 13.26), while interaction with the Figurative preference AI resulted in noticeably lower mean word count (M = 11.62, SD = 17.73). The BRM results presented in Table 53 indicate that interaction with the Literal preference AI is associated with a credibly higher estimated word count (Estimate = 9.08, 95% CI [2.43, 15.49]) compared to the Figurative AI baseline. However, interaction with the Semi-Figurative preference AI did not result in a credible difference in estimated word count (Estimate = 0.80, 95%CI [-5.51, 7.15]) compared to baseline interaction with the Figurative preference AI. Therefore, indicating that interaction with the Semi-Figurative AI preference did not cause a credible difference in estimated word count compared to the baseline interaction with the Figurative preference AI.

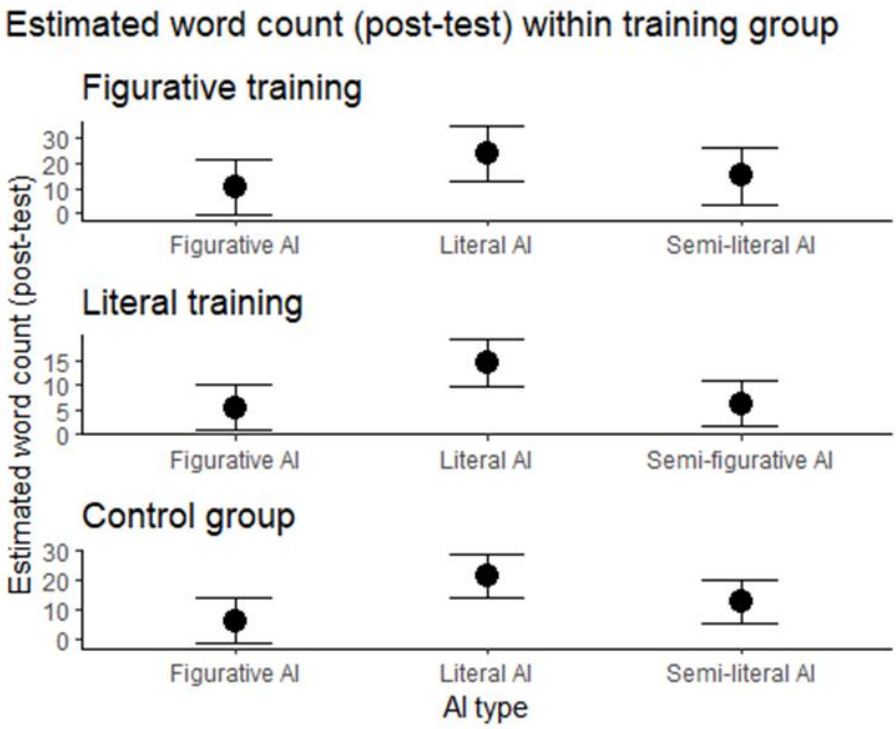
Table 53: Estimated word count (post-test) - non-native speaker Literal training

	estimate	std.error	95%CI
Intercept (Figurative AI)	5.35	2.30	[0.84, 9.92]
Literal AI	9.08	3.32	[2.43, 15.49]
Semi-Figurative AI preference	0.80	3.23	[-5.51, 7.15]

Note: Brms model formula is `post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian()`. `AI preference` is treatment-coded with Figurative preference, the baseline to which other AI preferences are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language, but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

The following plots estimate how word counts in the post-test (following initial training and interacting with an AI during the Tangram game) differ among participants who interacted with different types of AIs but had the same training (or no training in the case of the Control group). For participants with a given type of language experience (i.e. training condition), what is the effect of interacting with an AI who is either i/ congruent (same language preference as training), ii/ incongruent but cooperative (uses different language from training, but accepts either type from its interlocutor), or iii) incongruent and uncooperative (uses a different language type from training, and only accepts that language type from its interlocutor).

Figure 27. Estimated word count in the post-test - non-native speaker plots



Note: Plots represent conditional effects and 95% Credible Intervals from Brms model formula post_test_word_count ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = gaussian(). Plot A runs on the subset of participants who underwent figurative training, Plot B on participants with literal training, and Plot C on the control participants. Semi-Literal or semi-Figurative AI preference represents an AI who communicates using the named language type but accepts either language type from its interlocutor.

12.4.2 Post-test language type use results

Tables 54, 55, and 56 present the proportion of language type use, analysed by AI language type preference within each training group (Control group, Figurative training, and Literal training respectively).

Table 54: Proportion of language type use (post-test) - non-native speaker Control group

	Figurative language use	Literal language use	Mixed language use
Figurative AI	80%	2%	18%
Literal AI	0%	82%	18%
Semi AI	36%	24%	40%

Note: The table displays the proportion of language type use (post-test)

Table 55: Proportion of language type use (post-test) - non-native speaker Figurative training

	Figurative language use	Literal language use	Mixed language use
Figurative AI	88%	0%	12%
Literal AI	34%	28%	38%
Semi AI	80%	0%	20%

Note: The table displays the proportion of language type use (post-test)

Table 56: Proportion of language type use (post-test) - non-native speaker Literal training

	Figurative language use	Literal language use	Mixed language use
Figurative AI	68%	0%	32%
Literal AI	0%	84%	16%
Semi AI	54%	36%	10%

Note: The table displays the proportion of language type use (post-test)

For the dependent variable language type use within the Control group, interaction with the Figurative preference AI resulted in the highest proportion of figurative language use (proportion = 80%), while interaction with the Literal preference AI resulted in the highest proportion of literal language use (proportion = 82%). The BRM results presented in Table 57 indicates that, compared to the baseline interaction with Figurative AI, interacting with Literal AI is associated with a credibly higher log-odds of using literal language post-test (Estimate = 30.33, 95%CI [13.76, 62.64]). The interaction with Semi-Literal AI is associated with a credible higher log-odds of using literal language compared to the baseline interaction with Figurative AI (Estimate = 10.45, 95% CI [0.94, 26.37]). These credible increases in the estimated log-

odds for literal language use compared to the Figurative preference AI baseline indicate that interaction with both the Literal and Semi-literal preference AIs resulted in increased log-odds of literal language use in the post-test for Control group participants.

Table 57: Estimated language type use (post-test) - non-native speaker Control group

	estimate	std.error	95%CI
Intercept (Literal language)	-13.30	6.42	[-29.77, -4.87]
Intercept (Mixed language)	-4.40	2.32	[-9.71, -0.84]
Literal language:Literal AI	30.33	12.40	[13.76, 62.64]
Literal language:Semi-literal AI	10.45	6.48	[0.94, 26.37]
Mixed language:Literal AI	13.99	6.49	[3.75, 29.72]
Mixed language:Semi-literal AI	4.04	3.05	[-1.16, 10.84]

Note: Brms model formula is `post_test_language_type ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = categorical(link = "logit")`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

Within the Figurative training group AI preferences resulted in noticeably different proportions of language type use with the Figurative and Semi-Literal preference AIs resulting in high proportion use of figurative language use (proportion = 88%, figurative) and (proportion = 80%, figurative) respectively; while interaction with the Literal preference AI resulted in an almost even spread of language type use amongst the three types of language used within the experiment. The BRM presented in Table 58 indicates that, compared to the baseline interaction with Figurative AI, interacting with Literal preference AI is associated with a credibly higher log-odds of using literal language post-test (Estimate = 69.98, 95% CI [27.50, 158.10]). Similarly the model also indicates that interaction with the Semi-Literal preference AI results in credibly higher log-odds for literal language use (Estimate = 30.11, 95% CI [5.25, 71.49]).

Table 58: Estimated language type use (post-test) - non-native speaker Figurative training

	estimate	std.error	95%CI
Intercept (Literal language)	-34.73	16.76	[-77.07, -11.64]
Intercept (Mixed language)	-3.03	2.85	[-9.86, 1.57]
Literal language:Literal AI	69.98	34.29	[27.50, 158.10]
Literal language:Semi-literal AI	30.11	17.17	[5.25, 71.49]
Mixed language:Literal AI	19.48	11.64	[2.98, 47.61]
Mixed language:Semi-literal AI	-4.21	4.89	[-16.22, 2.84]

Note: Brms model formula is `post_test_language_type ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = categorical(link = "logit")`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Literal AI preference represents an AI who communicates using literal language, but accepts figurative language from its interlocutor. 95%CI are Credible Intervals.

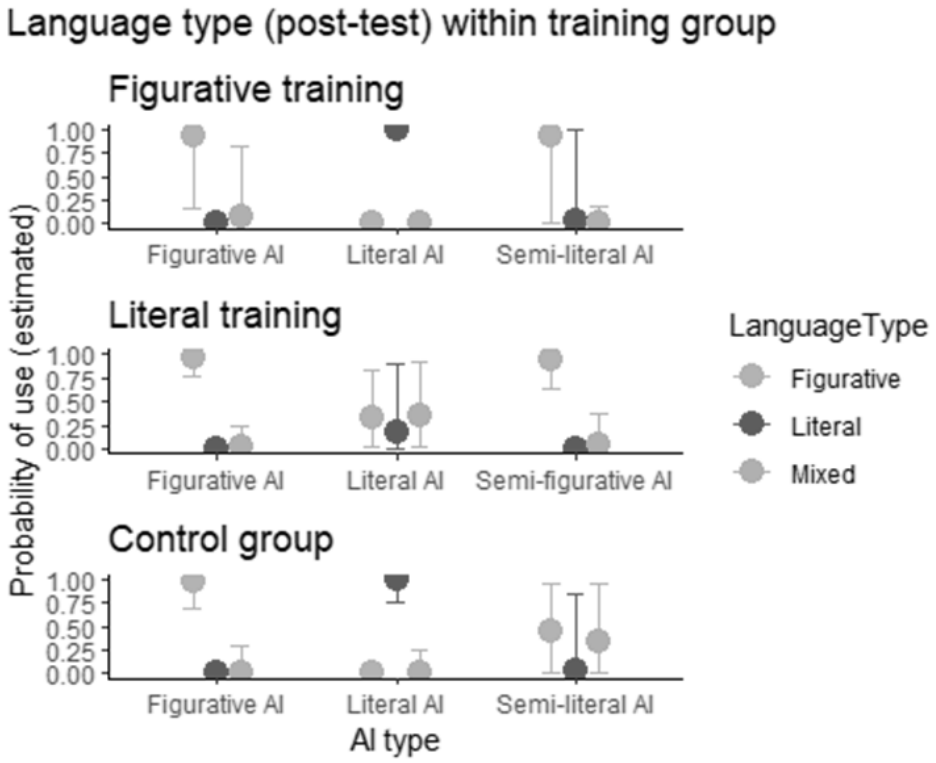
Within the Literal training group, AI preferences resulted in noticeably different proportions of language type use, with the Figurative and Literal preference AIs resulting in a high proportion use of their preferred language type (proportion = 68%, figurative) and (proportion = 84%, literal); while interaction with the Semi-Figurative AI resulted in a noticeably high proportion of figurative language use (proportion = 54%).The BRM presented in Table 59 indicates that, within the Literal training group, compared to the baseline interaction with Figurative AI, interacting with Literal AI is associated with a credibly higher log-odds of using literal language post-test (Estimate = 34.57, 95% CI [4.96, 181.15]). However, interaction with the Semi-Figurative preference AI did not result in a credible difference in the log-odds of literal language use compared to the Figurative AI baseline (Estimate = -1.86, 95%CI [-123.73, 122.64]. Therefore, indicating that interaction with the Semi-Figurative preference AI results in similar log-odds of literal language use compared to the baseline interaction with the Figurative preference AI.

Table 59: Estimated language type use (post-test) - non-native speaker Literal training

	estimate	std.error	95%CI
Intercept (Literal language)	-35.39	45.48	[-184.36, -5.93]
Intercept (Mixed language)	-3.90	1.63	[-7.52, -1.20]
Literal language:Literal AI	34.57	45.28	[4.96, 181.15]
Literal language:Semi-figurative AI	-1.86	60.18	[-123.73, 122.64]
Mixed language:Literal AI	3.93	2.22	[-0.08, 8.79]
Mixed language:Semi-figurative AI	0.68	2.13	[-3.59, 4.85]

Note: Brms model formula is `post_test_language_type ~ AI_Preference_Code + (1 | ParticipantID) + (1 | post_test_Image), family = categorical(link = "logit")`. `AI preference` is treatment-coded with Figurative AI, the baseline to which other AI types are compared. Semi-Figurative AI preference represents an AI who communicates using figurative language but accepts literal language from its interlocutor. 95%CI are Credible Intervals.

Figure 28. Estimated language type use in post-test - non-native speaker plots



12.5 Summary of Study 3: Non-native speaker results

This section presents a summary of the results of study 2 for non-native speakers of English (n=90). This section covers the following research question and hypotheses.

- Research question 1: Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?
- Hypothesis 1: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that in the pre-test that without training participants should favour the use of figurative language-based descriptions.
- Hypothesis 2: Based on the assertion that figurative language used in descriptions is less effortful (see Study 1 for findings that figurative descriptions require fewer words) it is hypothesised that untrained participants should show a bias towards figurative descriptions and perform similarly to the figurative description pre-training group throughout the experiment.
- Hypothesis 3: Without an increase in communicative effort due to breakdowns, participants will maintain their trained or untrained language type used for describing images (in the case of control group participants) unless participants come across less effortful means of achieving the goals of the communicative task.
    - For example, figurative and control group participants do not switch to literal descriptions in the semi condition since there are no breakdowns by design, while the literal group will switch to figurative descriptions in the semi condition even when there are no breakdowns due to an economy of effort.
- Hypothesis 4: Participants will switch to the description type of the AI that causes breakdowns when the AI refuses to understand the participants' descriptions.
    - For example, figurative and control participants switching to literal descriptions and literal participants switching to figurative.

