Bootstrapping Vocal Communication Systems in Virtual Reality

iconicity and compositionality in vocalization

by
Kotryna Motiekaitytė
S1083700

MA degree programme in
Linguistics and Communication Sciences (research)

Nijmegen, The Netherlands, March 2024

Supervisor(s): dr. Limor Raviv
Assessor: prof. Asli Özyürek

Radboud University

Abstract

Understanding the origins of modern human communication has long been a goal for researchers investigating language evolution. A core questions within the field asks which modality – vocal or visual-gestural – played a larger role in the emergence of language, or were modality roles similar and human communication multimodal in its origins like it is today? Prior experimental research on this question shows an advantage of gestural communication, as its high affordance for iconicity (i.e., a strong connection between the form and the meaning of a signal) can aid the creation and learning of signal meanings, grounding the language. However, recent typological and experimental research indicates that human speech and vocalization can have more affordance for iconicity than previously thought. At the same time, iconicity can slow down the development of compositional structure – another core feature of modern human languages – as less reliance on grammatical structure is needed. Still, cognitive limitations have been shown to bias learners to create compositional structure, as re-combining pre-established signals is less cognitively demanding than creating new iconically motivated signals for an expanding meaning space. In my Research Master's thesis, I investigate this question further by focusing on the development of iconicity and compositionality in vocal communication systems created by human participants over repeated dyad interaction and expanding meaning space. I have collected and analyzed a dataset of vocal languages created by 8 pairs as a part of a larger project that explored the development and change of novel communication systems across different modality conditions (non-linguistic vocalization-only, gestural-only, or multimodal, i.e., an option to use both) and semantic features (stimuli shape, size, movement type, and speed). To accommodate both high experimental control and ecological validity, the project utilized an interactive Virtual Reality (VR) environment, where participants communicated face-to-face without computer interference. Additionally, to reduce participants' reliance on conventional signals, a novel set of 32 stimuli differing in the four semantic domains was created. The thesis results suggest that vocal communication systems have affordance for iconicity via modulations of acoustic features of vocal signals. Participants mapped the given semantic features of the stimuli by modulating the duration, pitch, loudness, speech rate, and number of syllables of their vocalizations. In addition, varying levels of compositional structure emerged across pairs' languages. Both compositional structure and convergence between participants on their language increased over time. Furthermore, increasing compositional structure resulted in improved communicative success. Together, these results show that dyadic communication can be enough to create vocal communication systems that are simultaneously iconic and compositional when the interaction takes place repeatedly and over an expanding meaning space.

Table of Contents

List of Figures iii
List of Tables iv
Chapter 1. Introduction 1
1.1. The mystery of language evolution 1
1.2. The multimodal origins of human language 2
1.3. Experimental approaches to language evolution 4
Chapter 2. The present thesis 9
Chapter 3. Methods 11
3.1. Participants 11
3.2. Materials 11
3.3. Procedure 11
3.4. Measurements and Predictions 12
3.5. Data preparation 15
3.6. Analyses 16
Chapter 4. Results 18
4.1. Communicative success 18
4.2. Iconicity 19
4.3. Compositional structure 25
4.4. Convergence 27
Chapter 5. Discussion 28
5.1. Communicative success 28
5.2. Iconicity 29
5.3. Compositional structure 31
Chapter 6. Conclusions 33
References 34
Appendices 39
A. Praat scripts 39
B. Models for statistical analysis 43
C. Total averages of compositional structure scores for all pairs 49

List of Figures

Figure 1. Stimuli and their four shape types. 11
Figure 2. Participants interacting in the CAVE. 12
Figure 3. Accuracy over time in all conditions over all pairs (6 pairs per condition). 18
Figure 4. Accuracy over time by pairs that used vocalizations. 19
Figure 5. Iconic mappings on vocalization duration by pair over time 21
Figure 6. Iconic mappings on F0 levels across pairs over time 22
Figure 7. Iconic mappings on vocalization intensity levels across pairs over time. 23
Figure 8. Iconic mappings on the number of syllables per vocalization across pairs and over time. 24
Figure 9. Iconic mapping of target speed on speech rate across pairs and over time. 24
Figure 10. Changes in compositional structure over time 25
Figure 11. Example of a highly compositional language (A) and less compositional language (B). 26
Figure 12. The correlation between compositional structure and communicative success across participants. 27

List of Tables

Table 1. Summary of iconicity results. Only significant two-way and three-way interactions are included in the table. 20
Table 2. Summary of compositional structure results. 25

Chapter 1. Introduction

1.1. The mystery of language evolution

How did modern human language come to be? How did the early stages of our ancestors' communication look like: was it spoken, gestured, or a combination of both? How did it change over time? These are all critical questions posed by scholars interested in human language origins and change, a field termed language evolution. As Dediu and de Boer (2016) note, the field of language evolution methodologically is still coming of age, even though people have long attempted to explain the origins of the uniqueness of human language through myths and philosophical postulations (see Fitch, 2010 for a historical overview). In fact, some of the first theories on the origins of language grew so speculative, that in the 19th century, the influential *Socie´te´ de Linguistique de Paris* denounced further research relating to this topic. However, this ban did not diminish scholars' interest in language evolution, and this can be noted in the growth of modern hypotheses about social, cognitive, and biological adaptations required for the emergence of language, or typological comparisons of existing languages in search of universal linguistic features. Notably, the main source of problems in the modern research of evolutionary linguistics relates to the inability to directly investigate the phenomena under study. It is impossible to study the minds of early hominins, nor can we observe the natural emergence of a language from scratch, as all existing languages are products of gradual and slow linguistic change. Crucially, human communication systems occupy the spoken and visual modalities, hence they do not leave archaeological traces. Some historic developments of natural languages can be recreated by comparing languages within and across linguistic families, as well as analyzing written records (Scott-Phillips & Kirby, 2010). Nevertheless, the archaic forms of modern languages can only be traced back to the development of their written systems, and with writing being a relatively recent development in human cultural evolution, the truly ancient forms of existing languages, let alone human communication in its origins, are forever lost. In turn, historically the field of language evolution was theoretical by nature, providing an open ground for far-reaching speculations.

Despite the lack of direct accessibility to the early forms of human communication, modern research attempts to provide empirical, data-driven evidence to the questions of language origins and change. Scholars research indirect naturalistic sources, such as pidgins in creoles – full natural languages created by mixing languages of different linguistic communities in unique socio-historic contexts (Botha, 2006). Meanwhile, work on emerging sign languages created by communities not exposed to any conventional sign system provides insights into how systematic communication develops in the visual modality (Senghas et al., 2004). However, as beneficial pidgins, creoles, and emerging sign languages are in offering cues about the natural formation of communicative systems, they might not represent the contexts in which human communication originated, they are not entirely novel linguistic systems, but rather combinations of existing languages. In the case of pidgins and creoles, they are invented by humans who already possess the experience of acquiring and using a language. In the case of emerging sign systems, they are created by individuals restrained in their use of the vocal modality. This way, emerging sign languages may not provide the full picture of language development, as the majority of the human population has access to and utilizes both vocal and gestural modes of communication. In addition, all naturalistic studies are constrained by difficulties in discerning and controlling causal variables of interest.

Hence, the past two decades have seen a surge of experimental methods making their way into language evolution research (Nölle, Hartmann, & Tinits, 2020a). Computational models of conversational agents have allowed us to simultaneously mimic the evolutionary contexts of language onset and test causal factors affecting further linguistic change (Kirby,

2002). Meanwhile, Communication Games where human participants are restricted from using a common language have become a robust experimental paradigm to test how people create novel symbol inventories from scratch, and how these inventories change over time and become efficient communication systems. It is worth noting, however, that neither naturalistic nor experimental studies are better than the other and sufficient alone. No contemporary experimental and computational model can cover all aspects of the origins and change of language (Scott-Phillips & Kirby, 2010). Research on communication systems in the wild is crucial to forming theories about linguistic evolution, which can be tested and refined with experimental approaches and vice versa. As Nölle and colleagues (2020a) note, a combination of interdisciplinary methodologies is needed to answer the complex and intricate phenomena that are language emergence and change. The following sub-section discusses the question in evolutionary linguistics specifically of interest to the present thesis.

1.2. The multimodal origins of human language

One of the core debates in the study of language evolution is the role of vocal and visual-gestural modalities in the emergence of human communication (Christiansen & Kirby, 2003). Historically, two opposing poles have been proposed regarding this issue, one putting forward speech and the other putting forward gesture as the primary modality in which language originated. The speech-first view refers to the centrality of the vocal domain in modern human communication and the astonishing diversity of spoken languages as indications that language was realized in the vocal modality right from its evolutionary roots (MacNeilage, 2010). The biological, cognitive, and sociocultural adaptations in humans that enable speech, along with the findings that children start exercising their vocal apparatus by babbling before learning to communicate, have been argued to support the speech-first view (Fitch, 2010). Further arguments for this theory are derived from comparative studies on non-human primate communication. Research indicates several shared properties between human languages and primate vocalizations: for instance, referentiality in primate alarm calls conveys semantic information about predator type and urgency of the threat (Cäsar, Zuberbühler., Young, & Byrne, 2013) and turn-taking exists in dyadic interaction between bonobos (Levréro, Touitou, Frédet, Nairaud, Guéry, & Lemasson, 2019). Moreover, auditory signals allow for communication when the visibility is low, e.g., in the dark or wooded areas, and when interlocutors are further away from each other. Regardless, the vocal behavior of the great apes and other non-human primates still appears to be confined to a fixed repertoire of calls and environmental, emotional, and other triggers, while gestural communication is prevalent among these closest relatives of humans. Therefore, the pitfall of the speech-first view lies in the tendency to consider human speech and vocalizations in isolation, neglecting the visual-gestural modality and overlooking a substantial part of data important for understanding the formation of communication systems.

On the opposing side, the gesture-first view puts forward the idea that human communication in its first stages was mostly visual and only later supplanted by vocal communication (Arbib, Liebal & Pika, 2008, Corballis, 2012, Tomasello, 2010). Support for this view comes from the evidence for the importance of gestures and other visual modality cues in human and non-human primate communication. Starting with the latter, research shows that vocalizations of non-human primates are too limited in their functionality to bootstrap language alone, while primate gestures offer higher flexibility in their use and learnability (Pollick & De Waal, 2007). In fact, primates were shown to adapt their gestures to the age and status of their interlocutors (Arbib et al., 2008). The adaptability and learnability of primate gestures make it well-suited for transfer across generations of users and thus suggest a similar role in the evolution of human communication.

For humans, gestures, especially ones accompanying speech, are used across linguistic backgrounds and cultures (Feyereisen & de Lannoy, 1991). Hearing children start gesturing from an early age, in some cases preceding the onset of spoken language acquisition (Iverson & Goldin-Meadow, 2005). Meanwhile, hearing children exposed to sign languages from birth use signs to babble (Petitto & Marentette, 1991), suggesting that babbling is not only a speech-based phenomenon related to the maturation of the vocal apparatus. Moreover, when deaf children lack a reliable sign language model in their environment, they readily create spontaneous sign inventories to communicate with their family members (Senghas, Kita, & Özyürek, 2004). Over time and with an increasing number of users, homesigns can become conventional linguistic systems (Hunsicker & Goldin-Meadow, 2012). This demonstrates that efficient communication can and does emerge in the visual-gestural modality alone. However, the theory's major shortcoming is the difficulty explaining why and how early human communication would switch from being mostly gestural to mostly spoken as seen today. That is, considering that gestural communication is prevalent in non-human primates and can develop into efficient and structured systems for non-hearing human communication, why do modern hearing humans use gestures to supplement spoken communication, not the other way around (Kendon, 2017)?

Taking note of the strengths and pitfalls of both speech-first and gesture-first theories, a mediating and more nuanced perspective over the past few decades suggests that the roots of human communication are multimodal by nature (Kendon, 2014; Nölle et al., 2020a). The core notion of the multimodal view of language origins comes from the increasing understanding that modern human communication is inherently multimodal and the deep integration of multimodal signals therein (Perniss, 2018). People across languages and cultures make use of a combination of various vocalizations, hand gestures, facial expressions, and eye behavior to aid information comprehension and production (Levinson & Holler, 2014). In fact, experimental research demonstrates that the removal of one modality hinders the processing of information in other modalities (Kelly et al., 2010). Other studies indicate that multimodality plays a facilitating role in language acquisition, as pre-vocal children can understand multimodal signals but struggle with signals that are presented in a single modality (Bohn et al., 2019).

Alongside observations on the multimodal character of modern human communication, evidence in favor of the multimodal view of language origins is drawn from archaeological, genetic, and comparative animal studies. Recent research indicates that modern non-human primates use a combination of vocal and visual signals to increase communicative efficiency instead of dominantly communicating in a single modality (Fröhlich, Sievers, Townsend, Gruber, & van Schaik, 2019; Slocombe, Waller, & Liebal, 2011). Additionally, the ability of non-human primates to learn the gestures of their interlocutors seems to extend to vocalizations somewhat more than previously thought – a behavior that is considered important for the development of a spoken language (for an overview, see Perlman, 2017). From the phylogenetic perspective, previous work indicates an intricate connection between the neurological apparatus that controls manual actions and the apparatus that controls actions used in speaking. This coordination between mouth and hand movements implies that our ancestors may have possessed pre-adaptations sufficient for multimodal communication deep in the hominin evolutionary line (Gentilucci & Corballis, 2006).

A major issue that remains open is how early human vocalizations turned into structured and relatively arbitrary linguistic systems, while the visual-gestural behavior of non-deaf humans remains largely iconic. Historically, spoken languages are thought to be fundamentally arbitrary, that is, opaque in the connection between the form and meaning of their signals (Pinker & Bloom, 1990). In contrast, sign languages and gestures are considered to have high affordance for iconicity (i.e., their form can intuitively resemble the expressed meaning),

primarily due to the visual nature of signs and gestures, allowing interlocutors to mimic the characteristics of referents observed in the real world (Cartmill, Beilock, & Goldin-Meadow, 2012; Perniss, Thompson, & Vigliocco, 2010). Thus, the visual-gestural modality has been postulated to play an important role in grounding the emergence of human communication: its high affordance for iconic representation of meanings makes it easier for learners to create and acquire new signals (Fay & Ellison, 2013).

However, recent studies demonstrate that speech and vocalizations can be more iconic than traditionally assumed (Perlman, Clark, & Johansson Falck, 2015a; Perlman, 2017; Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015). Across linguistic families, there is a widespread prevalence of ideophones – words such as onomatopoeias "bang" and "squish" in English – that convey sensory domains via close form-to-meaning mappings. Ideophones are shown to express a variety of semantic domains related to senses, including texture, luminance, manner of movement, shape, temperature, taste, and emotional and psychological characteristics (Dingemanse, 2012, for an overview). Additionally, multiple cross-linguistic studies demonstrate that some words outside of the ideophones class can also be iconic (Blasi, Wichmann, Hammarström, Stadler, & Christiansen, 2016). Here, iconicity is realized by encoding features of semantic domains via acoustic modulations in a speech signal (e.g., pitch and loudness). Similar to ideophones, cross-linguistic examples of iconicity in words outside the ideophones class are found in the domains of size (e.g., Haynie, Bowern, & LaPalombara, 2014), shape (e.g., Sidhu, Westbury, Hollis, & Pexman, 2021), manner of motion (e.g., Shintel & Nusbaum, 2007), texture (e.g., Winter, Sóskuthy, Perlman, & Dingemanse, 2022) and emotional/psychological characteristics (e.g., Aryani, Conrad, Schmidtke, & Jacobs, 2018; Kilpatrick, Ćwiek, Lewis, & Kawahara, 2023).

Therefore, as Perlman and colleagues (2018) argue, it is worth considering not whether speech signals can be iconic, but to what extent they are iconic. Research suggests that semantic domains are more readily available for iconic expression depending on the resources for iconic expression in the given modality (Dingemanse et al., 2015; Perlman & Cain, 2014). For instance, across sign languages, iconicity is high in signs signaling semantic domains such as spatial relations and visual shapes. Meanwhile, in spoken languages, iconicity is found in vocalizations referring to qualities of sound (Dingemanse et al., 2015). Qualities like "size", "repetition", "intensity", and "temporal unfolding" seem to have the potential for iconicity equally in both modalities. This suggests that vocalizations and gestures might have complemented each other early on, with iconic vocalizations used for certain meanings where gestures were less productive. Moreover, the awareness that labels can be used to refer to entities that are not the labels themselves may have enabled our ancestors to communicate about things that are not present in the communication act, i.e., the concept of displacement (Perniss & Vigliocco, 2014). In this way, iconicity would lay the ground for the emergence of modern human communication by the creation of labels that are motivated and easily understood within the community. The established vocal and gestural systems would gradually co-evolve to become more expressive and systematic, ultimately leading to the multimodal system used today (Levinson & Holler, 2014). However, without direct access to the languages of our ancestors, it is hard to get data-driven insights into how vocal communication evolved to become more arbitrary and systematic, while co-speech gestures remained for the most part iconic. The next sub-section will present previous experimental work that has investigated this question.

1.3. Experimental approaches to language evolution

As discussed above (see sub-section 1.1.), the field of language evolution has experienced a surge in empirical approaches in the past decades, now possessing a rich toolkit of experimental and computational methods. Among these, Communication Games are a robust paradigm that

has been used to study the specific questions related to the emergence and change of language. In this paradigm, pairs or groups of human participants complete communicative tasks, such as director-matcher game (e.g., Kirby, Tamariz, Cornish, & Smith, 2015; Motamedi, Schouwstra, Smith, Culbertson, & Kirby, 2019; Raviv, Meyer, & Lev-Ari, 2019a). Importantly, participants are deprived of their ability to use a common language and must invent a novel communication system, which, depending on the aims of a study, can be spoken, gestural, or graphical (drawings, written labels). This restricted behavioral context of communicative games allows researchers to tailor the causal factors of interest and target any wanted features of human languages to investigate their formation and change over time (Nölle & Spranger, 2022).

Early research making use of the Communication Game paradigm has investigated the cognitive underpinnings for bootstrapping communication, such as establishing common ground (e.g., Scott-Phillips, Kirby, & Ritchie, 2009), or planning and recognizing communicative intent (e.g., Noordzij, Newman-Norlund, De Ruiter, Hagoort, Levinson, & Toni, 2009). Subsequent studies have expanded the work by examining whether and how key features of natural languages emerge in artificial communication systems. Of interest to the present thesis are iconicity and compositionality. In terms of iconicity, it has been demonstrated that iconicity in both gestural and vocal signals helps to bootstrap communication. For instance, Fay and colleagues (2013, 2014) had pairs of participants communicate about a set of items depicting emotions, actions, and objects by using either gestures, non-linguistic vocalizations, or a combination of both modalities. The studies indicated a higher communicative success for the gestural compared to the vocal modality: gestures were either as effective as vocalizations (Fay et al., 2013) or more effective (Fay et al., 2014). The authors proposed several interpretations to the findings: first, the visual-gestural modality has more potential for creating iconically motivated signs, and second, iconic signals can aid the grounding of a communication system. Following this interpretation, the visual-gestural modality is argued to be more communicatively successful when interlocutors are deprived of a common language than the vocal modality.

It is worth stressing, however, that in both experiments, communication between participants in the vocalization-only condition was effective enough to denote the possibility of their success being only accidental. As Fay et al. note (2014), vocal signals increased communicative performance when they were iconic and shared a motivated relationship with the concept being described, e.g., making a yawning noise to communicate the emotion *tired*. This point was further supported by Perlman, Dale, and Lupyan (2015b), who showed that participants communicating in the vocalization-only condition did create vocal mappings that were iconically motivated, and that iconicity helped to successfully differentiate semantic antonyms.

A more recent study by Macuch Silva, Holler, Özyürek, and Roberts (2020) indicates that participants produce iconic vocalizations when referring to abstract, unconventional visual and audio stimuli (e.g., making splashing sounds for an image that resembles ink spots). The study reported an advantage of the modality of the stimuli on the accuracy of communication: for the auditory stimuli, accuracy was comparable across all three conditions (vocal-only, gesture-only, and multimodal), while for the visual stimuli, participants were less accurate in the vocal modality than in either of the other two. Moreover, the efficiency of communication (the time needed to convey a label to a partner) was the highest in the multimodal condition for the auditory stimuli and improved faster than in other conditions for the visual stimuli. However, the signals in the multimodal condition were often unimodal, with participants creating either gestural or vocal labels. Based on these findings, the authors made an inference that the efficiency advantage in the multimodal condition derived from interlocutors' ability to flexibly

use modalities to take advantage of the (iconicity) affordances of either visual or auditory stimuli and to meet the current communicative demands.

The development of iconic signals is also explored in Communication Game studies employing a different, graphical modality. Here, instead of communicating by non-linguistic vocalizations or gestures, participants have to produce drawings to refer to stimuli. These studies show that participants start with highly iconic drawings depicting detailed representations of target items. Over repeated use, these inventories of iconic drawings become conventionalized and symbolic systems when participants are allowed to give each other feedback (Fay, Garrod, Roberts, & Swoboda, 2010).

To sum up, previous experimental work on iconicity in novel communication systems indicates that gestural communication is often more successful than vocal or multimodal communication (Fay et al., 2014; 2013; Macuch Silva, et al., 2020), and that the higher affordance for iconically motivated signals in this modality in part attributes to the success. Nevertheless, some evidence for iconic expression in participants' vocalizations does emerge (Perlman et al., 2015b), especially when the stimuli are auditory and match the affordance capabilities of the vocal modality (Macuch Silva, et al., 2020). Further work is needed to investigate the iconic affordances of vocal systems created in the lab across different semantic domains to put together a fuller picture of iconicity in human vocalizations.

Another key aspect of natural languages that is often addressed in experimental research on language evolution is compositionality, i.e., the systematic recombination of smaller linguistic units to form larger units with novel meanings. Investigating the emergence of compositionality in artificial communication systems deals with the question of how inventories of iconically motivated and unstructured signals evolve into systems in which sub-parts of signals are systematically recombined to communicate about new referents, leading to the creation of compositional systems. As some studies indicate, iconicity may hinder or delay the emergence of compositionality, as less reliance on grammatical structure re-using established units is needed to convey new meanings (Dingemanse et al., 2015; Roberts, Lewandowski, & Galantucci, 2015; Verhoef, Kirby, & De Boer, 2016).