Key results are summarised in Table 60 for ease of reference. These include the hypothesis tests for all dependent measures in the Tangram task, and descriptives for language type use in the post-test. credible differences between conditions are indicated by use of bold.

Table 60: Summary of dependent measure in Tangram and Post-test - non-native speakers

	Figurative training	Literal Training	Control group
Tangram: number of turns	**Literal AI > Fig AI** = Semi AI	Lit AI = Fig AI = Semi AI	Lit AI = Fig AI = Semi AI
Tangram: breakdowns	**Lit AI > Fig AI** = Semi AI	**Lit AI = Semi AI < Fig AI**	**Lit AI > Fig AI** = Semi AI
Tangram: Switches	**Lit AI > Fig AI** = Semi AI	**Lit AI = Semi AI < Fig AI**	**Lit AI > Fig AI** = Semi AI
Post-test: word count	**Lit AI > Fig AI** = Semi AI	**Literal AI > Fig AI** = Semi AI	**Literal AI** > **Figurative AI** = Semi AI
Post-test: language type	**Literal AI = 28% literal language use Figurative AI = 88% Figurative language use Semi AI = 80% Figurative language use**	**Literal AI = 84% literal language use Figurative AI = 68% Figurative language use Semi AI = 54% Figurative language use**	**Literal AI = 82% literal language use Figurative AI = 80% Figurative language use Semi AI = 24% Literal language use**

12.5.1 Hypothesis 1

In terms of the first hypothesis, the pre-test results show that non-native speaker participants used a mean word count of 12.28, with a standard deviation of 13.94. Figurative language was used in 237 descriptions (53%) and Mixed in 198 descriptions (44%). The BRM confirms that without training non-native speaker participants are less likely to use literal language compared to figurative language, but that figurative and mixed language are equally likely to be used. This result confirms hypothesis 1 as purely figurative language was the most used language type in the pre-test.

12.5.2 Hypothesis 2

In terms of hypothesis 2, non-native speakers (Control and Figurative training group participants) performed very similarly to native speakers as discussed below with this performance, generally confirming the hypothesis even for non-native speakers interacting with the use of L2 English.

The BRM results for all the dependent variables for the Tangram task mirrored native speaker results to a very similar degree, with the major difference being a slightly smaller observable difference in effect size for non-native speakers compared to native speakers. Furthermore, post-test results for the dependent variable word count also indicated a very similar pattern for the influence of AI interaction to native speakers, where interaction with the Literal preference AI resulted in a credible increase in estimated word count for both groups compared to the baseline interaction with the Figurative preference AI. While interaction with the Semi-Literal preference AI did not result in a credible increase in estimated word count compared to the baseline interaction i.e. the estimates were similar.

However, post-test language type use reveals that interaction with the Semi-Literal preference AI did result in a credible increase in the log-odds of literal language use for both groups which was not the case for native speakers. This indicates that non-native speakers were more susceptible to the influence of AI language type preference, although these results may also reflect the issue of a small sample size as the proportion of figurative language use was still considerably higher than literal language use for both groups. When taken in conjunction with the estimates for word count it appears that participants may have used abbreviated literal descriptions, under the assumption that this was what the task required. Which indicates that the addition of an exit interview to contextualise participants' performance can help with the interpretation of results. Overall, these results suggest that it is likely that Control group participants perform similarly to Figurative training participants in terms of non-native participants but require a larger sample size to confirm the hypothesis for non-native speakers.

12.5.3 Hypothesis 3

In terms of hypothesis 3, interaction with the Semi preference AI across both trained participant groups indicates that participants were influenced by an economy of effort, but this does not conclusively confirm this hypothesis when taking into account model results, likely due to the influence of the small sample size available. Furthermore, the results for non-native speakers for both Figurative and Literal training participants again showed similar patterns in

performance to native speaker participants, with the major differences being a smaller effect size.

In terms of the Tangram Task the BRM results did not indicate that interaction with the Semi-preference AI for both training groups resulted in a credible increase in the estimate when compared to baseline interaction with the Figurative preference AI. When considering that the Semi-preference AI does not result in breakdowns by design, this indicates that for both groups interaction with the Semi-preference AI did not result in credible differences in the log-odds of language type switch when compared to the baseline interaction with the Figurative preference AI. For Figurative training participants this means that the Semi-literal preference AI did not cause participants to switch to literal language use since the baseline interaction indicated that the Figurative preference AI did not show a tendency towards language type switch. However, the inverse is true for Literal training participants, where the baseline interaction resulted in a credible increase in the log-odds of observing a language type switch for this group. Meaning that when interaction with the Semi-Figurative preference AI for Literal training participants caused participants to switch language type and use figurative language since the Semi-Figurative preference AI resulted in performance that was not credibly different compared to the baseline interaction i.e. the Semi-Figurative preference AI resulted in log-odds of language type switch similar to the baseline.

Similarly to native speakers' post-test results also show that interaction with the Semi-preference AI resulted in a similar estimate in word count to baseline interaction with the Figurative preference AI for both Figurative and Literal training participants. However, the BRM results for language type use for the Figurative training group indicate that interaction with the Semi-preference AI did result in a credible increase in the log-odds of literal language use, although this may reflect the influence of sample size as only 4 of 10 participants were observed to switch from figurative to literal language type use during interaction in the Tangram task. Conversely, interaction with the Semi-preference AI resulted in log-odds of literal language type use that were not credible in difference compared to the baseline interaction with the Figurative preference AI indicating similar use of figurative language descriptions.

In terms of the hypothesis, Figurative training participant results in the post-test indicate that they were susceptible to the influence of the Semi-preference AI language preference, but that this may be an issue in sample size influencing the credibility of the results. Literal training participants performance, however, does indicate that participants that were trained on the use of higher effort literal descriptions opportunistically adopted the use of lower effort figurative descriptions without added effort in the form of communicative breakdowns. Overall, the

similarities in the patterns of results to native speakers indicate that a larger sample size is likely to confirm this hypothesis in potential future replications.

12.5.4 Hypothesis 4

In terms of hypothesis 4, the descriptive and summary results indicate the potential influence of an economy of effort and confirm the hypothesis. As trained participants that interacted with an AI that had the same language type preference (e.g. figurative trained participant interacting with Figurative preference AI) were observed to have less breakdowns and less language type switches than their counterparts that interacted with an AI that did not have the same language type preference (e.g. literal trained participant interacting with Figurative preference AI).

The BRM results indicate that Figurative group participants showed a credible increase in the estimate of total turns taken after interaction with the Literal preference AI; however, interaction with the Figurative preference AI did not show a credible increase in the estimate of total turns taken for the Literal training group, as interaction with the Literal preference AI resulted in a 95% credible interval that included zero, indicating weak evidence or a lack of credibleincrease in estimated number of turns taken, with a small magnitude of effect. However, the BRM result for breakdowns indicate that interaction with the Figurative preference AI did result in a credible increase in the estimate for breakdowns. This increase was relatively small compared to the credible increase in the estimate for breakdowns after interaction with the Literal preference AI for Figurative training participants, indicating that these participants were more resistant to switching and attempted to use figurative language descriptions more times than their Literal training counterparts.

In terms of language type switch, for Literal training participants interaction with the Figurative preference AI resulted in a credible increase in the log-odds of language type when compared to interaction with the Literal preference AI. Similarly, the inverse occurred for Figurative training participants where interaction with the Literal preference AI resulted in a credible increase in the log-odds of language type switch compared to the baseline interaction with the Figurative preference AI that did not indicate a tendency towards language type switch. Post-test results also revealed a similar pattern where interaction with the Literal preference AI resulted in a credible increase in the estimate for word count and a credible increase in the log-odds of literal language use for Figurative training participants; while interaction with the Figurative preference AI resulted in a credible decrease in estimated word count and a credible

decrease in the log-odds of both literal and mixed language use thereby indicating the use of figurative language.

Overall, the comparison of trained participant groups indicates that participants that interacted with an AI preference that matched their training were observed to have less breakdowns and were less likely to switch language type use; however, when they interacted with an AI preference that did not match their training they were observed to have credibly more breakdowns and were credibly more likely to switch language type use. These observations are in line with the expected influence of a tendency towards an economy of effort and indicates that the experimental paradigm is able to operationalise, manipulate, and measure the influence of an economy of effort in communication.

12.6 Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?

Based on the overall results for non-native speakers and their similarity to the results of native speakers, it again appears that the experimental paradigm used in studies 2 and 3 shows good potential for use in experimentally operationalising and manipulating an economy of effort for both native and non-native speakers.

13. Summary comparison of native and non-native speaker results

This section presents a summary comparison of the results of studies 2 and 3 in order to answer research question 2 Does an economy of effort influence communicative interaction differently for native and non-native speakers? The results are compared by performance across each of the 4 stages of the experimental paradigm.

13.1 Stage 1 Pre-test

In terms of the pre-test the dependent variables were word count and language type use, the results of which were both very similar for native and non-native speaker participants alike. Both native and non-native speakers used a higher proportion of figurative language and these results were confirmed by the Bayesian Regression Model that without any training or instructions, people are less likely to use literal language than figurative language, but that they are equally likely to use mixed language and figurative language.

13.2 Stage 2 Training

In terms of the training stage, both native and non-native speakers showed near identical mean performance on accuracy overall, but non-native speakers were marginally less accurate in second half performance overall.

13.3 Stage 3 Tangram task

In terms of the Tangram task stage, the dependent variables were total turns taken, total breakdowns, and language type switch. Performance on these dependent variables were extremely similar for both native and non-native speakers, with non-native speakers showing similar patterns of performance across all groups and dependent variables to native speakers but with a smaller effect size. This is evidenced by the near identical performance of native and non-native speakers across all Bayesian Regression Models used in analysis.

13.4 Stage 4 Post-test

In terms of the post-test the dependent variables were again word count and language type use. Again, both native and non-native speakers showed similar patterns of performance across all groups for both dependent variables. This was the case for word count in terms of AI

preference where for example interaction with the Literal preference AI increased word count for all three group types across both native and non-native speakers with a smaller effect size for non-native speakers. This was also the case for language type use in the post-test, where both the proportions of language type use and the Bayesian Regression Models also showed very similar patterns of performance but a smaller effect size for non-native speakers.

13.5 Concluding remarks for research question 2 discussion

As noted across the previous comparison of results between native and non-native speakers, performance is similar for both groups with a smaller effect size for non-native speakers overall. This indicates that within the context of this experimental paradigm, an economy of effort appears to influence the performance of both participant groups similarly; meaning that with regards to research question 2, an economy of effort does not appear to influence communicative interaction differently for native and non-native speakers. Furthermore, the similarity in performance for both groups suggests that an economy of effort may not necessarily influence native speakers differently to non-native speakers, but that context may be why non-native speakers are observed to be overly explicit as noted by studies of non-native collaborative interaction (e.g. Ryan 2015).

14. Conclusion

The aim of this research was to explore the notion of an economy of effort as a variable with the potential to influence communicative interaction and subsequently the potential end-state of adult usage based second language acquisition. As this variable is currently under explored in the literature, this thesis carried out two studies to answer the following research questions:

- General research question: Does an economy of effort influence communicative interaction?
- Research question1: Can the notion of an economy of effort be experimentally operationalised and manipulated in the context of communicative interaction?
- Research question 2: Does an economy of effort influence communicative interaction differently for native and non-native speakers?

The Tangram Task results for study 2 show that the experimental paradigm was successful in operationalising and manipulating the notion of an economy of effort in a communicative context for native speakers, thereby answering research question 1. The pre-test results indicated that purely figurative descriptions were the most used descriptions followed by mixed (literal and figurative descriptions), which was reflected in the results of the control group. For these participants, the highest mean number of turns taken and breakdowns when interacting the Literal preference AI with breakdowns showing a credible increase in the likelihood of experiencing a breakdown. Similarly, the figurative and literal training groups both experienced higher mean turns taken when interacting with the incongruent preference AI (i.e. figurative training with Literal preference AI). Furthermore, the experienced a credible increase in the likelihood of experiencing breakdowns during these incongruent interactions. In terms of the log-odds of language type switch when interacting with an incongruent preference AI that caused increased breakdowns. However, unlike the control and figurative training groups, the literal training group also showed a credible increase in the log-odds of language type switch when interacting with the Semi preference AI that were in line with the log-odds that resulted from interaction with the Figurative preference AI. This occurred without breakdowns which was by design since the Semi preference AI only models language that runs counter to participant training and accepts all descriptions from participants regardless of training group.

Taken together, these results answer the genera research question above in terms of native speakers, that the notion of an economy of effort does influence communicative interaction. This

becomes especially clear when considering that the Semi preference AI merely models language the language that runs counter to participant training (i.e. figurative descriptions for literal training participants). In these interactions, participants did not make use of this different descriptive behavior in their own descriptions when it did not represent a reduction in communicative effort, as figurative and control group participants maintained the use of figurative descriptions during the Tangram task stage and into the post-test, while literal training participants switched to using figurative descriptions when they were modeled by the AI when they represented a reduction in communicative effort. The results showed that breakdowns were required for participants to switch from the lower effort strategy of using figurative description to the more effortful literal descriptions. Furthermore, these results highlight the ecological validity of interaction-hypothesis" based theories. Development only occurs after breakdown, and breakdown only occurs when an interlocutor (in this case the confederate AI) refuses to accept language which it classes as "inaccurate".