However, experimental work indicates that cognitive biases and memory limitations of learners are important in regulating the trade-off between iconicity in compositionality, as creating and learning new motivated signals for each meaning increases the cognitive load (Kirby, Griffiths, & Smith, 2014). Following this proposition, a discussion emerges of what communicative conditions are required for the learnability pressures to bootstrap compositional systems. Classic Iterated Learning studies argue for the importance of generational transmission for the development of compositionality: over generations, communication systems become more systematic and hence easier to learn because they have to be continuously acquired by new learners (Kirby, Cornish & Smith, 2008; Kirby, Tamariz, Cornish & Smith, 2015). For instance, in Kirby et al. (2015), generations of participants learning labels created by previous generations and producing new output that was used as a learning input for the following generations created compositional structure over time. Meanwhile, for closed groups communicating over rounds without generational shift, the signal inventories remained mostly unstructured, with different items acquiring unique labels.

Simultaneously, Communication Game studies indicate that generational transmission is not necessary for the creation of compositionality. It appears that compositional re-se of linguistic units can also emerge within a single generation if the interaction takes place between more than two participants (Raviv, Meyer, & Lev-Ari, 2019b) when stimuli are not present in the moment of communicative act (Nölle, Staib, Fusaroli, & Tylén, 2018), and when the meaning space of stimuli is continuously expanding (Nölle et al., 2018; Raviv et al., 2019b). Furthermore, Motamedi, Schouwstra, Smith, Culbertson, and Kirby (2019) indicate a

combined effect of repeated interaction and generational transmission in the increase of compositionality over time.

In general, experimental work on communication systems created in the lab indicates that interlocutors create compositional structures when cognitive pressures bias them against iconically motivated and unstructured signal inventories. Thus, compositionality can be seen as a fundamental feature of modern human communication: we can express an unlimited set of novel meanings with a limited number of established smaller linguistic units and be easily understood. However, the conditions facilitating the cognitive pressures needed for the emergence of compositionality remain elusive: some studies indicate the importance of introducing new learners to the system (Kirby et al., 2008; Kirby et al., 2015), while others show that in some cases repeated communication without generational transmission is enough (Raviv et al., 2019b; Nölle et al., 2018). Further comparison of the development of compositionality across modalities and semantic domains is important for a more comprehensive understanding of this question.

Taken together, experimental work on language evolution illustrates that both iconicity and compositionality are important aspects in the grounding and establishment of novel communication systems. Iconic signals may have a significant role in bootstrapping communication, while pressures introduced by repeated communication and generational transmission can prompt individuals to re-composition the established labels to express new meanings.

Nevertheless, the discussed experimental studies have several limitations concerning ecological validity that make them not necessarily respective to the contexts in which human communication originated. First, in most of these experimental paradigms which are typically administered via a computer interface, human communication evolved during face-to-face interaction (Levinson, 1983) and is embedded in a dynamic, 3D world. Second, typical everyday communication is multimodal, with auditory and visual signals interacting in a close and complex manner (Perniss, 2018). Third, many of the experiments discussed above employ tasks that lack realistic communicative motivation and context. That is, classic experimental approaches typically involve participants communicating about 2D referents presented on computer screens (Nölle et al., 2018; Raviv et al., 2019a; Raviv et al., 2019b; Macuch Silva et al., 2020) and in a single modality (Fay et al., 2010; Perlman et al., 2015; Nölle et al., 2018). Undoubtedly, high experimental control is valuable as it allows us to tap into the relationships between phenomena under study and factors that may influence them. However, too much experimental control strips studies of the natural communicative environment resembling that present in the early stages of communication. Ideally, researchers would seek to combine high experimental control with high ecological validity in methods that reflect the intricate and complex phenomena that is the language evolution (Peeters, 2019).

Another common limitation in experimental research on language evolution is noted by Macuch Silva et al. (2020). Many experiments include referents that are familiar to participants, such as the act of sleeping (Fay et al., 2014) or the quality ‘ugly’ (Perlman et al., 2015)). This presents a bias: familiar concepts can be associated with conventional signals (e.g., the gesture of laying the head on hands and closing the eyes to denote the meaning ‘sleep’) and, in turn, affect the development of the system. Macuch Silva and colleagues addressed this observation by introducing a set of visual and audio stimuli that represent abstract shapes and ambiguous sounds. However, abstract concepts might not resemble “classic” communicative situations between individuals without any shared symbol system. Specifically, conveying information about abstract concepts or items may be less important/crucial than communicating about contexts ensuring survival, such as warning about different predators or describing the location of resources. In fact, these specific situations were shown to elicit multimodal labels among non-human primates, likely to strengthen the robustness of the signals (Ratcliffe & Nydam,

2008; Rigaill, Higham, Lee, Blin, & Garcia, 2013). It is therefore fair to assume that predators, mating, food, and other domains crucial for survival were also essential in the early stages of establishing referential conventions in humans, with abstract items following after. Thus, novel stimuli that address these contexts and do not invoke conceptualized signals in any modality may be more representative of communication emergence.

To summarize, two important and interchangeably related questions emerge from the experimental work regarding the creation of novel communication systems that are fundamentally important to the overall field of language evolution. First is the question of affordance for iconicity between modalities when the stimuli are non-conventionalized and equally matched for iconic expressions in both gesturing and vocalizing. Second, is the relation between iconicity and compositionality across different modalities: how iconicity may hinder the emergence of compositionality, and in what conditions repeated dyadic interaction presents enough learning pressures for the creation of compositional structure without generational transmission. In addition, to better touch upon these questions, what is needed are methods that combine high ecological validity with high experimental control, as well as a novel set of stimuli that is both commutatively motivated and non-conceptualized across modalities. The current thesis shifts its focus to the vocal modality and investigates the emergence of iconicity and compositionality in vocal communication systems created by dyads communicating about different semantic features in a novel and interactive Virtual Reality environment.

Chapter 2. The present thesis

The study described herein makes use of the Communication Game paradigm (discussed in section 1.3.) to investigate the advantages or disadvantages of vocal and gestural signaling in bootstrapping communication emergence and learning, as well as the tradeoff between iconicity and compositionality. To examine this, dyads of participants played a producer-guesser game in one of three different modality conditions: novel vocalization-only, gesture-only, or both vocal and gestural (i.e., multimodal). Since comparing the vocal and gestural modalities is unfortunately beyond the scope of the present thesis[1], here I focus on analyzing the vocalization and combined conditions in depth. Specifically, the current thesis examines the evolution of vocal labels in the newly emerging communication systems with the following research questions in mind:

RQ1: To what extent communication is successful when participants are not allowed to use their shared spoken languages and gestures, and may only communicate using novel vocalizations (vocalization-only condition), novel gestures (gesture-only condition), or vocalizations and gestures together (multimodal condition)? How does communication success change over time?

RQ2: To what extent different semantic features of the stimuli (such as size or shape) are iconically expressed in participants' vocalizations through modulations in selected acoustic measurements?

RQ3: To what extent does compositional structure emerge in vocal communication systems created by participants repeatedly interacting in dyads? How does compositionality change (increase) over time? Additionally, to what extent do interlocutors converge on their systems?

The study was designed in a way that mimics the communicative contexts of early language emergence and learning. Dyads played the communication game face-to-face without any object obstructing their view of each other. Furthermore, the study made use of a novel Virtual Reality (VR) paradigm. The paradigm implemented in the current study is further discussed in section 3.3. Procedure. Notably, prior research demonstrates that 3D and Virtual Reality (VR) environments are robust tools for investigating linguistic behavior (Peeters, 2019). Such environments are shown to be more immersive and naturalistic compared to typical communicative tasks employed in experimental settings like tabletop games (Nölle et al., 2020a; Nölle, Kirby, Culbertson, & Smith, 2020b). Simultaneously, they allow researchers to manipulate variables of interest that would otherwise be hard or nearly impossible to utilize, such as the complexity of communication or changing environmental conditions (Peeters, 2019). Therefore, by employing the VR paradigm studies can opt for both high ecological validity and tight experimental control. The current experiment immersed dyads in a 3D forest. The environment was interactive, meaning that participants were surrounded by the 3D forest and could select stimuli by pointing a laser pointer.

Another way the study simulated naturalistic communicative situations is by introducing stimuli that simulated prey or predators Specifically, a set of fantasy creatures was designed for the experiment, containing different creature shapes, in different sizes, moving in different

[1] This is a multi-authored project, seeing as analyzing each condition requires extensive annotation, manual coding, and specific hands-on experience using the relevant tools (Praat). Since my background is in acoustic analysis, my contribution to this project was focused on the vocalization condition. The task of analyzing the gestural condition was assigned to another student with a relevant background in gesture analysis using automated motion detection, with the end goal of combining our insights into one comprehensive journal submission in the future.

ways and at different speeds (see Section 3.2. Materials). This resulted in stimuli that were closer to the contexts of early communication. Furthermore, the stimuli were novel and unconventional, meaning they were unlikely to elicit vocal and gestural signals that are conventional cross-culturally.

The presence of iconic mappings in vocal signals produced by participants in the vocalization-only and combined conditions was examined through several prosodic elements: pitch (measured as fundamental frequency), loudness (measured as intensity), duration, harmonics-to-noise ratio, number of syllables per vocalization, and speech rate. The compositionality of vocalizations was measured by comparing the pair-wise correlation between string distances (based on annotated forms) and corresponding semantic distances of produced labels. Finally, communicative success was measured as the accuracy of participants' guesses. Descriptions of all measurements and the motivation for their inclusion are presented in section 3.4. Measurements.

Regarding RQ1, the thesis predicts that the accuracy of participants' guesses will increase over time, with novel gestural communication systems being more successful than vocalization-only, or multimodal communication. Relating to RQ2, a prediction is made that different semantic features of the stimuli, such as shape and size, will be iconically mapped through modulations in acoustic measurements of participants' vocalizations based on the iconic expressivity of those elements. Finally, for RQ3 it is hypothesized that varying levels of compositional structure will develop in participants' vocalization systems, with compositionality increasing over repeated interaction and expanding meaning space. Additionally, interlocutors were expected to increase in their convergence over time, i.e., increasingly use similar labels for the same items. A detailed overview of predictions for each acoustic measure is presented in section 3.4. Measurements and predictions. Furthermore, some established labels will be re-combined to refer to shared semantic features between stimuli, turning them into syllabic structures and grounding the emergence of compositionality.

Chapter 3. Methods

3.1. Participants

A total of 36 native Dutch speakers (M_{Age} = 22, SD = 4; 29 women) were recruited for the study, resulting in 6 pairs for each of the three experimental conditions. Of these, 12 (M_{Age} = 22, SD = 3; 11 women) participated in the vocalization-only condition, and another 12 (M_{Age} = 23, SD = 5; 11 women) in the combined condition. All participants had normal-to-corrected vision and had no hearing problems, dyslexia, or language development disorders. Participants were tested in pairs and did not know each other before the experiment. One pair in the combined condition was excluded from the analysis due to incomplete audio recordings. All participants signed informed consent in compliance with the local ethics committee and were paid €20 for their attendance. The Ethics was given by the Social Sciences Ethics Committee at Radboud (Ref: ECSW-LT-2022-5-25-43541).

3.2. Materials

During the experiment, participants were immersed in the Cave Automatic Virtual Environment (CAVE). The stimuli were projected against a 3D forest background. To increase the immersive effect, the lights in the experiment room were dimmed to 20 Lux. Participants could interact with the environment (i.e., select target items) by pointing with a laser pointer. Video data was recorded with the Marshall CV346 cameras. Audio recordings were made with the RØDE Wireless GO II wireless microphones. The experiment was programmed with the Unity software.