In terms of research question 2, non-native speakers showed a similar general tendency towards minimising their communicative effort as their native speaker counterparts, but with a smaller effect size. Non-native speaker participants showed similar performance in terms of turns taken and breakdowns to their native speaker counterparts within the same groups, while the log-odds of language type switch also mirrored their native speaker counterparts. Furthermore, the influence of the Semi preference AI was also mirrored where control and figurative training group participants did not switch without breakdowns in this condition, while literal training participants had a credible increase in the log-odds for language type switch without breakdowns. Therefore, under the conditions of this experimental paradigm, it can be inferred that both native and non-native speakers are influenced by a tendency towards an economy of effort and reduction of total effort in communicative interaction in a similar but smaller manner.

The following sections present the conclusions based on the summary of results for research question 1 Can the notion of an economy of effort be experimentally operationalised in the context of communicative interaction? And research question 2 Does an economy of effort influence communicative interaction differently for native and non-native speakers? Furthermore, this section also presents the contribution of the results of studies 2 and 3 for the field of applied linguistics and usage-based adult second language acquisition. This is followed by a discussion of the limitations of the current research.

14.1 Conclusions based on the summary of results for research question 1

In terms of research question 1 Can the notion of an economy of effort be experimentally operationalised in the context of communicative interaction? The results indicate that the paradigm itself was able to operationalise, measure, and manipulate the notion of an economy of effort across several experimental stages.

For the pre-test stage, the results indicated that participants were more likely to use figurative language descriptions without prompting or intervention. Based on the findings of study 1 that showed that literal language was indeed more effortful to produce with a higher word count. Therefore, the use of description type appears to be a suitable measure to examine the unprompted descriptive tendencies of participants and investigate if these tendencies align with the characteristics of a tendency towards an economy of effort, which again does appear to be the case based on the observed preference for figurative language use in descriptions.

For the training stage, similarly to the results of Ellis and Sagarra (2010b, 2011) the 2-answer forced choice design of the stage was able to influence participants language type use during the Tangram task. This allowed the manipulation of an economy of effort through the preferences of the AI participants interacted with as an independent variable. Where participants now enter a communicative interaction stage with a manipulated baseline for descriptive language choice that can now be further manipulated and interacted with based on if the AI preferences align with participants' descriptive language choice.

For the Tangram stage, the results indicated that AI preferences were able to manipulate participants' descriptive language choice and use, as an independent variable. With participants switching between figurative and literal language descriptions in response to their tendency towards an economy of effort in conjunction with AI preferences that enforced their preference through communicative breakdowns or that offered the opportunity to use a different description type without enforcing their preference through breakdowns. Furthermore, the influence of AI preferences as an independent variable was observable via the dependent variables total turns taken, total breakdowns, and language type switch, indicating that an economy of effort was successfully operationalised and manipulated in this stage, and that the influence of AI preferences as an independent variable were measurable. Where participants that were trained on figurative language use were more resistant to switching to literal language use when faced with breakdowns and were highly unlikely to switch to literal language in Semi-Literal preference AI conditions indicating a tendency towards an economy of effort. While participants that were

trained on literal language use were less resistant to switch to figurative language when faced with breakdowns and were highly likely to switch to figurative language use in Semi-Figurative preference AI conditions again indicating a tendency towards an economy of effort.

Therefore, the dependent variable of total breakdowns adds important contextualisation to the variable language type switch on how and why a switch occurs and subsequently how an economy of effort was manipulated to incentivize a language type switch through breakdowns or as an unprompted response to the availability of satisficing solutions that offer a reduction in overall effort. Therefore, distinguishing between switches that occurred due to breakdowns, thereby switching to avoid them, and switches that occurred when participants noticed an opportunity to use less effortful descriptions when it appeared that the AI preference allowed it or preferred it. Finally, the dependent variable language type switch also offers important insight when switching does not occur when participants interact with AI preference types that match their trained description type or natural description tendencies, especially in the case of literal training participants interacting with a Literal preference AI. Where these did not encounter breakdowns and did not switch language type use, indicating that when an available means of satisficing is good enough it is likely that an economy of effort will not incentivize the search for less effortful satisficing solutions as the search may be more effortful overall than reusing currently effective means of satisficing their communicative goals.

For the post-test stage, the results indicated that interaction with the AI during the Tangram task influenced the number of words and which language type participants used to describe images in the post-test, i.e. the influence carried over to a novel scenario. This finding itself indicates that the dependent variables word count and language type in the post-test were sufficient to operationalise an economy of effort in terms of measuring the influence of manipulations during the Tangram task. Since participants results showed that interaction with the Literal preference AI resulted in increased word count and the use of literal language in their descriptions in the post-test, while interaction with the Figurative preference AI for all participants and the Semi-Figurative preference AI for literal training participants resulted in a smaller word count and the use of figurative language in their descriptions in the post-test.

Overall, these results indicate that the experimental design used was successful in operationalising, measuring, and manipulating an economy of effort in communicative interaction. Furthermore, these results indicate that controlling for interlocutor preferences is a valid means of manipulating the economy of effort in communicative interaction for participants engaged in such an experimental paradigm; with this manipulation being successful whether it was through communicative breakdowns or offering implicit opportunities to switch language

type in the Semi-preference condition. Finally, these results speak to the ecological validity of ‘interaction-hypothesis” based theories. Development only occurs after breakdown, and breakdown only occurs when an interlocutor (in this case the confederate AI) refuses to accept language which it classes as “inaccurate”. As such it would be interesting to investigate the influence of manipulating an economy of effort in communication in terms controlling interlocutor preference for the accuracy of morphological cues and its influence on learning and acquisition.

14.2 Conclusions based on the summary of results for research question 2

In terms of research question 2 Does an economy of effort influence communicative interaction differently for native and non-native speakers? The comparison of results in section 13 between native and non-native speakers across studies 2 and 3 indicate that overall, both populations largely follow similar patterns of performance within this experimental paradigm. With the main difference being that non-native speakers show a smaller effect size compared to their native speaker counterparts. However, it must be taken into account that this similar pattern in performance occurred within the context of a manipulated communicative interaction where participants were under the impression that they were interacting with an artificial interlocutor. Meaning that this similarity is not necessarily evidence that runs counter to the observed differences in an economy of effort between native speakers and non-native speakers in naturalistic contexts mentioned in previous studies (e.g. Cheng & Warren, 1999; Feng, 2022; Önen & İnal, 2019; Ryan, 2015); but under the circumstances of this experimental paradigm native and non-native speakers both follow similar patterns of performance.

14.3 Contribution and implications of this research

The main contribution of this research is the development and testing of a promising experimental paradigm that allows the notion of an economy of effort to be further explored in terms of its influence on communicative interaction and its subsequent potential influence on the end-state of naturalistic adult section language acquisition. Overall, this experimental paradigm brings together aspects of paradigms that investigate communicative interaction (i.e. the Tangram task) and paradigms that test the effects of blocking (i.e. the training stage) in associative learning and using various operationalizations to define independent variables to manipulate an economy of effort and dependent variables that show the influence of these manipulations. The results of studies 2 and 3 indicate that this paradigm was able to

successfully bring together these aspects to demonstrate the influence of an economy of effort in communicative interaction. Meaning that this experimental paradigm has the potential to provide the next step for research into usage-based language acquisition by providing a link between associative learning and experimentally inducing blocking of certain linguistic cues and testing the influence of interaction that can be manipulated to interface with blocking. Essentially, if linguistic cues can be mapped to dimensions similar to those of figurative and literal language descriptions used in the current experimental paradigm, and AI preferences can be similarly mapped to these dimensions, the experimental paradigm used in studies 2 and 3 offers a promising means of testing the influence of interaction and an economy of effort on the acquisition of these cues.

This means that questions such as Does an economy of effort in communicative interaction influence the outcomes of usage-based adult second language acquisition? Can be potentially answered with further refinement of this experimental paradigm. One potential means of mapping these dimensions is to first consider figurative language descriptions as a representation of more salient and less effortful linguistic cues that are effective at satisficing communicative goals and may potentially block the acquisition of other less salient cues; while literal language represents the less salient and more effortful linguistic cues that are made contextually redundant by the availability of figurative language descriptions. A possible means of substituting these dimensions with linguistic cues is the use of numeral adverbials as the more salient and less effortful cue dimension and noun inflections of quantity as the less salient and more effortful cue dimension.

This would be similar to the use of temporal adverbials and verbal inflection cues of temporal reference used in Ellis and Sagarra (2010b, 2011), and participants can therefore be similarly trained on their use during a 2-answer forced choice training stage. Furthermore, these can be similarly mapped to the dimensions of figurative and literal language use, AI preferences, and can be operationalised through the switch and breakdown dependent variables (i.e. switch in cue type use and total breakdowns associated with cue type use). If these cue dimensions are created in a miniature artificial language for example, with the goal of describing images based on the number of objects in them in a Tangram task setting; it becomes possible to test the influence of communicative interaction and an economy of effort on the blocking of linguistic and the potential to recover from said blocking in associative learning. Again, this means that the experimental paradigm used in studies 2 and 3 shows promising potential to help in answering more questions about usage-based adult second language acquisition.

However, it is imperative for the validity of results gathered on the basis of remapping both cue dimensions of temporal reference can be remapped to cues of numeral reference, that the experiment used in Ellis in Sagarra (2011) is closely replicated i.e. repeating the original experiment with the exception of one major change involving one of the variables of the experiment (McManus, 2022). This replication should specifically focus first on the influence of cue frequency on the learnability of verb inflections and adverbial cues of temporal reference. As McManus (2022) suggests, the frequency of adverbial cues of temporal reference could be brought in line with that of verbal inflectional cues of temporal reference by increasing the number of adverbial cues, thereby reducing the number of times each individual cue is seen. As discussed in section 4, the size of the cue dimension has a potential link to an economy of effort where a smaller set of physically more salient cues such as adverbial cues is likely more economical to use and maintain for the purpose of correctly disambiguating temporal reference.

Therefore, a series of replications focused on the influence of frequency can help further disambiguate how manipulating the set size of a cue dimension influences the results of the experiment. This will subsequently serve to guide future manipulations regarding the frequency of cues and how they influence the results of the experiment and what to expect from successive runs of the experiment. This then allows a further replication with the substitution of both cue dimensions of temporal reference can be remapped to cues of numeral reference. This replication would then confirm if these cues function similarly to cues of temporal reference used in Ellis and Sagarra (2010b, 2011) and the validity of their use in the aforementioned adaptation of this experimental paradigm. Where the cues are learned through a training stage, and the influence of interaction with AI preferences on acquisition in general and blocking specifically in a subsequent Tangram stage.

14.4 Limitations of the current research

While every effort was taken to ensure the robustness of this research and the studies within, it is important to acknowledge the limitations of this work. A key limitation of the results of studies 2 and 3 was found in the performance check of poisson models showing a poor fit in some cases such as for counts of total turns taken and total breakdowns because of over dispersion. For example, this occurs due to a lack of breakdowns observed in conditions where an AI preference matches participants training on language use. These checks indicated that another model family, the negative binomial family, would have been a better fit for response distribution. However, due to the extensive time Bayesian Regression Models take to run, it was

not feasible to rerun these models, and subsequent checks currently, leaving the current results as the best approximation that could be currently produced.

Another key limitation of the results of the study for Tangram task dependent variables is the lack of manual data review. Due to the limitations of time, it was not feasible to manually review the chat data for a total of 180 participants that took part in studies 2 and 3. As such this led to instances where the automation of data processing conducted via R within the Rstudio IDE registered false positives in some instances for these dependent variables. For example, instances where participants apologise for inadvertently placing images in the incorrect drop zone may have registered as a breakdown, and a subsequent increase in turns taken. This should be taken into account in future iterations of the experimental paradigm, where manual review of chat data is required to ensure the soundness of data processing.

An additional key limitation of the results of studies 2 and 3 was the relatively small number of participants in each group. Overall, both studies included 180 total participants (native = 90, non-native =90), however each training group was subset by AI preference resulting in 10 participants per group. This means that current results, while informative, still require more data from more participants to improve the quality of the results generated from the Bayesian Regression Models used. However, this was again difficult to do due to time constraints, as I was the confederate playing the role of the artificial interlocutor interacting with all participants across both studies.

This highlights the next important limitation of the study; the role of the artificial interlocutor being assigned to a confederate. Despite following a strict script for interaction with participants it cannot be discounted that a confederate researcher playing the role of the artificial interlocutor has the potential to influence the results of these studies. Whether this is due to any small bias, or even fatigue from interacting with multiple consecutive participants, the use of a confederate will always be to some degree problematic. However, with the surge in prominence of large language models in artificial intelligence such as ChatGPT that can be trained to play the role of the confederate participant, it is likely the case that future iterations of this experimental paradigm can depend on such language models to fulfil this role and avoid any issues of bias from human confederates.

Finally, the lack of an exit interview following the completion of the post-test task presents an additional limitation. As data gathered from such an interview may provide additional contextualisation to how participants performed, and what influenced their choice to switch language type for instance. Furthermore, participants could be questioned on the quality

of the experiment and what can be done to improve the participant's experience for future applications of the experimental paradigm.

References

Andringa, S., & Curcic, M. (2015). How explicit knowledge affects online L2 processing: Evidence from differential object marking acquisition. *Studies in Second Language Acquisition, 37*(2), 237-268.

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*(3), 292-305.

Badger, R. (2018). From input to intake: Researching learner cognition. *tesol QUARTERLY, 52*(4), 1073-1084.

Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological science, 19*(3), 241-248.

Bannard, C., Lieven, E., & Tomasello, T. (2009). Modeling Children's Early Grammatical Knowledge. *Proceedings of the National Academy of Sciences, 106*, (41), 17284-17289.

Bardovi-Harlig, K. (2000). Tense and aspect in second language acquisition: Form, meaning, and use. Language Learning: *A Journal of Research in Language Studies, 50*, 1.

Bartlett, F. C. (1995). *Remembering: A study in experimental and social psychology*. Cambridge university press.

Bauerly, M., & Liu, Y. (2008). Effects of symmetry and number of compositional elements on interface and design aesthetics. Intl. *Journal of Human–Computer Interaction, 24*(3), 275-287.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of personality and social psychology, 79*(6), 941.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of communication, 52*(3), 566-580.

Bavelas, J., Gerwing, J., & Healing, S. (2017). Doing mutual understanding. Calibrating with micro-sequences in face-to-face dialogue. *Journal of Pragmatics, 121*, 91-112.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence, 35*(8), 1798-1828.

Berger, C. R., & Calabrese, R. J. (1974). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human communication research, 1*(2), 99-112.

Berlyne, D. E. (1958). The influence of complexity and novelty in visual figures on orienting responses. *Journal of experimental psychology, 55*(3), 289.

Berlyne, D. E. (1974). The new experimental aesthetics. *Studies in the new experimental aesthetics*, 1-25.

Beukeboom, C. J. (2009). When words feel right: How affective expressions of listeners change a speaker's language use. *European Journal of Social Psychology*, *39*(5), 747-756.

Birdsong, D. (2009). Age and the end state of second language acquisition. *The new handbook of second language acquisition, 17*(1), 401-424.

Botvinick, M., & Braver, T. (2015). Motivation and cognitive control: from behavior to neural mechanism. *Annual review of psychology, 66*, 83-113.

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive science, 38*(6), 1249-1285.

Boyd, J. K., & Goldberg, A. E. (2009). Input effects within a constructionist framework. *The Modern Language Journal, 93*(3), 418-429.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. Journal of experimental psychology: *Learning, memory, and cognition, 22*(6), 1482.

Brown, D. (2007). *Principles of language learning and teaching*. Fifth Edition. PearsonLongman,USA.

Brown, R. (1970). Derivational complexity and order of acquisition. *Cognition and Development of Language*.

Cameron, D. (2001). Sequence and structure: Conversation Analysis. In D. Cameron (Ed.), *Working with Spoken Discourse* (pp. 87–105). Sage.

Carroll, S. (1989). Second-language acquisition and the computational paradigm. *Language Learning, 39*(4), 535-594.

Cheng, W., & Warren, M. (1999). Inexplicitness: What is it and should we be teaching it?. *Applied Linguistics, 20*(3), 293-315.

Chevrot, J. P., Dugua, C., & Fayol, M. (2009). Liaison acquisition, word segmentation and construction in French: a usage-based account. *Journal of child language, 36*(3), 557-596.

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in cognitive science, 11*(3), 468-481.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences, 39*, e62.

Cintrón-Valentín, M. C., & Ellis, N. C. (2016). Salience in second language acquisition: Physical form, learner attention, and instructional focus. *Frontiers in psychology, 7*, 195253.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. In L. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 127-149). Washington, DC: American Psychological Association.

Clark, H. H., & Schaefer, E. F. (1987). Concealing one's meaning from overhearers. *Journal of memory and Language, 26*(2), 209-225.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive science, 13*(2), 259-294.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1-39.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological review, 97*(3), 332.

Davies, B. L. (2007). Least collaborative effort or least individual effort: Examining the evidence. Univ. Leeds Work. *Pap. Linguist. Phon, 12*, 1-20.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language learning*.

Divjak, D. (2019). *Frequency in language: Memory, attention and learning*. Cambridge University Press.

Dye, M., & Ramscar, M. (2009). Error and expectation in language learning: An inquiry into the many curious incidents of" mouses" in adult speech. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 31, No. 31).

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition, 24*(2), 143-188.

Ellis, N. C. (2005). At the interface: Dynamic interactions of explicit and implicit language knowledge. *Studies in Second Language Acquisition, 27*, 305– 352.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics, 27*(2), 164-194.

Ellis, N. C. (2007). Blocking and learned attention in language acquisition. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 29, No. 29).

Ellis, N. C. (2008a). The associative learning of constructions, learned attention, and the limited L2 endstate. *Handbook of cognitive linguistics and second language acquisition*, 372-405.

Ellis, N. C. (2008b). The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The modern language journal, 92*(2), 232-249.

Ellis, N. C. (2008c). Implicit and explicit knowledge about language. *Encyclopedia of language and education*, *6*, 1-13.

Ellis, N. C. (2017). Salience in language usage, learning and change. *The changing English language: Psycholinguistic perspectives*, 71-92.

Ellis, N. C., & Sagarra, N. (2010a). Learned attention effects in L2 temporal reference: The first hour and the next eight semesters. *Language Learning, 60*, 85-108.

Ellis, N. C., & Sagarra, N. (2010b). The bounds of adult language acquisition: Blocking and learned attention. *Studies in Second Language Acquisition, 32*(4), 553-580.

Ellis, N. C., & Sagarra, N. (2011). Learned attention in adult language acquisition: A replication and generalization study and meta-analysis. *Studies in Second Language Acquisition, 33*(4), 589-624.

Ellis, R. (1997). Second language acquisition. *The United States: Oxford*, *98*.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in second language acquisition, 28*(2), 339-368.

Ellis, R., Sheen, Y., Murakami, M., & Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System, 36*, 353–371.

Feng, S. (2022). L2 tolerance of pragmatic violations of informativeness: Evidence from ad hoc implicatures and contrastive inference. *Linguistic Approaches to Bilingualism.*

Feng, S. F., Schwemmer, M., Gershman, S. J., & Cohen, J. D. (2014). Multitasking versus multiplexing: Toward a normative account of limitations in the simultaneous execution of control-demanding behaviors. *Cognitive, Affective, & Behavioral Neuroscience, 14*(1), 129-146.

Fernald A., Taeschner, T., Dunn, J., Papousek, M., Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of ChildLanguage, 16*, 477–501.

Fillmore, C. J., Kempler, D., & Wang, W. S. (Eds.). (2014). *Individual differences in language ability and language behavior*. Academic Press.

Fisher, C., & Tokura, H. (1996). Acoustic cues to grammatical structure in infant-directed speech: Cross-linguistic evidence. *Child development, 67*(6), 3192-3218.

Foster, P., & Ohta, A. S. (2005). Negotiation for meaning and peer assistance in second language classrooms. *Applied linguistics*, *26*(3), 402-430.

Gass, S. M., & Mackey, A. (2006). Input, interaction and output: An overview. *AILA review*, *19*(1), 3-17.

Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Routledge.

Geissler, G. L., Zinkhan, G. M., & Watson, R. T. (2006). The influence of home page complexity on consumer attention, attitudes, and purchase intent. *Journal of Advertising, 35*(2), 69-80.

Gergle, D., Kraut, R.E., & Fussell, S.R. (2004). Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language and Social Psychology, 23*, 491-517.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Goldberg, A., & Suttle, L. (2010). Construction grammar. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 468-477.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. *Phraseology: Theory, analysis and applications*, 145-160.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217-229.

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, *47*(4), 589-606.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, *1*(2), 151-195.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. In *The 39th Annual Conference of the Cognitive Science Society (CogSci 2017)* (pp. 564-569). Cognitive Science Society.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different?. Linguistic *Approaches to Bilingualism*, *4*(2), 257-282.

Kamin, L. J. (1967, June). Attention-like processes in classical conditioning. In *SYMP. ON AVERSIVE MOTIVATION MIAMI* (No. TR-5).

Klein, W. (1998). The contribution of second language acquisition research. *Language Learning*, *48*, 527– 550.

Klein, W., & Perdue, C. (1992). *Utterance structure: Developing grammars again* (Vol. 5). John Benjamins Publishing.

Klein, W., & Perdue, C. (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). *Second language research*, *13*(4), 301-347.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*(9), 1321-1333.

Kramer, M. W. (1999). Motivation to reduce uncertainty: A reconceptualization of uncertainty reduction theory. *Management Communication Quarterly*, *13*(2), 305-316.

Kruschke, J. K . (2006 , June). Learned attention. Paper presented at the Fifth International Conference on Development and Learning, Bloomington, IN.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636-645.

Langacker, R. W. (1987). Foundations of cognitive grammar: Volume II: Descriptive application.

Lieder, F., & Griffiths, T. L. (2015). When to use which heuristic: A rational solution to the strategy selection problem. In *CogSci*.

Lieder, F., & Griffiths, T. (2016). Helping people make better decisions using optimal gamification. In *CogSci*.

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in neural information processing systems* (pp. 2870-2878).

Lieven, E., Pine, J., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*, 187–220.

Lieven, E.V.M., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*(2).

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*. 481–507.

Lightbown, P. M., Spada, N., & White, L. (1993). The Role of Instruction in SLA: Introduction. *Studies in Second Language Acquisition*, *15*(2), 143-145.

Long, M. H. 1985. 'Input and second language acquisition theory' in S. Gass, and C. Madden (eds): *Input and Second Language Acquisition Rowley*. MA: Newbury House, pp. 268–86.

Long, M. H. (1990). The least a second language acquisition theory needs to explain. *TESOL Quarterly, 24*, 649–666.

Long, M. H . (1991). Focus on form: A design feature in language teaching methodology . In K. de Bot , R. Ginsberg , & C. Kramsch(Eds.), *Foreign language research in cross-cultural perspective* (pp. 39 – 52). Amsterdam : Benjamins.

Long, M. H. 1996. 'The role of the linguistic environment in second language acquisition' in W. C. Ritchie, and T. K. Bhatia (eds): *Handbook of research on Language Acquisition: Second language acquisition.Vol. 2*. New York: Academic Press,

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*, 471–497.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82*, 276–298.

MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, *43*(1/2). 1–123.

Madan, C. R., Bayer, J., Gamer, M., Lonsdorf, T. B., & Sommer, T. (2018). Visual complexity and affect: ratings reflect more than meets the eye. *Frontiers in Psychology*, *8*. 2368.

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

Mariscal, S. (2009). Early acquisition of gender agreement in the Spanish noun phrase: starting small. *Journal of Child Language*, *36*(1), 143-171.

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science, 9*, 637–652.

McManus, K. (2022). Crosslinguistic influence and L2 grammar learning: Proposed replications of Ellis and Sagarra (2011) and Tolentino and Tokowicz (2014). *Language Teaching*, *55*(4), 565-573.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological bulletin*, *117*(3), 363.

Mills, G. J., Purver, M, and Healey, P. G. (2013). *A dialogue experimentation toolkit*. https://dialoguetoolkit.github.io/chattool/

Mishra, R. K. (2015). Interaction between attention and language systems in humans. *A cognitive science perspective/RK Mishra.—New Delhi: Springer*.

Musslick, S., Dey, B., Özcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016a, August). Controlled vs. Automatic Processing: A Graph-Theoretic Approach to the Analysis of Serial vs. Parallel Processing in Neural Network Architectures. In *CogSci*.

Musslick, S., Dey, B., Ozcimder, K., Patwary, M. M. A., Willke, T. L., & Cohen, J. D. (2016b). Parallel processing capability versus efficiency of representation in neural networks. *Network*, *8*(7).

Nobre, A. C., and S. Kastner. 2014. 'Attention: Time Capsule 2013.' In Anna C. Nobre and Sabine Kastner (eds.), *The Oxford Handbook of Attention* (Oxford University Press: Oxford).

Önen, S., & İnal, D. (2019). A Corpus-Driven Analysis of Explicitness in English as Lingua Franca. *Journal of Curriculum and Teaching*, *8*(3), 73-83.

Onnis, L. (2012). The potential contribution of statistical learning to second language acquisition. *Statistical learning and language acquisition*, 203-235.

Pawley, A., & Syder, F. H. (1980). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R.Schmidt (Eds.), *Communicative competence* (pp. 191–225). London: Longmans.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(2), 169-190.

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*(2-3), 203-228.

Pine, J. M., & Lieven, E. (1997). Slot and frame patterns and the development of the determiner category. *Applied Psycholinguistics, 18*, 123–138.

Ramscar, M., & Gitcho, N. (2007). Developmental change and the nature of learning in childhood. *Trends in cognitive sciences, 11*(7), 274-279.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive science, 31*(6), 927-960.

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive science*, *34*(6), 909-957.

Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., & Gajos, K. Z. (2013, April). Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2049-2058).

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think it is. *American psychologist*, *43*(3), 151.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Classical conditioning II: *Current research and theory*, *2*, 64-99.