The stimuli consisted of 32 novel animations depicting fantasy creatures. The creatures were characterized by four semantic features: shape (4 different types, see Figure 1), size (large versus small), movement type (walking versus hopping), and movement speed (fast versus slow). This allowed us to match the stimuli in a way that all features could be equally expressed via both vocal and gestural signaling. Each participant saw all possible combinations of features twice, thus leading to an overall set of 128 stimuli. Across participant pairs, the succession of target items was fully randomized.

Figure 1. Stimuli and their four shape types.

3.3. Procedure

Pairs of participants were seated across each other in the CAVE in such a way that they could not see the screen behind them but faced the screen behind the other participant (see Figure 2). In the vocalization-only condition, the participants were instructed to create a ‘fantasy language’ to communicate about unfamiliar creatures encountered in a VR forest. They were not allowed to use any existing spoken languages or names and were not allowed to gesture. In the gesture condition, only gesturing was allowed, and in the combined condition communication was possible in all modalities, but the use of spoken languages and names was forbidden.

Figure 2. Participants interacting in the CAVE.

Before the experiment, a trial creature that did not share any features with the experimental stimuli was presented in the middle screen so that it would be visible to both participants. Depending on the condition, pairs were asked to create a non-linguistic vocalization, a gesture, or any of the latter to refer to the creature. They were encouraged to come up with the signal themselves and were not given any examples by the experimenter to avoid prompting. In case participants violated the rules of the game (e.g., by describing the creature using Dutch or English words or using a modality that was not allowed), the experimenter corrected them and gave them another attempt, until participants understood what was expected from them.

After the trial, the director-matcher game would start following such order: first, the producer sees a target creature on the screen behind the guesser and creates a vocalization, gesture, or any of these based on the experimental condition. Once the label is produced, the matcher selects a creature they think best matches their partner's label from an array of eight creatures (the target creature + seven distractors) displayed behind the producer. Distractors were selected such that at least one creature always was the same shape as the target, and one shared the same feature (e.g., small size) as the target. After the guesser selects their creature using a laser pointer, both participants receive feedback: both the selected and target items move to the middle screen so that both participants can compare them. This was accompanied by a visual cue (green ✓ or red X) and an acoustic cue (high-tone beep for correct and low-tone beep incorrect choice) to explicitly indicate whether the interaction was successful or not. After each trial, the participants would switch roles, so the producer became the matcher and so forth. This way, both participants took turns producing and guessing labels an equal number of times (64 productions per participant). In total, the experiment consisted of 128 trials equally divided into two blocks. At all times during the experiment, the examiner was sitting in another room and could see and hear the participants. Video, audio, and accuracy data were recorded, alongside all trial information (e.g., which distractors were presented, their location on the screen, RTs, etc.). At the end of the experiment, participants were asked to fill in debriefing forms with questions about the language they created.

3.4. Measurements and Predictions

3.4.1. Communicative success

Communicative success was measured as the accuracy of guesses between participants. The success of each communicative turn was coded as 1 (correct) if the guesser picked the target item, or 0 (incorrect) if the guesser picked a different item. Following previous studies looking at the accuracy of participants developing communication systems across different modalities (Fay et al., 2014; Macuch Silva et al., 2020), I predicted that gestural communication would be

more successful than vocal or multimodal communication. Moreover, I expected the accuracy of participants' guesses to improve over repeated interaction.

3.4.2. Pitch

Pitch is a perceived speech feature measured by fundamental frequency (F0), which indicates the frequency of vibration of the vocal cords during a speech signal, which is then perceived as high or low sound, i.e., high or low pitch (Hirst & de Looze, 2021). F0 is measured in Hertz (Hz). In studies investigating links between acoustic measurements and semantic meanings, F0 has been most widely linked to iconic expressions of semantic domains relating to size. Across linguistic families, there is an observed relationship between the use of low vowels (e.g., *a* and *o*) and meanings of largeness, and high vowels (e.g., *i*) and meanings of smallness (Blasi et al., 2016; Ohala, Hinton, & Nichols, 1997). Importantly, vowel height, which is differentiated based on the rise of the tongue, also correlates with F0 levels, that is, low vowels have low F0, while high vowels have high F0 (Kawahara, 2021).

The linkage between F0 and size domain also extends to experimental findings on vocalization production (e.g., Perlman et al., 2015 for vocal reading tasks; Nygaard, Herold, & Namy, 2009 for production of nonce words) and assessment of utterances (e.g., Perlman et al., 2022 for assessments of non-linguistic vocalizations; Pisanski & Rendall, 2011 for assessments of speech relating to speaker's body size). This association seems to form at a young age, as preschool children successfully match high-pitched sounds to smaller items and low-pitched sounds to larger items (Mondloch & Maurer, 2004). The F0-size relationship has also been explored in marketing research, which demonstrates that consumers infer the size of the product from pitch levels in advertisement speech (Lowe & Haws, 2017).

Apart from the size domain, F0 has been linked to the meaning of movement and directionality. Experimental work demonstrates that participants increase or decrease their pitch as they describe an upward or downward-moving item, respectively (Shintel, Nusbaum, & Okrent, 2006). Similarly, participants can match nonce words varying in pitch levels to stimuli movements with either downward or upward trajectories (Ekström, Nirme, & Gärdenfors, 2022).

Taking together the discussed studies on sound symbolism in modulations of F0, the present thesis predicted that participants would produce higher-pitched vocalizations for smaller creatures than for large creatures. Additionally, participants could use modulations in F0 to map different movements of the stimuli, with jumping creatures eliciting higher pitch than walking creatures.

3.4.3. Loudness

The perceived loudness is measured as intensity, with loud sounds corresponding to high-intensity levels and vice versa. Intensity is measured in decibels (dB). In terms of iconicity in the spoken modality, intensity has been linked to the size domain. Higher intensity is demonstrated to characterize meanings of largeness, while lower intensity implies meanings of smallness (Perlman et al., 2022; Nygaard et al., 2009, however, not in Parise & Pavani, 2011). Additionally, vocalization loudness is found to change according to the referent's shape, with rounder objects eliciting higher intensity than sharper objects (Parise & Pavani, 2011).

For the current study, it was hypothesized that participants would produce vocalizations with higher intensity levels when referring to large compared to small creatures. Participants may also use higher intensity for round creatures (such as the rhino- and dyno-like shapes) and low intensity for the sharp-shaped creatures (spider-and dragonfly-like shapes); however, as to the author's knowledge, there is only one study demonstrating a link between stimuli shape and vocalization intensity (Parise & Pavani, 2011) the prediction is not strongly supported.

3.4.4. Duration

Duration refers to the length of time the given speech signal is produced. Prior studies point to a linkage between the duration of vocalizations and meanings of size. That is, longer durations

of the speech signal indicate larger items and vice versa (Nygaard et al., 2009; Perlman et al., 2022). Moreover, research reports a relationship between vocalization duration and movement speed: fast events elicit short speech times while slow events relate to longer speech times (Perlman et al., 2015). Thus, the present thesis predicted that participants would use modulations of vocalization duration to refer to creature size, with larger creatures eliciting longer vocalizations than small creatures. Dyads may also use duration for mapping stimuli speed, but the prediction is not strongly supported.

3.4.5. Harmonics-to-noise ratio

Harmonics-to-noise ratio (HNR), or harmonicity, measures the ratio between the periodic – typical – element of the speech signal and the aperiodic – noise – element (Teixeira & Fernandes, 2014). Thus, HNR provides an estimate of the overall periodicity of the sound expressed in dBs. As the noise ratio increases, the utterances become hoarser and breathier, and the HNR value decreases. Not many works have directly investigated the relationship between HNR and semantic meanings, but it has been implemented as a constituent element of other variables. For instance, HNR, among other acoustic measurements, has been used to measure vocal roughness as the iconic expression of item shape (Lacey, Jamal, List, McCormick, Sathian, & Nygaard, 2020). It has been found that as vocal roughness, and in turn noise ratio in HNR, increases, ratings of pseudowords increasingly shift to pointed shapes as opposed to rounded shapes. Therefore, seeing as some of our stimuli were round-shaped, while others were more sharp-shaped, HNR was included in the present analysis to further explore the measurement's linkage with a referent's shape; however, no strong predictions were made.

3.4.6. Number of syllables

Here, the number of syllables refers to the total number of syllabic units per vocalization. To the author's knowledge, no previous work had directly investigated the possible iconic mappings through modulations in syllable number in existing or artificial languages, but several studies have looked into reduplication in natural languages – the repetition of a word or its part to convey various semantic and grammatical functions. It is thus intuitive that reduplication is found in nouns to express the meanings indicating plurality, verbs to indicate repetition, and adjectives and adverbs to indicate intensity, more than one degree of a trait (for overview: Blake, 2017; Dingemanse et al., 2015). Following these findings, we predicted that participants would use re-duplication to refer to the two meanings signifying intensity: stimuli size and speed. Consequently, this would increase the total number of syllables, with large or fast creatures eliciting more syllables than small or slow creatures.

3.4.7. Speech rate

Speech rate refers to the speed at which speech units are produced at a given time frame. Here, speech rate is measured as the number of syllables per second. Previous studies mostly point to the possibility of modulating speech rate to express meanings associated with speed. Experiments, where speakers are tasked with reading or narrating stories, report that participants speak faster across descriptions of fast events compared to descriptions of slow events (Taremaa, Kiik, Toots, & Veismann, 2022; Perlman et al., 2015). Additionally, speech rate is distinctly modulated when adverbial phrases about speed, such as "really fast" or "very slow," are introduced (Perlman et al., 2015). Therefore, for the current study, speech rate was predicted to increase in vocalizations for fast stimuli and decrease in vocalizations for slow stimuli.

3.4.7. Compositionality

Here, the compositional structure score indicates the extent to which participants systematically re-used similar parts of strings to express the same semantic feature of the stimuli. Following Raviv et al. (2019a; 2019b), the compositional structure was measured as raw correlations between string distances and semantic distances between the labels in participants' languages. Semantic differences were calculated in RStudio (R Core Team, 2016) as follows: target items

that differed in shape scored 1 in difference score, and items that were of the same shape scored 0. All other semantic features (size, motion type, speed) were considered to be less perceptually prominent than creature shapes (types), therefore, they were given a score of 0.5 if the feature differed and 0 if it was the same between given items. Then, difference scores for all features per item were added, resulting in semantic distances that ranged from 0.5 (one shared feature) to 2.5 (all features are shared). Meanwhile, string distances were calculated as normalized Levenshtein distances between all possible string pairs produced by a given participant in a given block. Finally, the correlations of string distances and semantic distances were measured with Pearson product-moment correlation. The final compositionality score could range between 0 and 1, with 0 representing no compositionality and 1 indicating high compositionality in a language. In addition to raw correlations, z-scores for compositional structure were calculated as well. Performing the analysis with either raw scores or z-scores did not change the significance and direction of the results (see compositionality results in section 4.3.).

Different levels of compositional structure were predicted to develop in participants' languages, with some pairs developing more compositional systems than others. Moreover, compositionality was expected to increase over time, that is, over repeated interaction and expanding meaning space, as these conditions were shown to push the development of compositional structure (see section 1.3.).