Ryan, J. (2015). Overexplicit referent tracking in L2 English: Strategy, avoidance, or myth?. *Language Learning*, *65*(4), 824-859.

Sabbah, S. (2015). Negative transfer: Arabic language interference to learning English. *Arab World English Journal (AWEJ) Special Issue on Translation*, (4).

Sacks, H., Schegloff, E. A., & Jefferson, G. (1978). A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction* (pp. 7-55). Academic Press.

Saryazdi, R., Bannon, J., Rodrigues, A., Klammer, C., & Chambers, C. G. (2018). Picture perfect: A stimulus set of 225 pairs of matched clipart and photographic images normed by Mechanical Turk and laboratory participants. *Behavior Research Methods*, *50*(6), 2498-2510.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks. In *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 1271–1276).

Schegloff, E. A. (1997). Practices and actions: Boundary cases of other-initiated repair. *Discourse processes*, *23*(3), 499-545.

Schmidt, R. W. (1990). The role of consciousness in second language learning1. *Applied linguistics*, *11*(2), 129-158.

Schmidt, R. (2001). Attention in P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge, UK: Cambridge Applied Linguistics.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive psychology*, *21*(2), 211-232.

Schumann, J. H. (1987). The expression of temporality in basilang speech. *Studies in Second Language Acquisition, 9*, 21–41.

Selten, R. (1999). What is bounded rationality? Paper prepared for the Dahlem Conference 1999. In *Dahlem Conference*.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*(2), 217-240.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, *40*, 99-124.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review*, *84*(2), 127.

Simon, H. A. (1957). *Models of man: Social and rational*. New York : Wiley.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161-176.

Simon, H. A. (1990). Bounded rationality. In *Utility and probability* (pp. 15-18). Palgrave Macmillan, London.

Slobin, D. I. (1993). Adult language acquisition: A view from child language study. *Adult language acquisition: Cross-linguistic perspectives*, *2*, 239-252.

Swain , M.(2005). The output hypothesis: Theory and research . In *E. Hinkel(Ed.), Handbook of research in second language teaching and learning* (pp. 471 – 484). Mahwah, NJ : Erlbaum.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Trudgill, P. (2002a). Linguistic and social typology. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 707–728). Oxford, UK: Blackwell.

Trudgill, P. (2002b). *Sociolinguistic variation and change*. Edinburgh, Scotland: Edinburgh University Press.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, *247*(4940), 301-306.

Weiner, B., (1995). Aspiration level. In: Manstead, A.S.R., Hewstone, M.(Eds.), *The Blackwell Encyclopedia of Social Psychology*. Blackwell, Oxford, p. 362.

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.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge UniversityPress.

Wray, A. (2004). 'Here's one I prepared earlier': Formulaic language learning on television. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use*(pp. 249–268). Amsterdam: John Benjamins.

Wray, A. (2008). The puzzle of language learning: From child's play to 'Linguaphobia'. *Language Teaching, 41*, 253–271.

Wu, W. 2014. *Attention* (Routledge: London).

Wurm, L. H., & Cano, A. (2010). Stimulus norming: It is too soon to close down brick-and-mortar labs. *The Mental Lexicon*, *5*(3), 358-370.

Yorio, C. A. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam & L. K. Obler (Eds.), *Bilingualism across the life-span* (pp. 55–72). Cambridge: Cambridge University Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison Wesley.

Appendix A: Ethics documentation, consent form, and exit debriefing

Ethical Issues Audit Form for Research Students

This questionnaire should be completed for each research study that you carry out as part of your degree.

A. Surname / Family Name: Al-Abdul Razzaq

B. First Name/ Given Name: Mohammad

C. Programme: PhD in Applied Linguistics

D. Supervisor (of this research study): Dr. Cylcia Bolibaugh

E. Topic (or area) of the proposed research study: Second Language learning.

F. Where the research will be conducted: Online research.

G. Methods that will be used to collect data: Rating questionnaire for norming materials; online learning experiment.

H. If you will be using human participants, how will you recruit them? Online via social media and personal connections for the piloting of studies, and then via paid online participant panels (Prolific Academic) for online studies. If in-person testing resumes over the next 12-18 months, recruitment may take place via advertisements on campus.

Supervisors, please read Ethical Approval Procedures – for students (Oct2018).

The application is a joint one by the research student and supervisor(s). It should be submitted to the TAP member for initial approval and then to the Higher Degrees Administrator who will seek a second opinion from a designated member of Education Ethics Committee. Forms may also require review by the full Ethics Committee (see below).

Approvals

(This section is to be completed by a staff member.)

First approval: by the TAP member (after reviewing the form):

Please select one of the following options.

Ethics statements	Tick one box
I believe that this study, as planned, meets normal ethical standards. I have checked that any informed consent form a) addresses the points as listed in this document, and b) uses appropriate language for the intended audience(s).	☒
I am unsure if this study, as planned, meets normal ethical standards	☐
I believe that this study, as planned, does not meet normal ethical standards and requires some modification	☐

Add TAP member's name: Leah Roberts

Add date: 31/01/2021

Approval: by a designated Ethics Committee member:

Please select one of the following options:

Ethics statements	Tick one box
I believe that this study, as planned, meets normal ethical standards. I have checked that any informed consent form a) addresses the points as listed in this document, and b) uses appropriate language for the intended audience(s).	☒
I am unsure if this study, as planned, meets normal ethical standards	☐
I believe that this study, as planned, does not meet normal ethical standards and requires some modification	☐

Add the name of Ethics Committee member: Ruggero De Agostini

Add the date: 23/03/2021

Consent form for Study 1 Image norming

Welcome!

You are being invited to take part in a research study. This research has been approved by the Department of Education, University of York Ethics Committee. If you have any questions or complaints about this research please contact the researcher (maar505@york.ac.uk) or the Chair of the Ethics Committee (education-research-admin@york.ac.uk).

Take your time to read the following information and decide whether or not you wish to take part.

Information about the study

What is the purpose of the study?

The goal of this study is to understand how people view and describe images. This study comprises a single session lasting approximately 20 min. You will be asked to perform a number of short tasks, for example describing images, or rating how appealing they are. Before each task you will be given detailed instructions about how to perform the task.

Who has access to your data and how will it be stored and used?

All of the data collected for this study will be anonymous. We will not ask for your name or any other identifying information. The anonymous data may be used in presentations, online, in research reports, in project summaries or similar. In addition the anonymous data may be used for further analysis. Your individual data will not be identifiable but if you do not want the data to be used in this way please do not complete the questionnaire.

What happens if I do not want to take part in the study?

You do not have to take part in this study if you do not want to do so. You are free to withdraw at any time during the study. Should you wish to withdraw, you are able to simply close the browser tab to withdraw from the study. You will not be able to withdraw your data after submission as we will be unable to trace your data.

By submitting this questionnaire, you are agreeing to all of the points above.

Many thanks for your help with this research.

Consent

By giving your consent you indicate that you understand and agree to the following.

1. I understand that my data is being collected fully anonymously.
2. I understand that my data will be collected and stored securely in a database, and agree to this.
3. I understand that my data may be used in presentations, online, in research reports, in project summaries, and further analysis; and agree to this.
4. I understand that I have the option to withdraw from the study and have my responses removed from the data base anytime before the completion of the study, and agree to this.

☑ I hereby give my consent that my anonymous data may be used for this study.

Next

Consent form for Studies 2 and 3

Welcome!

You are being invited to take part in a research study. This research has been approved by the Department of Education, University of York Ethics Committee. If you have any questions or complaints about this research please contact the researcher (maar505@york.ac.uk) or the Chair of the Ethics Committee (education-research-admin@york.ac.uk). Take your time to read the following information and decide whether or not you wish to take part.

Information about the study

What is the purpose of the study?

The goal of this study is to improve the ability of a learning digital personal assistant to describe images more naturally and identify those images using more natural descriptions. This study comprises a single session lasting approximately 30 min. You will be asked to perform a number of short tasks, that include describing images, some task based training and finally an interactive matching game with the personal assistant A.I.; Before each task you will be given detailed instructions about how to perform the task.

Who has access to your data and how will it be stored and used?

All of the data collected for this study will be anonymous. We will not ask for your name or any other identifying information. The anonymous data may be used in presentations, online, in research reports, in project summaries or similar. In addition the anonymous data may be used for further analysis. Your individual data will not be identifiable but if you do not want the data to be used in this way please do not complete the questionnaire.

What happens if I do not want to take part in the study?

You do not have to take part in this study if you do not want to do so. You are free to withdraw at any time during the study. Should you wish to withdraw, you are able to simply close the browser tab to withdraw from the study. You will not be able to withdraw your data after submission as we will be unable to trace your data.

By submitting this questionnaire, you are agreeing to all of the points above. Many thanks for your help with this research.

	consent
I hereby give my consent that my anonymous data may be used for this study.	☑

Next

Exit debriefing

Aim of the Study

The aim of this study was not to help develop a digital personal assistant; but in fact to study how interactions during communication influence how people communicate.

The main focus was to investigate if a tendency towards an economy of effort (i.e. finding a minimal effort solution to a problem) influenced communication such as changing how a person might describe an image to successfully convey what they intended to communicate.

The role of the digital assistant was played by the researcher. This premise was needed as it would not make sense to participants that another human participant would not be able to understand different types of natural descriptions.

Finally your data is completely anonymous, the anonymous Prolific ID is simply used to match your payment code to your contribution, to be able to authorise payment via the Prolific platform.

If you agree to allow your data to be used in this study please continue to the next page for your payment code. If not please close your browser at this point.

Next

Appendix B

Table B1:

Figurative naming agreement reported as proportion.

item	mean	sd	n
ant	0.96	0.37	81
bird_front	0.73	0.45	82
bird_over	0.87	0.34	83
butterfly	0.73	0.44	83
camel	0.89	0.32	81
dog	0.67	0.47	82
duck	0.68	0.47	81
elephant	0.93	0.26	82
fox	0.15	0.36	80
giraffe	0.88	0.33	80
hippo	0.88	0.33	83
panda	0.95	0.22	79
shark	0.85	0.36	80
sheep	0.94	0.24	81

spider	0.91	0.28	81
squid	0.81	0.45	81
turtle	0.83	0.38	81
whale	0.89	0.35	81

Note. Mean naming agreement represents the proportion of participants using the canonical name or genus variant to describe the image.

Appendix C

Table C1. Instances of responses other than canonical name by participant and item

subject	Response	item
S001	This image looks like a cat.	dog
S002	This looks like a pig	hippo
S002	This looks like a Fish	whale
S004	This is a chicken	bird_front
S005	It looks like a Cat	dog
S005	This looks like a moth or other similar winged insect.	butterfly
S006	This image looks like a cat.	dog
S007	it looks like a cat	dog
S008	This looks like an animal	elephant
S008	Looks like an animal	hippo
S008	Looks like a bug	ant
S009	The image looks like a cross between a pelican and a penguin.	bird_front
S010	It look like eleplant	elephant
S010	It look like a moth	butterfly

S010	it is a cat	dog
S015	Looks like an animal.	dog
S015	Looks like Cockroach	butterfly
S016	This looks like a black widow	spider
S017	This looks looks like a black and white moth	butterfly
S018	this is a graphically drawing of an insect	spider
S018	this is a drawing of a fish	whale
S018	this is a graphical drwaing of a black and white cat	dog
S019	this image looks liek a cat	dog
S019	this image looks like a fish	whale
S020	This is a penguin	bird_front
S021	This is a bug with 6 legs	ant
S021	This is a cat	dog
S023	This looks like a moth.	butterfly
S023	This looks like a crow.	bird_front
S023	This looks like a seagull.	bird_over
S023	This looks like a black widow.	spider

S024	It looks like an eagle crest.	bird_over
S024	It looks like a woodpecker.	bird_front
S026	It looks like a fox.	dog
S027	walrus	hippo
S028	Looks like a scorpion	spider
S030	image looks like a soider	ant
S032	looks like an eagle	bird_over
S033	It looks like an insect.	spider
S033	It looks like a decoration.	elephant
S033	It looks like a fish.	whale
S033	It looks like a cow.	dog
S033	It looks like a rabbit.	hippo
S034	looks like a wood pecker	bird_front
S035	this image looks like an Eagle	bird_over
S035	it looks like a shark	whale
S037	This image looks like a penguin	bird_front
S037	It looks like an eagle	bird_over

S039	This image looks like a buttefly	butterfly
S039	This image looks like a penguin	bird_front
S040	Insect	butterfly
S040	?? fish	whale
S043	This looks like a penguin.	bird_front
S043	This looks like an eagle	bird_over
S044	Its a black cat with a long face	dog
S044	it seems to be some sort of animal	hippo
S044	It looks like an insect	butterfly
S046	This looks like a bird	duck
S047	A square with triangles	duck
S047	An arch with a head on the left	camel
S047	A box with the bottom cut out, with an angular line to the right with a head	giraffe
S047	A rectangle with two feet a double diamond tail and a petagon face	fox
S047	A large white oval with two black rectangular feet and a black face	sheep
S047	Upside down pentagon spiked with a small square head all in black	turtle

S047	Black torpedo with two triangles one on top one underneath as fins, with a black triangular tail	shark
S049	This images looks like a cat.	fox
S050	Teddy bear	panda
S050	spider	turtle
S050	crocoach	squid
S050	dog	fox
S050	fish	shark
S050	hen	duck
S051	looks like a cat	fox
S052	I thin it supposed to be a cat walking sideways.	fox
S053	this looks like a cat	fox
S053	this looks like a llama	duck
S054	right facing profile of a cat	fox
S055	this look like an animal	camel
S055	this look like an animal	giraffe
S055	this looks like a fish	shark
S055	this looks like an animal	fox

S055	this looks like a bird	duck
S055	this looks like a fish	squid
S055	this loks like an animal	turtle
S055	this look like an animal	sheep
S055	this looknlike an animal	panda
S056	This is a picture of a cat	fox
S057	It looks like a cat	fox
S058	It loos like an animal made of rectangles and triangles	duck
S058	A very basic cat	fox
S059	looks like a dromedary	camel
S059	looks like a cat	fox
S060	It looks like a Girrafe.	giraffe
S060	It looks like a Tortise.	turtle
S060	It looks like a fish.	shark
S060	It looks like an image of an animal.	duck
S060	It looks like an insect.	squid
S060	It looks like a cat.	fox

S061	A pencil looking cat.	fox
S062	This is an image of a Jellyfish	squid
S062	This is an image of a cat	fox
S065	Looks like a girrafe.	giraffe
S066	Image looks like a cat	fox
S067	This looks like a cat	fox
S067	This image looks like a bird	duck
S069	it looks like a cat	fox
S070	It looks like a tortoise.	turtle
S070	It looks like a cat	fox
S070	It looks like a shorts.	shark
S071	This image looks to me like a cat	fox
S071	I think this is supposed to be a dog	duck
S072	Looks like a cat	fox
S072	Looks like a giant aquid	squid
S073	this looks like a cat	fox
S074	the image looks like a cat walking	fox

S075	The image is a cat.	fox
S076	It looks like a tortle.	turtle
S076	It looks like a fish.	shark
S076	It looks like a chiken	duck
S076	It looks like a cat.	fox
S076	It looks like Giratte	giraffe
S076	It looks like a prown.	squid
S077	It’s a cat with horns!	fox
S077	It’s a dog! Maybe!	duck
S078	a black camal	camel
S078	a four legged animal	fox
S080	Looks like a cat	fox
S081	The image is of a cat.	fox
S081	The image repesents an animal.	duck
S082	It looks like a walking cat.	fox
S083	This looks like a cat	fox
S084	Possibly a dog	duck

S084	Blacvk girraaf	giraffe
S084	Cat	fox
S085	This is a camle	camel
S085	This looks like a cat	fox
S086	this looks like a cat	fox
S086	this looks like a shrimp	squid
S087	This image looks like a cat	fox
S087	This looks like a camal	camel
S088	This image looks like a cat.	fox
S089	A large black rectangle, with a black and white triangle on the	fox
S090	it looks like a dog	hippo
S090	it’s looks like an ant	spider
S090	it looks like a cat	dog
S090	it looks like a parrot	bird_front
S090	it looks like a fish	whale
S091	This image looks like a cat.	dog
S092	It looks like an Eagle Crest	bird_over

S092	It looks like a woodpecker.	bird_front
S093	This image portrays a penguin	bird_front
S094	it looks like a bug	ant
S095	It looks like a pig	hippo
S095	It looks like a cat	dog
S096	this looks like a spide	ant
S097	This looks like a cat	dog
S097	This looks like some sort of insect	butterfly
S099	this image looks like a moth.	butterfly
S099	this image looks like a penguin.	bird_front
S099	this looks like a wale.	whale
S100	It looks like a penguin.	bird_front
S102	this image looks like a bug	ant
S102	this shape looks like an animal	hippo
S102	this shape looks like a penguin	bird_front
S103	sword fish	whale
S103	bat	butterfly

S104	It looks like an animal.	dog
S104	It looks like a penguin.	bird_front
S104	It looks like a moth.	butterfly
S105	Looks like a moth	butterfly
S105	Looks like a hawk	bird_over
S108	looks like an insect	butterfly
S108	looks like an animal	dog
S108	looks like an animal of some sort	hippo
S109	Looks like a buttefly	butterfly
S110	It looks like a potential penguin	bird_front
S112	moth	butterfly
S112	wasp	ant
S112	hawk	bird_over
S112	magpie	bird_front
S113	Possibly a fox	dog
S114	It looks like an emblem	elephant
S114	This image looks like a tortoise	hippo

S114	This image looks like an animal	dog
S114	The image looks like a bird	butterfly
S114	The image looks like a wing	bird_over
S115	looks like a fly	butterfly
S115	looks like a fish	whale
S116	It looks likes a penguin.	bird_front
S116	This appears to be an animal.	dog
S118	It looks like a moth	butterfly
S119	This image looks like a moth.	butterfly
S120	the image looks like a cat	dog
S120	the image looks like a sort of flying insect	butterfly
S121	Looks like a styalised bald eagle	bird_over
S121	Looks like a cat	dog
S122	This looks like an animal	dog
S122	This looks like a penguin	bird_front
S123	looks like a mammoth	elephant
S123	cat	dog

S124	it looks like a penguin	bird_front
S124	it looks like a cat	dog
S125	it looks like a tarantula	spider
S127	a small animal	dog
S128	moth	butterfly
S128	penguin	bird_front
S129	looks like mammoth	elephant
S130	THIS IMAGE LOOKS LIKE A DOG	fox
S130	THIS IMAGE LOOKS LIKE A LARGE TORTOISE	turtle
S130	THIS IMAGE LOOKS LIKE A CHICKEN	duck
S132	Kangaro	giraffe
S132	Ant	turtle
S132	Goat	fox
S132	Goat	sheep
S132	Ant	squid
S132	Dock	duck
S132	Aeroplane	shark

S133	This image looks like a turtoise	turtle
S133	This looks like a cat	fox
S134	It looks like a cat	fox
S134	It looks like a cow	sheep
S134	It looks like a dog	camel
S136	It looks like a shrimp	squid
S137	This image looks like an animal.	duck
S137	This image looks like an insect.	squid
S137	This image looks like an areoplant,	shark
S137	This image looks like a cat.	fox
S138	looks like a jellyfish	squid
S138	Like a cat	fox
S139	cat	fox
S139	crill?	squid
S140	looks like a goose	duck
S140	looks like a cat	fox
S140	looks like a girrafe	giraffe

S143	This looks like a cat	fox
S144	Jellyfish	squid
S144	Tortoise	turtle
S144	A carmel	camel
S144	Cat	fox
S145	This is a cat	fox
S146	it looks like a animal	duck
S146	looks like a fish	shark
S147	looks like a black silhouette of a boxy cat	fox
S148	looks like a bug	squid
S148	looks like an animal a cat	fox
S149	It looks like a cat.	fox
S150	looks like a dog	duck
S150	looks like a cat	fox
S151	The image loos like a tortoise	turtle
S151	looks like a chicken	duck
S151	looks like a cat	fox

S152	This image looks like a cat	fox
S153	it looks like a animal	giraffe
S153	it looks like a animal	turtle
S153	it looks like a animal	squid
S153	it looks like a animal	camel
S153	it look like a sea animal	shark
S153	looks like a large animal	panda
S153	looks like a small animal	fox
S153	looks like a wooly animal	sheep
S153	looks like a water bird	duck
S154	This image looks like a cat	fox
S155	This image looks like a camel.	giraffe
S155	This image looks like a cat.	fox
S156	this looks like a cat	fox
S157	It looks like a goose.	duck
S157	It looks like a cat.	fox
S158	looks like a bird	duck

S158	a cat	fox
S158	A fish	shark
S159	lools like chicken	duck
S160	This looks like a tortoise.	turtle
S160	The image looks like a cat.	fox
S160	The image looks like an insect..	squid
S161	Could be a cat or a dog	fox
S162	The image looks like a cat	fox
S163	cat	fox
S164	this looks like an sea creature	squid
S164	this looke like a dog	duck
S165	this image looks like a dog	duck
S165	this image looks like a cat	fox
S165	this image looks like a polar bear	panda
S165	this image looks like a fish	shark
S167	It seems as cat	fox
S167	It seems as Tortoise	turtle

S168	it looks like a cat	fox
S169	This looks like a cat	fox

Appendix D

Figure C1: Plots of random effects for items and participants for Analysis 1 (R model statement: wordCount ~ descripType + (descripType|subject) + (descripType|item)).

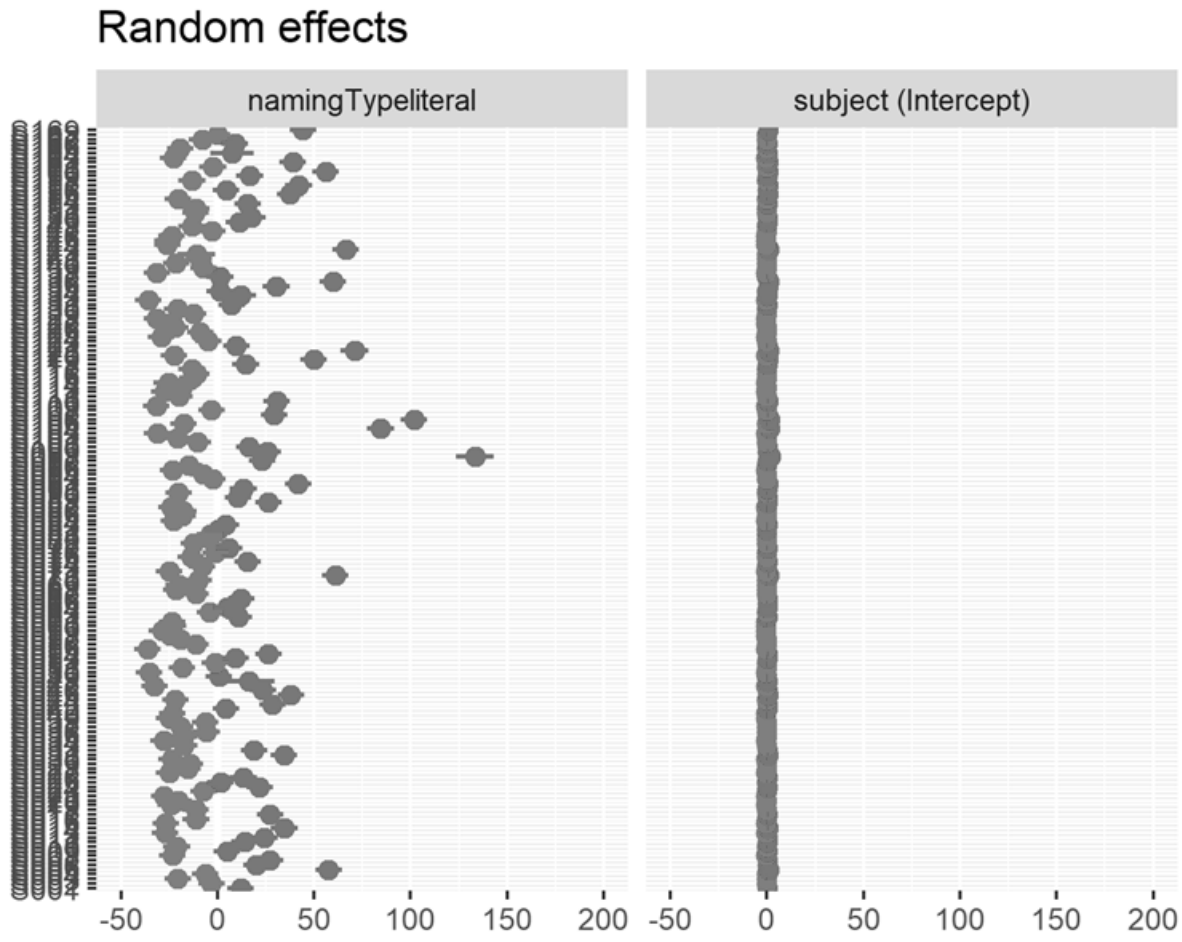
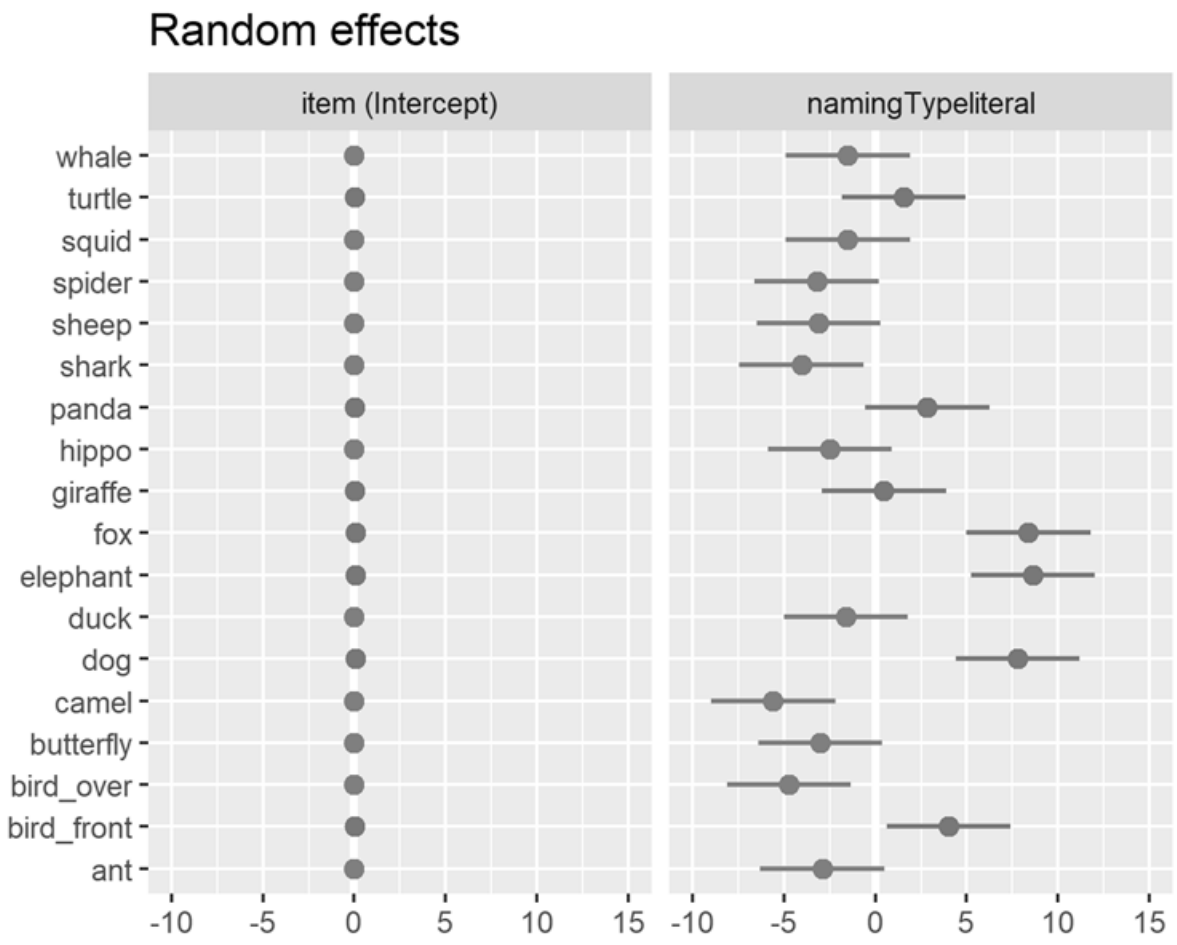