3.4.8. Convergence

In the present study, convergence scores reflect the degree of a shared lexicon in a dyad; it indicates how aligned interlocutors were on their label systems. Based on Limor et al., (2019a; 2019b), convergence was measured in RStudio (R Core Team, 2016) as string similarity between labels produced by participants in the same dyad for the same target. First, I calculated the normalized Levenshtein distances between two strings for the same target item in each block, which represent the minimal number of insertions, substitutions, and deletions of graphemes required to turn one string into the other, divided by the number of graphemes in the longer string (if the strings were not the same length). Then, the distance was subtracted from 1, the final score representing string similarity between two labels. This was repeated for each target item in each block. The convergence score could range between 0 and 1, with 0 indicating no convergence between labels produced by a pair for the same item in a given block, and 1 indicating full convergence. Following previous work (Limor et al., 2019a; 2019b), convergence between participants in a pair was predicted to increase over the course of the experiment.

3.5. Data preparation

Data collection resulted in a total of 24 audio and video recordings (1 recording per participant), obtained both from the vocalization-only and combined conditions. Each audio/video file was about 40 to 120 minutes long. Of the total recordings, 8 were not included in the analysis. This was due to 3 pairs in the combined condition not producing any communicative vocalizations, and 1 pair having an incomplete audio recording. The video data was checked to see if participants were not using gestures in the vocalization-only condition but was not analyzed further for the purposes of this thesis. Additionally, CSV files of each experimental session were obtained. These included information about the onset and offset of every trial, the target item, distractor items, participant turns, and accuracy.

The audio data was processed in several steps. First, a segmentation script developed by Hans Rutger Bosker (2022a) was run in the Praat phonetic analysis software (Boersma, 2001). The script matched each audio file to the corresponding CSV file. Based on the trial onset times marked in the CSV files, it then cut the full recordings into trial-long audio files. These were further cropped into audios encompassing only the vocalizations produced by participants. This

was done by modifying Hans Rutger Bosker's annotation script (2022b), which allowed to iterate through all audio files, create TextGrid files with boundaries for silent and sounding parts of the file, and edit these boundaries. It then saved the part of the audio annotated as 'sounding' into a separate file. The vocalizations were cut to their full length, meaning that all their repetitions were included unless they were separated by a significant silent pause, suggesting that the producer was repeating the vocalization for the matcher's clarity. Any loud noise or laugh that appeared at the start or end of such repeated vocalizations was cut off. Data cleaning resulted in a total of 1024 vocalization files, two of which were empty because guessers accidentally selected items before producers could generate a label.

To obtain the speech rate and number of syllables per speech signal, all vocalizations were orthographically annotated using Praat. Next, the annotation and audio files were run through the BAS WebServices (Kisler, Reichel, & Schiel, 2017) online sound-text forced aligner. In this step, the input language was indicated as Dutch, as all participants were native speakers and produced sounds common to the Dutch inventory. Other options that were specified were: 'Pipeline name' set to 'G2P->MAUS->PHO2SYL' and 'Output encoding' set to 'IPA'[2]. The service generated output in the form of TextGrid files that contained the orthographic annotations, as well as the word-, syllable-, and phoneme-level phonemic annotations.

In the final step, two Praat scripts were written by the thesis' author (see Appendix A). The first iterated through the audio and TextGrid files and extracted the measurements for F0, intensity, duration, and HNR. The script extracted the syllable number from the TextGrid files by calculating the number of intervals in the grid with syllabic annotation for each vocalization. Then, the extracted data was written into a CSV file. The speech rate for all vocalizations was calculated by dividing the duration of the vocalizations by the number of syllables using the Excel program. The second script iterated through TextGrid files and extracted all strings of orthographic annotations. Then, the annotations were written into a separate CSV file for the compositionality and convergence analysis.

3.6. Analyses

To test the emergence and change of iconicity and compositionality in the vocal communications systems created by our participants, mixed-effects regression models were used. All models were generated by the lme4 (Bates, 2010) and lmtest (Zeileis & Hothorn, 2002) packages in R (R Core Team, 2016). The full models can be found in Appendix B.

Regression models assessing communicative success and iconic mapping of semantic meanings on acoustic measurements (discussed in section 3.4. Measurements and predictions) had either accuracy (logical 1/0) or acoustic measure (duration, F0, intensity, HNR, speech rate, or number of syllables; numeric) as the dependent variable. The fixed effects were the total number trial (numeric, centered), experimental condition (vocal vs. multimodal, dummy coded with the vocal condition as the reference level), target size (small vs. large, small as the reference level), target movement (walking vs. hopping, walking as the reference level), and target speed (slow vs. fast, slow as the reference level), as well as two-way and triple interactions for all fixed effects except target speed since the model without target speed interactions had a marginally better fit. The random effects structure included random

[2] In 'Pipeline name' the option 'G2P->MAUS->PHO2SYL' performs three BAS WebServices actions in succession. First, 'G2P' turns the input annotation into a phonetic annotation, then 'MAUS' segments the transcriptions into words and phonemes, and finally 'PHO2SYL' creates syllabic segments. In 'Output encoding' option 'ipa' is selected as this takes the input orthographical annotations to generate phonemic annotations.

intercepts for different participants, target items, and target shape[3]. In lme4 syntax, this model corresponds to: Accuracy or Acoustic measurement ~ c.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape).

Since convergence was calculated per pair over all trials in a given block, and since compositional structure was calculated per participant over all trials and all items in a given block, the models for compositional structure and convergence as the dependent variable only had a fixed effect for block number (factor 1 or 2, centered). The random effects for the compositional structure model included a random intercept for different participants, while the random effects for the convergence model included random intercepts for target items and random by-pair slopes for the effect of block number. In lme4 syntax, the model for compositionality corresponds to: Structure (raw or z-scores) ~ c.BlockNr + (1|Participant); for convergence, the model is structured as follows: Convergence ~ c.BlockNr + (1+c.BlockNr|Pair) + (1|TargetItem).

The model measuring the effect of compositional structure on communicative success (i.e., testing whether more structure was predictive of better accuracy) included fixed effects for structure score (numeric), block number (centered), and interaction between structure score and block number. The random effects structure for this model was a random intercept for different participants. In lme4 syntax, the model looks like this: Accuracy ~ Structure (raw or z-scores) * BlockNr + TargetSpeed + (1|Participant).

[3] *Target shape* was not included as a fixed effect as no previous work shows a mapping between shape and acoustic features that is relevant for the stimuli of this study (see section 3.4. Measurements and predictions).

Chapter 4. Results

The following section reports the results of the participants' vocalizations analysis. First, communicative success results are reported, followed by iconicity, compositional structure, and convergence results.

4.1. Communicative success

Communicative success increased as the trials progressed (Model 1: $\beta = 0.003$, s.e. = 0.0008, $t = 3.82$, $p < 0.0001$), indicating that participants became more accurate at guessing the correct meanings over time. By the end of communication, participants in the vocalization condition reached 73% accuracy, and participants in the multimodal condition reached 80% accuracy (59% for vocal multimodal pairs; 99% for non-vocal multimodal pairs). In comparison, pairs in the gesture condition attained a ceiling accuracy of 99% (see Figure 3).

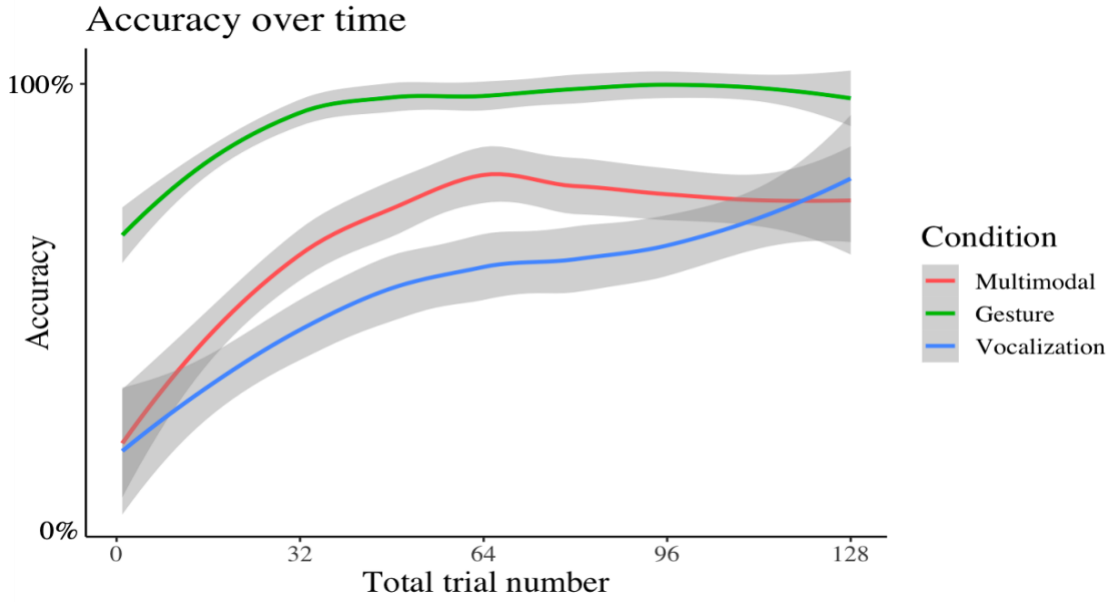

Figure 3. Accuracy over time in all conditions over all pairs (6 pairs per condition).

A t-test comparing mean accuracy scores between multimodal pairs that vocalized and multimodal pairs that did not use vocalizations showed a largely significant difference between the groups in terms of accuracy ($t(755.58) = 16.9$, $p < 0.0001$, Cohen's d = 1.038), suggesting that the overall accuracy scores for participants in the multimodal condition did not reach the ceiling level because pairs creating vocal labels were significantly less accurate. Accuracy levels for all vocal pairs are illustrated in Figure 4.

Apart from the effect of time, changes in participants' accuracy scores were not significantly modulated by any other fixed variables or their interactions in Model 1.

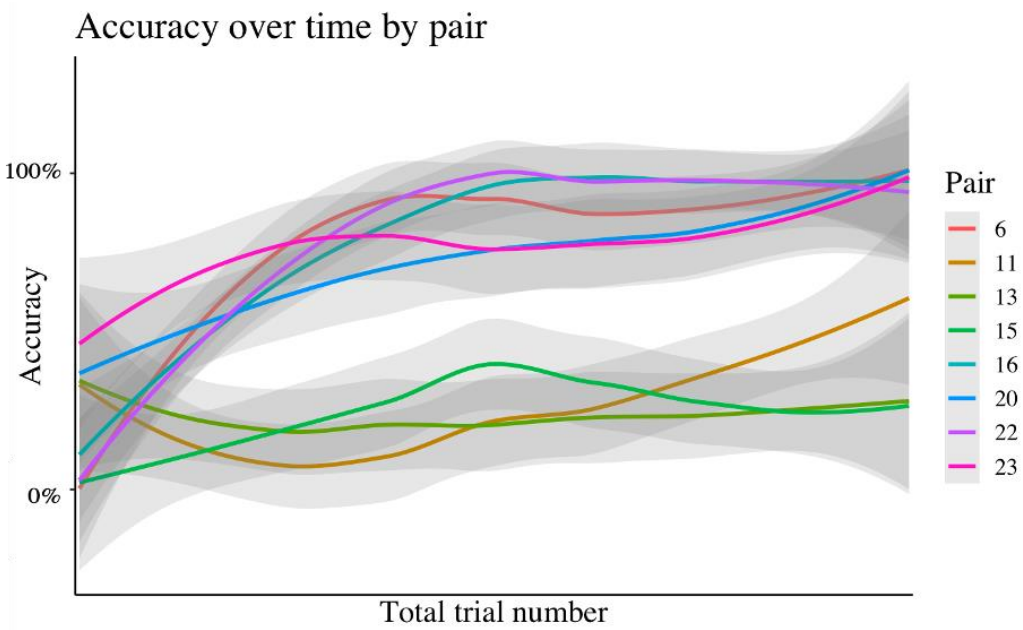

Figure 4. Accuracy over time by pairs that used vocalizations. Pairs 6, 11, 13, 16, 20, and 23 are from the non-linguistic vocalizations-only condition; Pairs 15 and 22 are pairs from multimodal condition that used vocalizations. The multimodal pair with missing audio recordings is not included in the graph.

4.2. Iconicity

Iconicity in participants' vocalizations was examined by looking at whether semantic features of the stimuli were mapped onto prosodic elements, that is, their acoustic measurements. As discussed in section 2.4. this included pitch (measured as F0), loudness (measured as intensity), vocalization duration, harmonics-to-noise ratio, number of syllables per vocalization, and speech rate. Each prosodic element is discussed separately. Table 1 summarises all the significant fixed effects and their interactions for the iconicity analysis.

	Vocalization duration	Pitch	Loudness	Harmonics-to-noise ratio	Number of syllables	Speech rate
Total trial number	✓ longer vocalizations over time	✗	✗	✗	✗	✗
Condition	✗	✗	✗	✗	✗	✗
Target size	✓ shorter vocalizations for large items	✓ lower pitch for large items	✓ louder vocalizations for large items	✗	✗	✗
Target movement	✓ longer vocalizations for walking items	✗	✗	✗	✓ more syllables for walking items	✗
Target speed	✓ shorter vocalizations for fast items	✓ higher pitch for fast items	✓ louder vocalizations for fast items	✗	✓ more syllables for fast items	✓ faster speech rate for fast items
Target movement x condition	✗	✗	✗	✗	✓ more syllables for walking items, difference larger in vocal cond.	✗
Target size x condition	✗	✓ higher pitch for small items, difference larger in vocal cond.	✗	✗	✗	✗

Table 1. Summary of iconicity results. Only significant two-way and three-way interactions are included in the table.

4.2.1. Vocalization duration

There was a significant effect of total trial number on vocalization duration (Model 2: $\beta = 0.004$, s.e. = 0.002, $t = 2.07$, $p = 0.04$), with participants producing longer vocalizations over time. This suggests that the change in participants' vocalization duration over time was gradual. The condition did not significantly affect vocalization duration (Model 2: $\beta = -1.15$, s.e. = 1.55, $t = -0.74$, $p = 0.48$).

All fixed effects for semantic properties had a significant effect on vocalization duration: slow creatures elicited longer vocalizations compared to fast creatures (Model 2: $\beta = 0.34$, s.e. = 0.06, $t = 5.56$, $p < 0.0001$); walking creatures elicited longer vocalizations than jumping creatures (Model 2: $\beta = 0.26$, s.e. = 0.1, $t = 2.65$, $p = 0.008$); and smaller creatures elicited longer vocalizations than large creatures (Model 2: $\beta = 0.26$, s.e. = 0.1, $t = 2.64$, $p = 0.008$). Moreover, there was a marginal negative interaction between target size and target movement (Model 2: $\beta = -0.28$, s.e. = 0.14, $t = -1.95$, $p = 0.051$), suggesting that participants produced marginally longer vocalizations for small creatures compared to large creatures when they were jumping, but less so when they were walking. The results are illustrated in Figure 5. No other two-way or three-way interaction was significant.

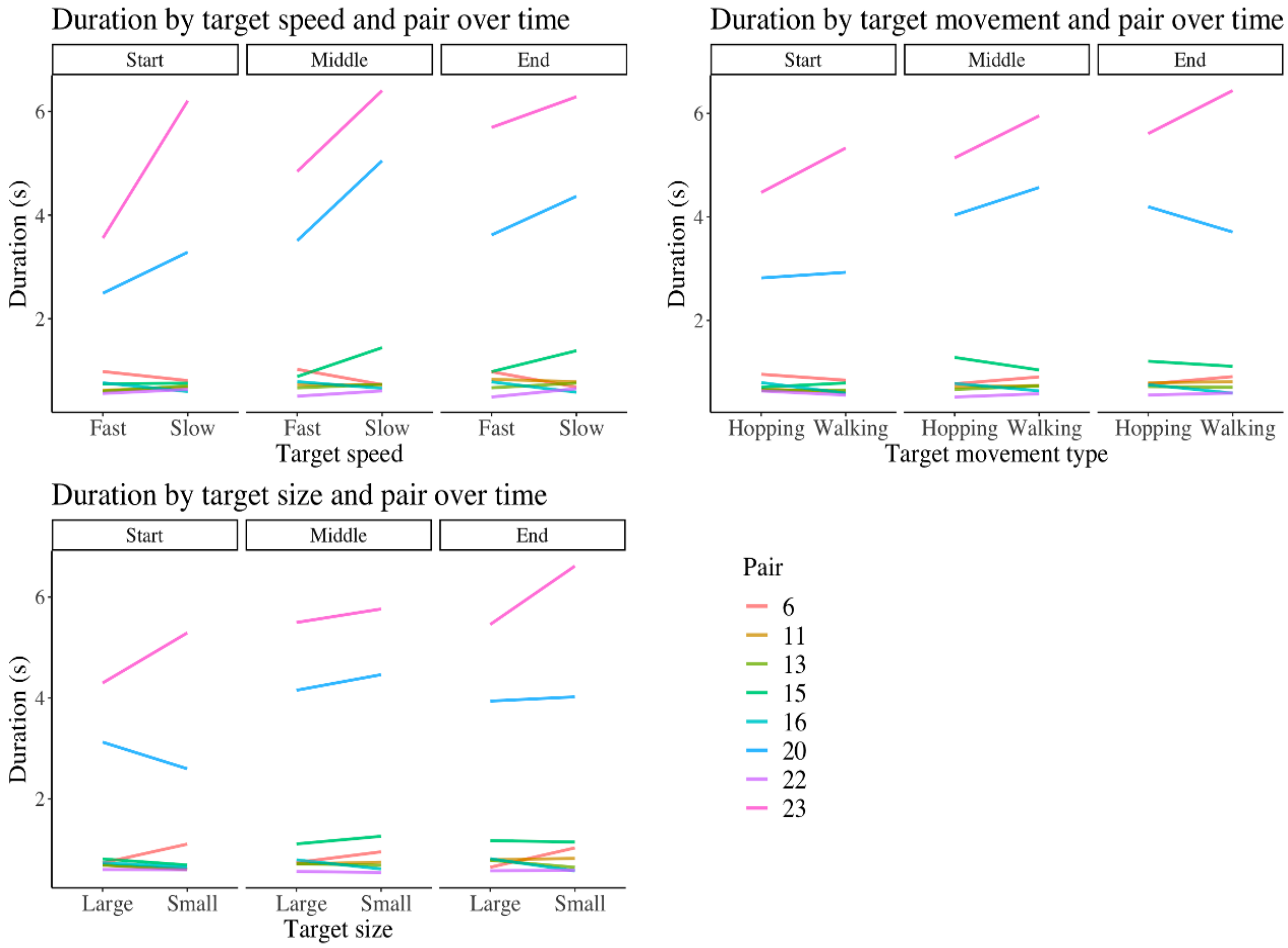

Figure 5. Iconic mappings on vocalization duration by pair over time. Here and in graphs following this, the total trial number was roughly divided into three parts, representing the start, middle, and end of the experiment. Overall, several pairs (6, 20, and 23) produced more iconic vocalizations in terms of duration, with slow, walking, and small creatures eliciting longer vocalizations than fast, hopping, and large creatures. Other pairs seem to be less iconic in their mappings on duration.

4.2.2. Pitch

Pitch levels were significantly affected by target size: participants produced vocalizations with significantly higher F0 levels when creatures were small compared to large (Model 3:

β = 17.97, s.e. = 5.61, t = 3.2, p = 0.0014). F0 levels were also affected by target speed (Model 3: β = -13.79, s.e. = 3.43, t = -4.02, p < 0.0001), such that fast creatures elicited higher pitch than slow creatures. These results are reflected in Figure 6. There was also a significant positive interaction between target size and experimental condition (Model 3: β = 39.94, s.e. = 11.12, t = 3.59, p = 0.0003), suggesting that, in the vocal condition, the difference in pitch levels for small as opposed to large creatures was larger than it was in the multimodal condition.

F0 levels were not affected by total trial number (Model 3: β = -0.03, s.e. = 0.1, t = -0.25, p = 0.8) or experimental condition (Model 3: β = -26.58, s.e. = 27.34, t = -0.97, p = 0.34), suggesting that overall, pairs had produced vocalizations with similar pitch levels throughout the experiment and across all conditions. Target movement also did not show a significant effect on F0 (Model 3: β = -8.08, s.e. = 5.63, t = -1.43, p = 0.15), with participants using similar pitch to refer to both walking and jumping creatures. None of the interactions between fixed effects were significant.

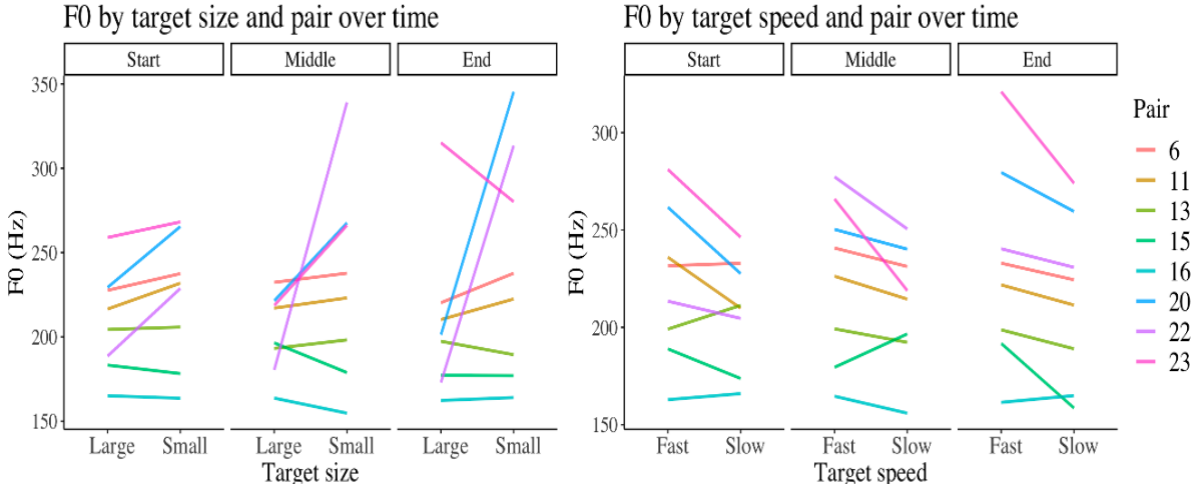

Figure 6. Iconic mappings on F0 levels across pairs over time. Notice that differences in F0 levels are present from the start of the experiment and tend to remain stable for most pairs over time.