Figure C2: Plots of random effects for items and participants for Analysis 1 (R model statement: wordCount ~ descripType + (descripType|subject) + (descripType|item)).



Appendix E

Table E1: Literal word count t.test matrix

group1	group2	statistic	df	p	p.adj
bird_front	camel	3.457797	196.1669	0.000668	0.0139091
bird_front	duck	2.825662	198.1631	0.005000	0.0450000
bird_front	shark	3.211501	192.5505	0.002000	0.0235385
bird_front	sheep	3.167604	180.1046	0.002000	0.0235385
camel	dog	-3.953384	186.4682	0.000109	0.0086190
camel	elephant	-3.914252	174.6422	0.000130	0.0086190
camel	whale	-2.801301	224.4353	0.006000	0.0483158
dog	duck	3.355392	188.2739	0.000959	0.0139091
dog	giraffe	2.911563	216.2678	0.004000	0.0408000
dog	shark	3.722637	183.2404	0.000262	0.0086190
dog	sheep	3.687736	172.1198	0.000303	0.0086190
dog	squid	3.276915	195.3528	0.001000	0.0139091
dog	turtle	2.779646	215.3875	0.006000	0.0483158
duck	elephant	-3.364748	176.1783	0.000940	0.0139091

elephant	giraffe	2.964662	200.8329	0.003000	0.0327857
elephant	shark	3.700278	171.9329	0.000290	0.0086190
elephant	sheep	3.662202	162.5835	0.000338	0.0086190
elephant	squid	3.295350	182.2485	0.001000	0.0139091
elephant	turtle	2.841960	199.9429	0.005000	0.0450000