4.2.3. Loudness

Intensity levels were significantly lower in vocalizations referring to slow creatures compared to vocalizations referring to fast creatures (Model 4: β = -1.78, s.e. = 0.26, t = -6.7, p < 0.0001). Intensity was also significantly modulated by target size, with large creatures eliciting louder vocalizations than small creatures (Model 4: β = -1.21, s.e. = 0.43, t = -2.8, p = 0.005). This indicates that participants produced louder vocalizations when the target creatures were fast and/or large, compared to slow and/or small. Condition showed a marginal effect (Model 4: β = 3.61, s.e. = 1.85, t = 1.95, p = 0.068), such that participants in the vocal condition produced marginally louder vocalizations than participants in the multimodal condition.

The total trial number also had a marginal effect on intensity levels (Model 4: β = 0.01, s.e. = 0.008, t = 1.69, p = 0.09), suggesting that participants' vocalizations had a marginal increase in loudness over time. No interactions between experiment time and the semantic features were significant.

Target movement type did not affect intensity levels (Model 4: β = -0.43, s.e. = 0.43, t = -1.004, p = 0.31), implying that participants did not map this semantic property through modulation in vocalization loudness. No interaction between the fixed effects was significant.

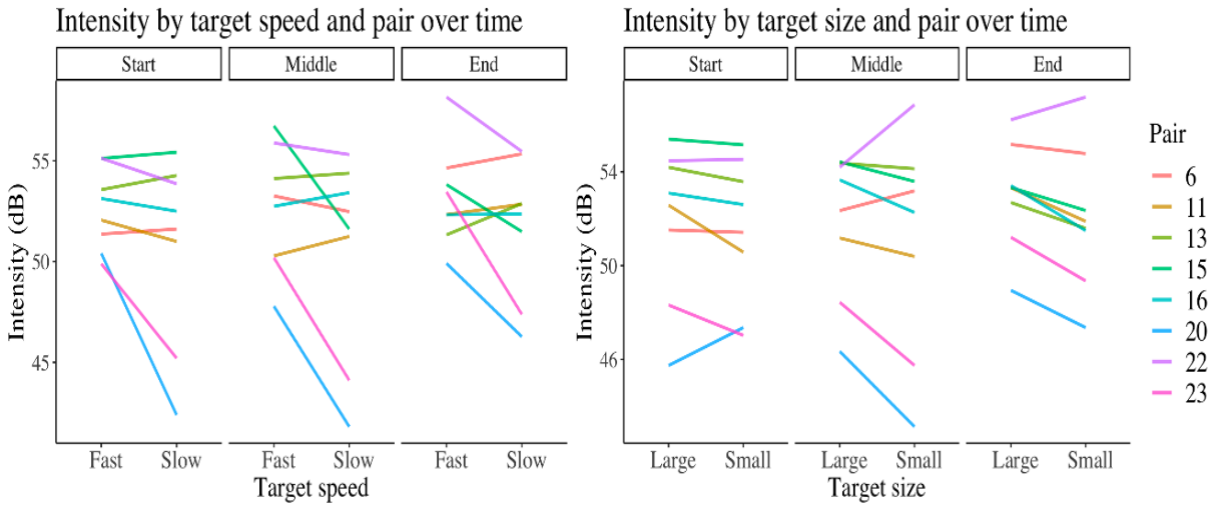

Figure 7. Iconic mappings on vocalization intensity levels across pairs over time. The trend of fast and large creatures eliciting louder vocalizations than slow and small creatures emerges right from the onset of the experiment and does not change significantly over time. Again, this difference in intensity levels depending on target speed and size is more prominent in some pairs (e.g., pairs 15, 20, and 23) than others, suggesting that these pairs utilized iconic affordance of modulating vocalization loudness to a greater extent.

4.2.4. Harmonics-to-noise ratio

There was a marginal effect of total trial number on harmonics-to-noise-ratio (Model 5: $\beta = 0.01$, s.e. = 0.007, $t = 1.87$, $p = 0.061$), suggesting that participants' vocalizations became more harmonic and less noise-like over time. No other fixed effect or their interaction was significant.

4.2.5. Number of syllables

Participants produced vocalizations with significantly more syllables when creatures were fast compared to slow (Model 6: $\beta = -2.29$, s.e. = 0.21, $t = -10.69$, $p < 0.0001$). The number of syllables also significantly varied with the target movement (Model 6: $\beta = 1.49$, s.e. = 0.35, $t = 4.27$, $p = 0.00002$): participants tended to use more syllables in their labels for walking creatures than for jumping creatures. Additionally, there was a significant interaction between target movement and experimental condition (Model 6: $\beta = -1.77$, s.e. = 0.7, $t = -2.53$, $p = 0.01$), so that, compared to pairs in the multimodal condition, pairs in the vocalization condition used more syllables when referring to walking creatures compared to jumping creatures.

The total trial number did not have a significant effect on the number of syllables (Model 6: $\beta = 0.003$, s.e. = 0.007, $t = 0.53$, $p = 0.59$), and neither did the experimental condition (Model 6: $\beta = 0.005$, s.e. = 0.013, $t = 0.37$, $p = 0.71$), indicating that, in all conditions, participants produced labels of similar length across the experiment. Target size also did not show a significant effect (Model 6: $\beta = 0.32$, s.e. = 0.35, $t = -0.64$, $p = 0.53$), which suggests that participants did not map this semantic property on the number of syllabic units. No interactions apart from the reported interaction for target movement and condition were significant.

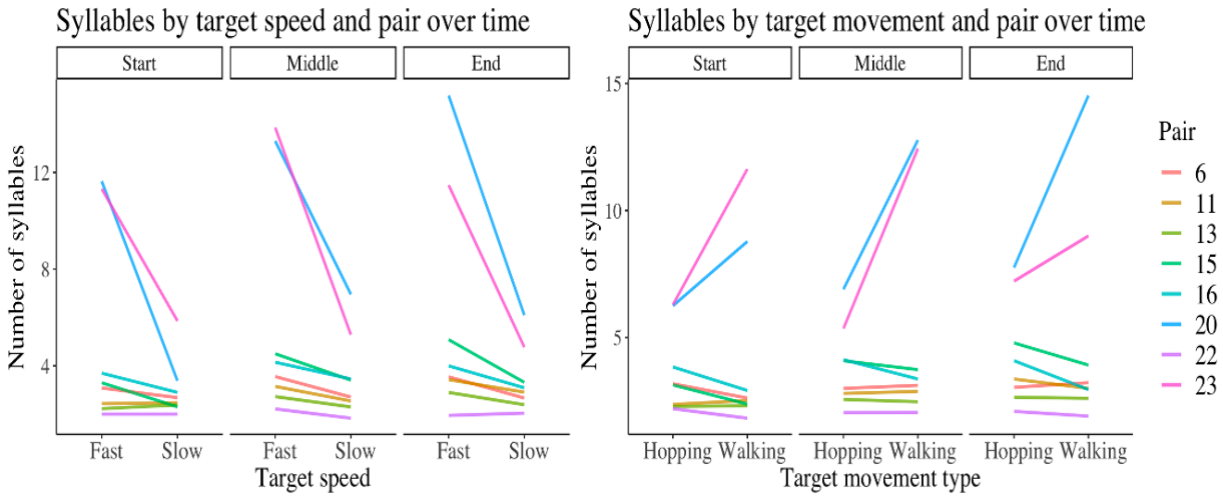

Figure 8. Iconic mappings on the number of syllables per vocalization across pairs and over time. Slow and walking creatures elicited vocalizations with a higher number of syllabic units than fast and hopping creatures. The differences in the number of syllables are extremely prominent in pairs 20 and 23. Overall, both trends are stable over time.

4.2.6. Speech rate

Target speed had a highly significant effect on speech rate (Model 7: $\beta = -1.05$, s.e. = 0.08, $t = -12.35$, $p < 0.0001$), with participants using faster speech rate for fast creatures than for slow creatures. The trend is illustrated in Figure 9. No other fixed effects or their interactions were significant for speech rate in Model 7.

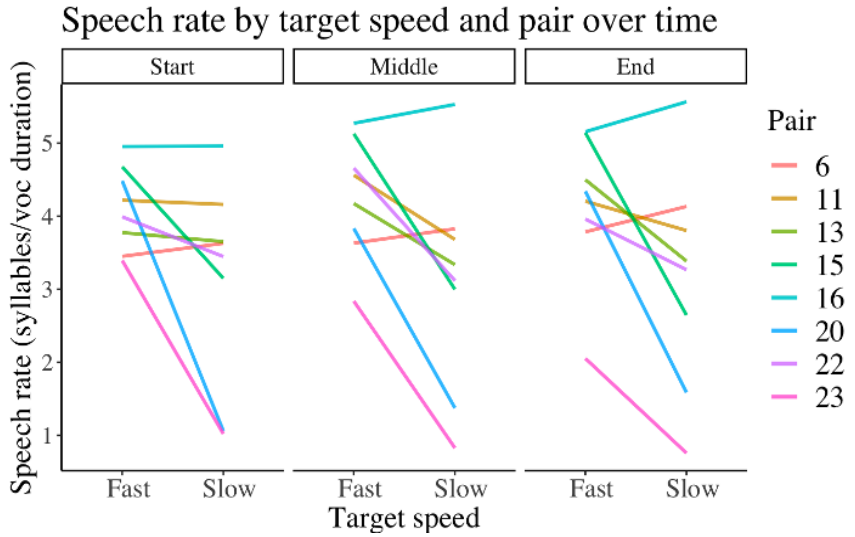

Figure 9. Iconic mapping of target speed on speech rate across pairs and over time. Participants produced faster speech rates for fast-moving creatures compared to slow creatures. Pairs 13, 20, and 23 showed the largest differences in their modulations of speech rate between different creature speeds. The trend is visible right from the onset of the experiment and remains similar throughout trials.

4.3. Compositional structure

	Compositional structure (raw scores)	Compositional structure (z-scores)
Block number	✓ compositionality increases over blocks	✓ compositionality increases over blocks

	Communicative success
Compositional structure (raw scores)	✓ accuracy increases with increasing compositional structure score
Compositional structure (z-scores)	✓ accuracy increases with increasing compositional structure score
Block number	✓ accuracy increases over blocks
Structure score x block number	✗

Table 2. Summary of compositional structure results.

Compositional structure (raw) significantly increased over blocks (Model 8: $\beta = 0.2$, s.e. = 0.04, $t = 5.18$, $p = 0.000113$), demonstrating that participants' languages became more systematic over time (see Figure 10). A similar trend emerged when testing changes in compositional structure z-scores over blocks (Model 9: $\beta = 4.67$, s.e. = 0.83, $t = 5.63$, $p < 0.0001$), with a significant increase in structure z-scores over time.