Appendix F

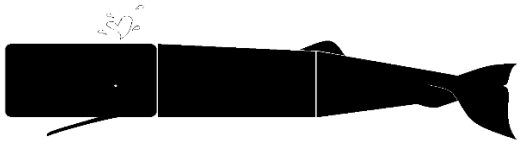


Image 1Whale

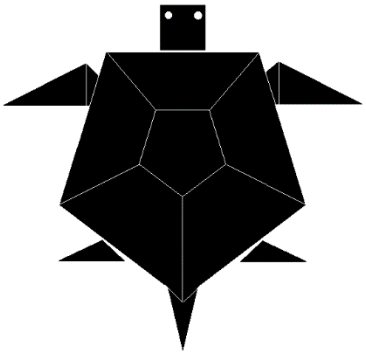


Image 2Turtle

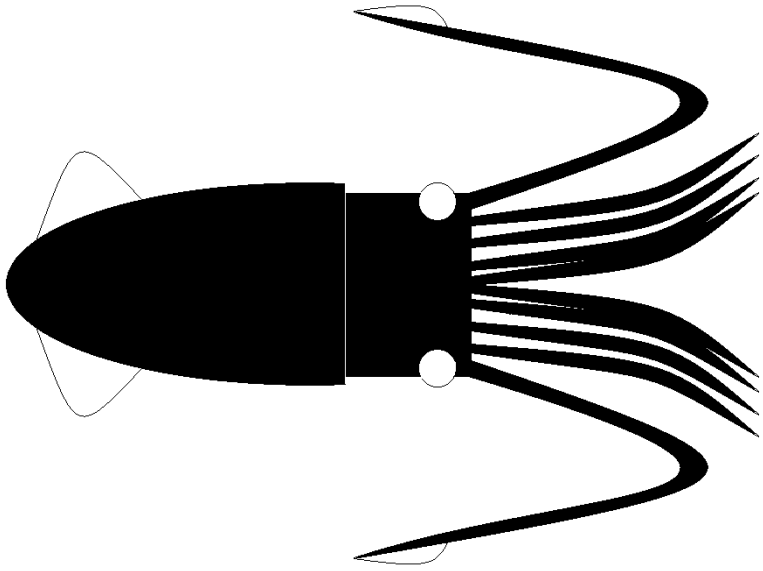


Image 3 Squid

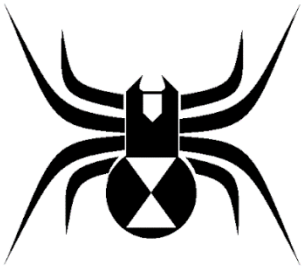


Image 4 Spider

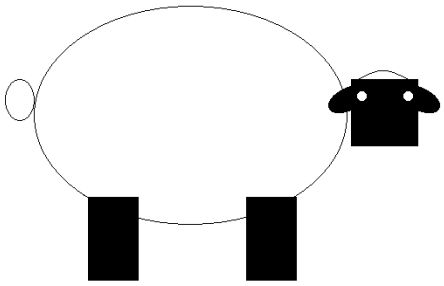


Image 5 Sheep

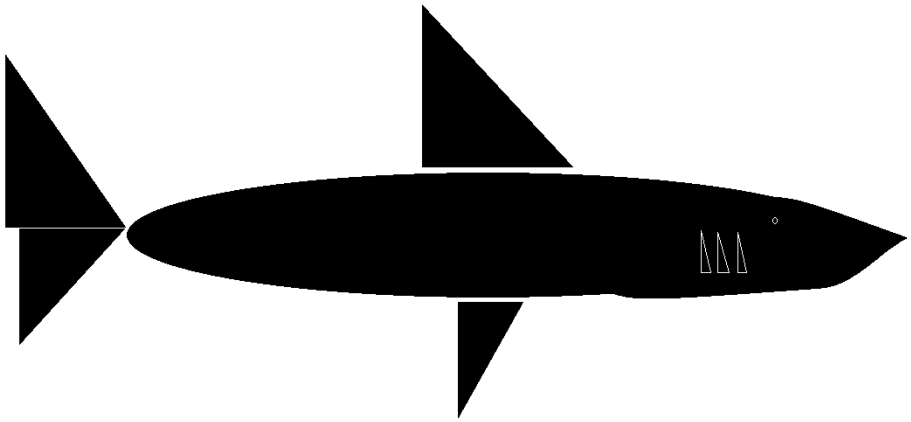


Image 6 Shark

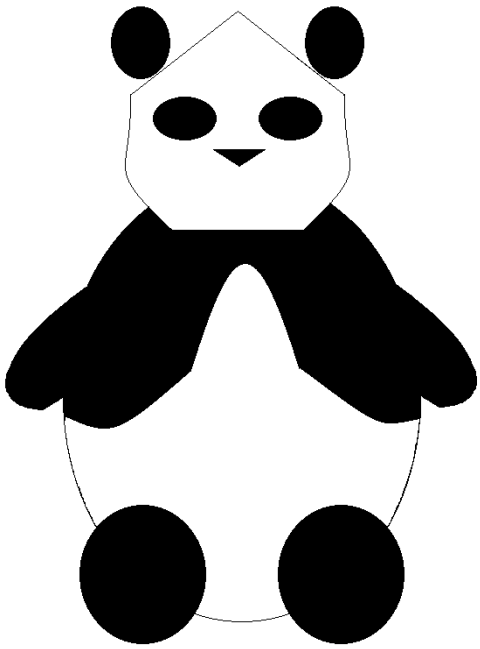


Image 7 Panda

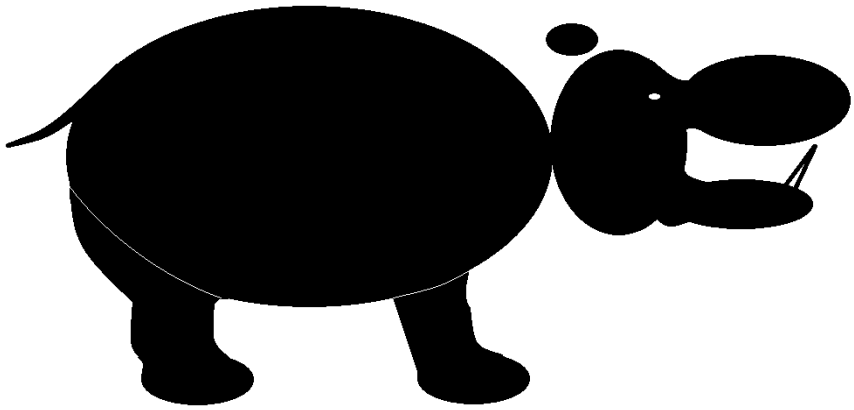


Image 8 Hippo

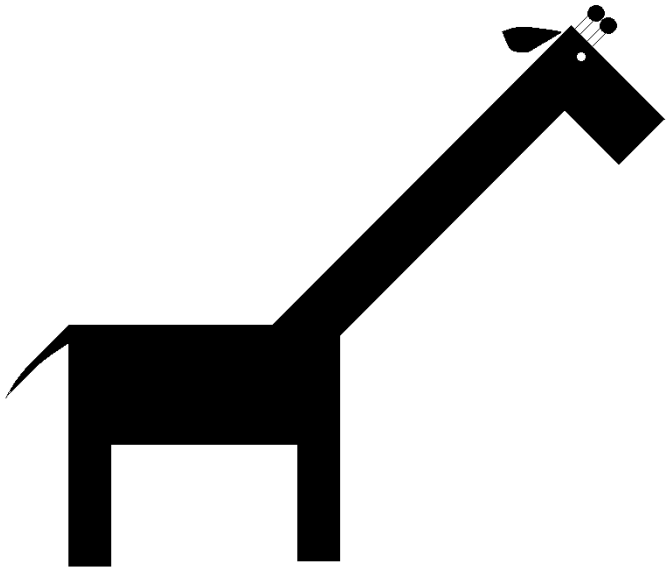


Image 9 Giraffe

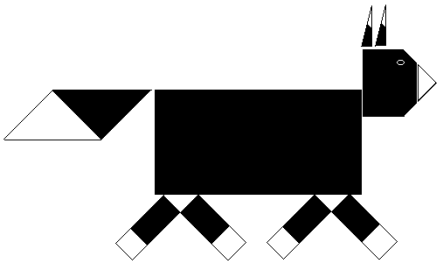


Image 10 Fox

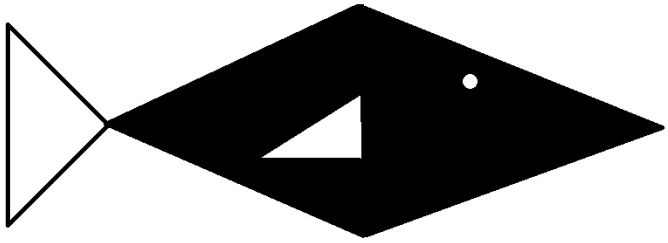


Image 11 Fish

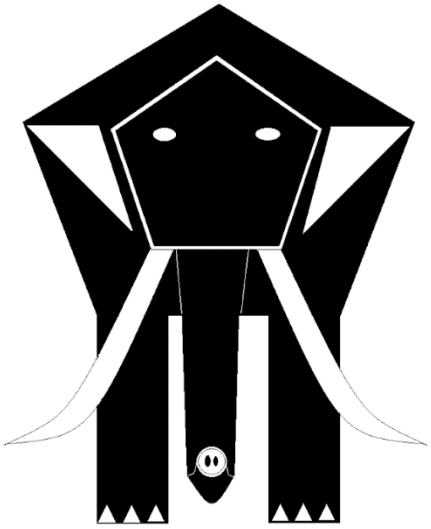


Image 12 Elephant

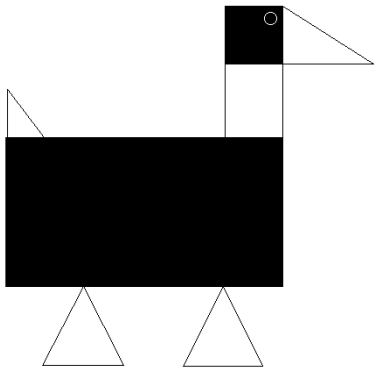


Image 13 Duck

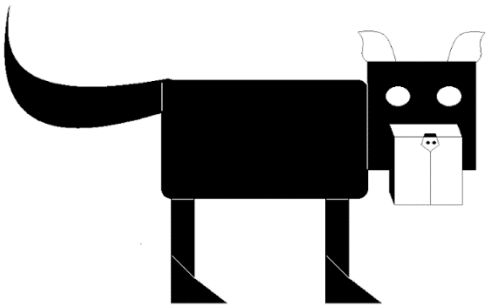


Image 14 Dog

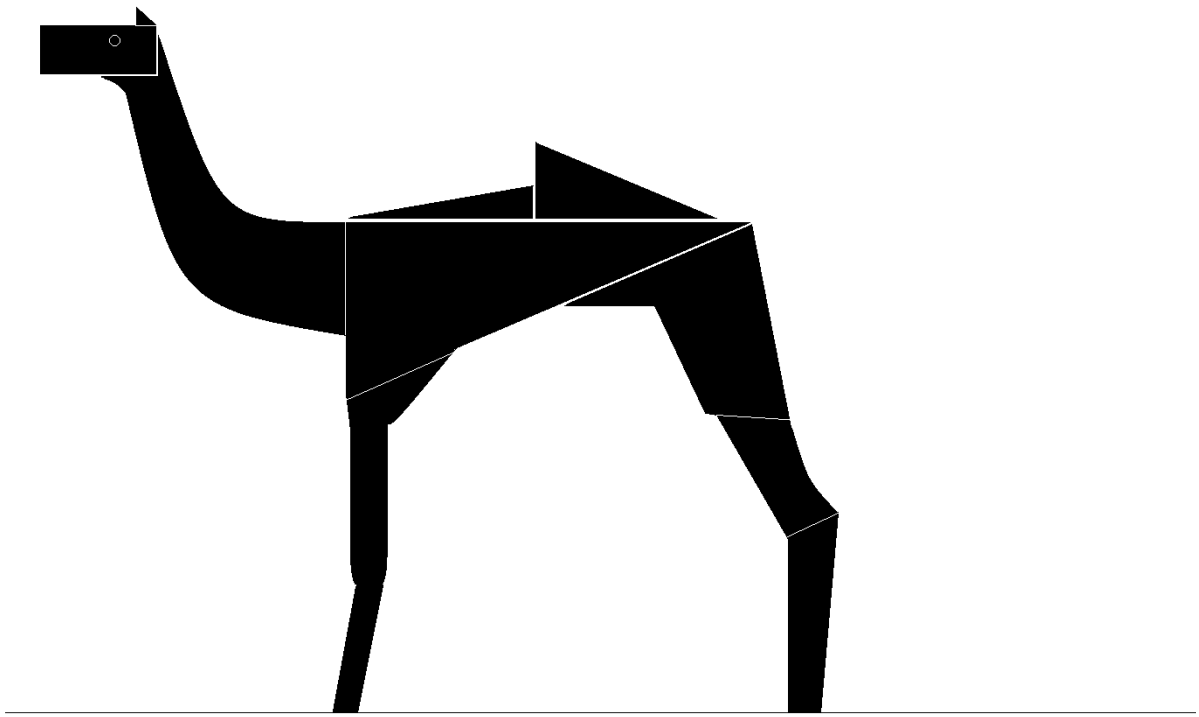


Image 15 Camel

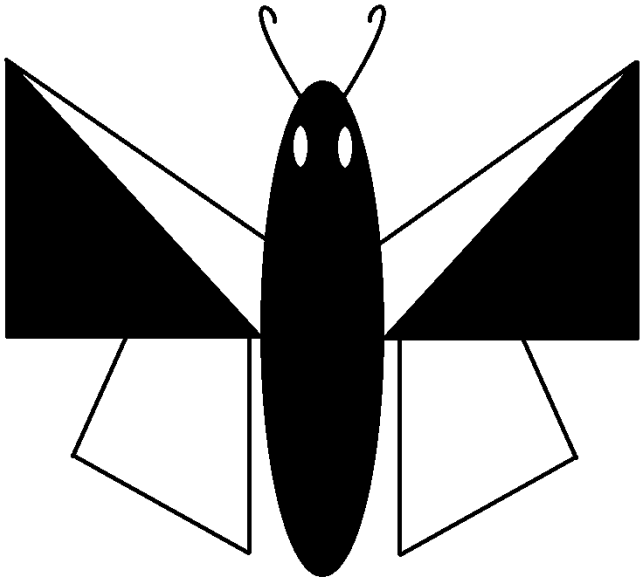


Image 16 Butterfly

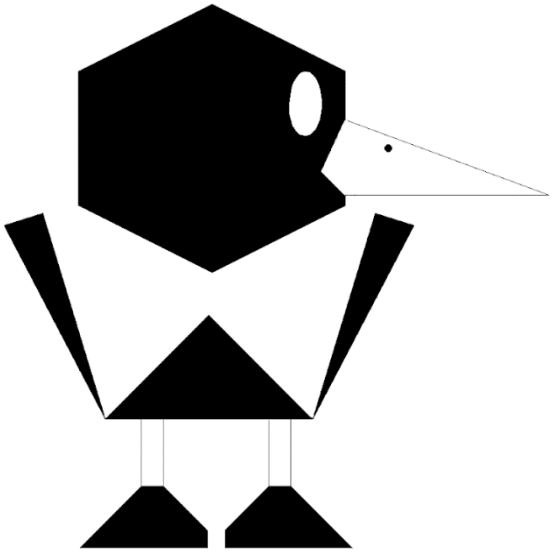


Image 17 Bird_front

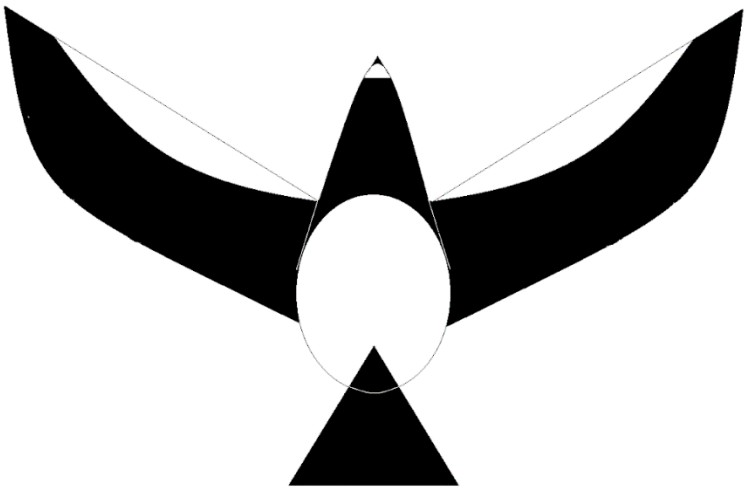


Image 18 Bird_over

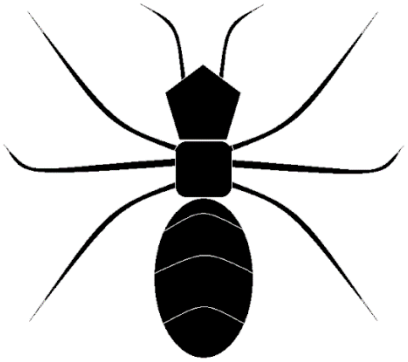


Image 19 Ant

Appendix G

Table G1: List of figurative and literal descriptions for visual stimuli used in studies 2 and 3

Figurative	**Literal**
It looks like an elephant.	There is a large black pentagon in the center with another smaller pentagon in the middle of it with two triangles on each side of it and two black rectangles below it with three small white triangles in each.
It looks like a butterfly	There is a large black oval shape in the middle with two smaller white ovals inside of it and two short curly lines above it. There are two black triangles on the left and right of the large black oval with two white triangles on top of them.
It looks like a fish	There is a large black diamond shape with a small white triangle in its center and a larger triangle attached to the left of the black diamond
It looks like a hippo	There is a large black horizontal oval in the middle with smaller black ovals attached to its right side, a small black line attached to its right side and two vertical shapes coming out of the bottom of the large oval. The ovals on the right have a small black oval floating over them and a small white triangle.
It looks like a dog	There is a large black rectangle in the middle with a thick curved line on its left, two vertical rectangles on the bottom of the large rectangle and a cube to the right of the large rectangle with two whole circles in it and two pointy curved triangle shapes on top of it
It looks like a camel	There is a large black triangle in the center of the image with two smaller black triangles on top of it and a series of long vertical shapes coming out of the bottom of the large triangle. On the left of the large central triangle there is a large thick curved line with a small black rectangle that has a small circle in it and a small triangle on top of it.
It looks like a spider	There is a black circle in the bottom center with two big white triangles inside and a pointy rectangle on top of the circle with a white pentagon shape inside of it. There are also four pointy curved lines on each side of the central shapes.

It looks like a panda	There is a large black and white oval with a pointy circular shape on top of it which has two circles and a triangle inside of it and two small ovals on top of it. The large oval also has to circles attached to it on the bottom and two thick curved lines on either side.
It looks like a bird	There is a large black Hexagon in the upper center of the image with a white circle inside of it and a big white triangle attached to the right of the hexagon. Under the large hexagon there are three black triangles surrounding a large white double triangle shape.
It looks like a bird	There is a white oval shape in the center of the image with a sharp black triangle attached to its bottom and a black curved triangular shape attached to its top that has a white tip. On either side of the central oval shape there are large black and white diagonal shapes with curved bottoms.
It looks like a cat	There is a large rectangle in the center of the image with four small black and white rectangles attached to its bottom. The central rectangle has two black and white triangles attached to its left and square like shape on its right that has a pointy white right end and two black and white triangles on top.
It looks like a squid	There is a large black half oval shape on the center left of the image with two small curved triangular shapes attached to it. On the right of the half oval there is a smaller black vertical rectangle that has many thin curved and pointy lines attached to it with two longer lines on the top and bottom of the rectangle. The rectangle also has two white circles in it.
It looks like a shark	There is a large horizontal black oval in the center with a pointy right end. The oval has two black triangles attached to the left side and two black triangles, one on the top and one on the bottom. The large oval also has three small triangles inside of it and one small circle.
It looks like a whale	There is a large horizontal black rectangular shape in the center of the image that is segmented. The left most segment has a small thin line protruding from the bottom, a small white circle inside of it and a set of white curved shapes on top. The rightmost segment has a small curved triangular shape on its top left and a fan shape protrusion on its right end.

It looks like a turtle	There is a large segmented pentagonal shape in the center of the image composed of five pieces that enclose a smaller pentagon in the center of the larger one. There are five small black triangles coming out of the sides and bottom of the larger pentagon, and there is a small black square on the top with two small white circles inside of it.
It looks like a giraffe	There is a horizontal black rectangle in the center of the image with a long diagonal rectangle attached to its right side that ends in a smaller rectangle with a small white circle inside of it and two dotted small white rectangles on its top. The central rectangle also has two smaller rectangles coming out of its bottom and a short thin pointy line protruding out of its left side.
It looks like a duck	There is a large black rectangle in the center of the image with two white triangles attached to its bottom. On top of the black rectangle there is a small white triangle on the left and a vertical white rectangle on the right with a black square on top of it. The black square has a small circle inside of it and a white triangle attached to its right.
It looks like a sheep	There is a large white circle in the center of the image with a small white circle on its left, two small black rectangles on its bottom and a small black square on its right. The small back square has two small circles inside of it and one white curved shape on top of it and two small black curved shapes on either side of the square.
It looks like an ant	There is a large black oval in the bottom center of the image with a small black square on top of it and a small black pentagon on top of the square with two thin curved lines on top of the pentagon. The black square has three thin curved lines protruding from its left and right.

Appendix H

Table H1: List of artificial interlocutor responses and phrases used in studies 2 and 3

Greetings, instructions and transitions	Responses for sorting actions as matcher	Instructions and responses for sorting actions as director
Hello!	Is that correct?	I will now describe the image that goes into Drop Zone A-F
You are the Director, and you will be describing the images for me to match.	Was that the right one?	Excellent!
Are you ready?	Was that correct?	Great Job!
Great! Please describe the first image!	Is that the right one?	Awesome!
Great! We will now proceed to the next screen where we will switch roles.	Excellent! Please describe the next image.	Well done!
Alright! I am the Director now and I will be describing the images for you to match.	Great! Please describe the next image.	Fantastic!
The next screen will be the final task of the study. Thank you!	Fantastic! Please describe the next image.	Correct!
	Great! Last one please!	

Calls for clarification and indication of misunderstanding	Phrases used for inadvertent errors made by the “artificial interlocutor”
Hmmm, I’m not quite sure which image you are referring to.	Sorry!
I’m sorry I don’t understand, could you try again please.	My mistake!
I can’t find a match for that description.	Let me try again.
Did you mean the one that looks like a/an	