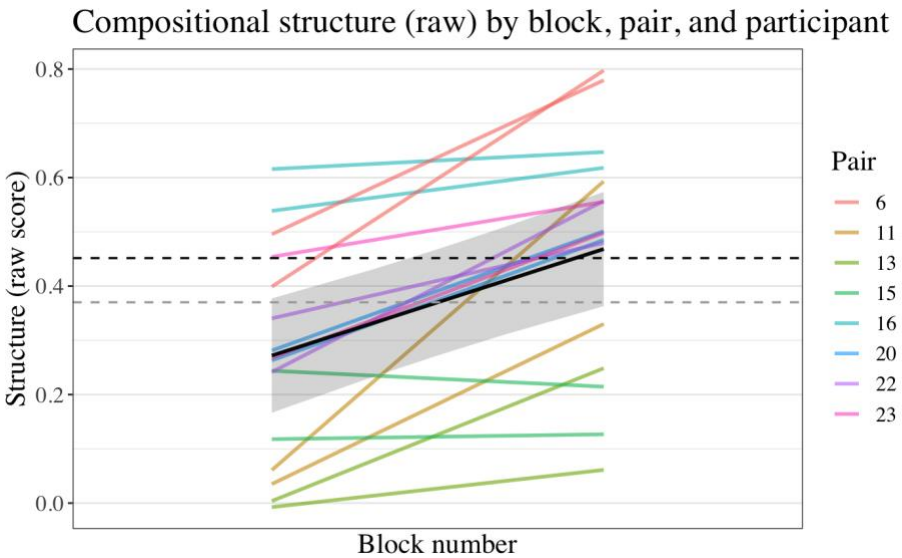

Figure 10. Changes in compositional structure over time. Each line represents a participant, participants from the same pair are given the same color. The solid black line indicates the effect of the increase in block number on structure scores, and their shadings show the reported effect's standard errors. The dashed dark lines demonstrate the upper boundary of 95% CI, and the dashed grey lines indicate the structure score mean.

In the first block, the mean compositionality score was 0.27 (SD = 0.19), and for the second block, it reached 0.47 (SD = 0.22). Overall, raw compositional structure scores had a mean of 0.37 (SD = 0.27). The 95% confidence interval was calculated as [0.2885471;

0.4516041], suggesting that there is a 95% chance the true mean structure scores of languages developed in our experiment fall between the scores 0.2885471 and 0.4516041. Two pairs (pair 6 and pair 11 from the vocalization condition) created compositional languages higher than the mean and confidence interval right from the onset of the first block. However notably, all pairs created significantly compositional languages by the end of the experiment.

Although all pairs developed ways to communicate with each other, these communication systems showed varying levels of compositional structure, with some pairs developing highly structured vocalizations. Figure 11 illustrates two examples of compositionality in languages created by participants: language A is the language with the highest compositional structure score (structure score: 0.62, pair 6); language B is an example of a language with some apparent but smaller compositional structure (structure score: 0.38, pair 20). The total averages of structure scores for all pairs can be found in Appendix C.

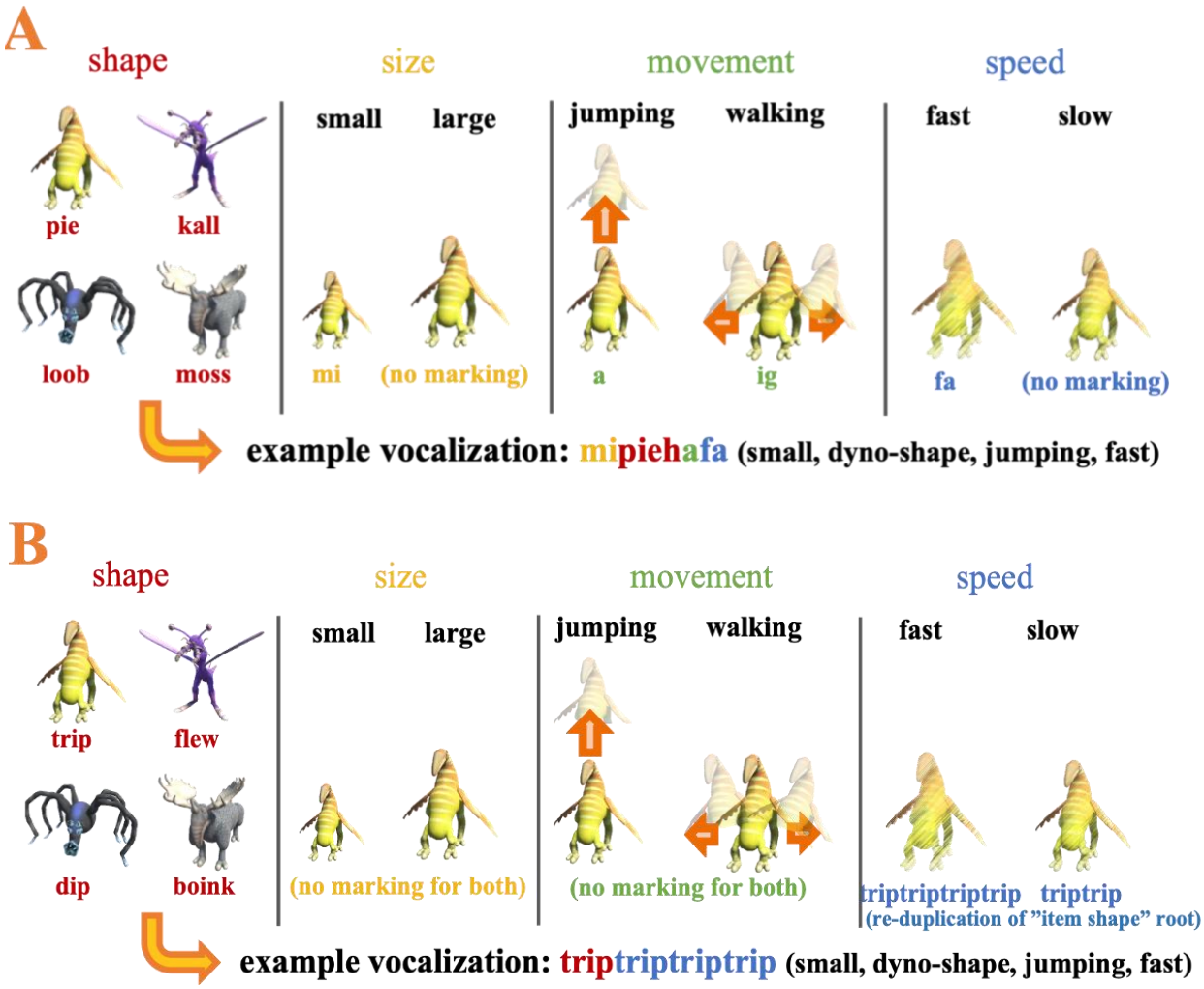

Figure 11. Example of a highly compositional language (A) and less compositional language (B). The structures are illustrated as they were at the end of the experiment (last 32 trials).

An additional model tested whether compositional structure is predictive of participants' accuracy scores. The model showed that raw compositional structure scores had a significant effect on communicative success (Model 10: $\beta = 0.87$, s.e. = 0.17, $t = 5.12$, $p < 0.0001$), demonstrating that more compositionality was associated with greater communicative success. In this model, there was no significant effect of block number on accuracy (Model 10: $\beta = -0.07$, s.e. = 0.1, $t = -0.7$, $p = 0.5$). In addition, there was no significant interaction between structure scores and block number (Model 10: $\beta = 0.27$, s.e. = 0.23, $t = 1.19$, $p = 0.26$), indicating that more compositionality was associated with better accuracy throughout the

experiment. A Pearson's correlation test also revealed a strong positive correlation between compositional structure and accuracy scores ($r(30) = 0.83$, $t = 8.29$, $p < 0.0001$). This relationship is illustrated in Figure 12.

Correlation between compositional structure and communicative success

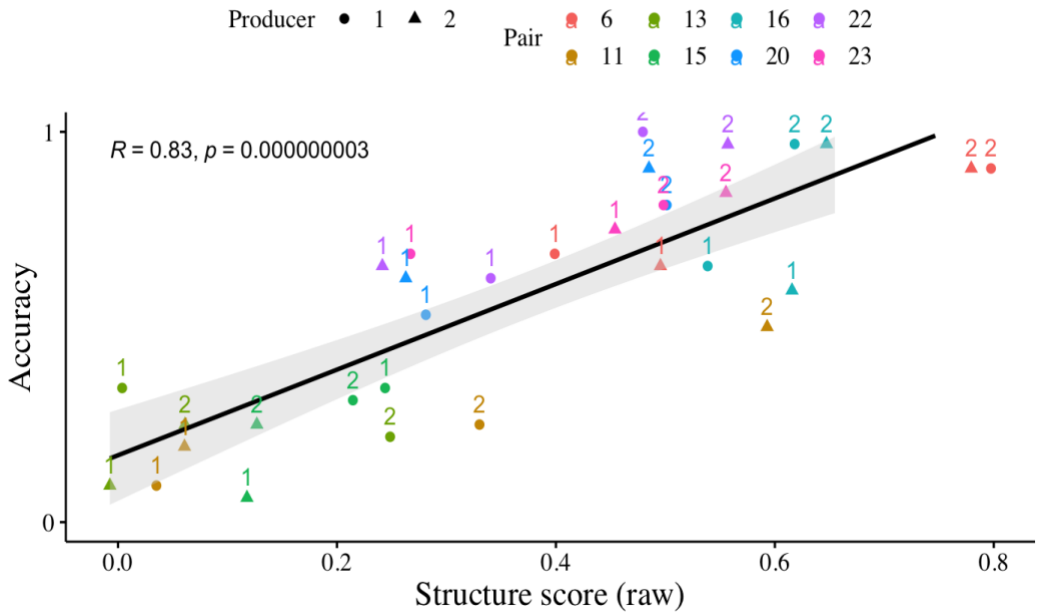

Figure 12. The correlation between compositional structure and communicative success across participants. The black line represents the correlation and its shading illustrates standard errors. Data points are colored by pair. Different data point shapes (points and triangles) represent different participants within a pair. The number above each data point indicates the block number the point belongs to.

4.4. Convergence

Convergence significantly increased over rounds (Model 12: $\beta = 0.24$, s.e. $= 0.05$, $t = 4.7$, $p = 0.002$), showing that participants used increasingly similar labels over time to refer to the same item.

Chapter 5. Discussion

The present thesis investigated the emergence of iconicity and compositional structure in *vocal* communication systems, developed by pairs of participants communicating in VR environment in one of three modality conditions: non-linguistic vocalization-only, gesture-only, or multimodal. It was hypothesized that interlocutors would create iconic mappings for different semantic features associated with our novel set of stimuli (shape, size, movement type, and speed) through modulations in acoustic features (vocalization duration, pitch, loudness, harmonics-to-noise ratio, number of syllables per vocalization, and speech rate). The results demonstrated that many acoustic features were significantly related to the semantic features of stimuli (except for HNR, which did not map any semantic features). Specifically, vocalization duration was used to express target size, movement type, and speed (i.e., longer signals for small, walking, or slow creatures), pitch was used to express target size and speed (i.e., higher frequency for small or fast creatures), loudness was used to express target size and speed (i.e., louder signals for large or fast creatures), the number of syllables was used to express target movement type and speed (i.e., more syllables for walking or slow creatures), and speech rate was used to express target speed (i.e., faster rate for fast creatures). In addition, it was predicted that compositional structure would emerge and increase over repeated interaction between participants. The analysis showed that compositional structure emerged in varying levels in pairs' languages and that in general, the structure increased as the experiment progressed. Notably, the accuracy of participants' guesses improved with the increase in compositional structure, supporting the idea that more grammatically systematic languages lead to higher communicative success. Moreover, participants showed higher levels of convergence in their languages over time. Finally, the data analysis showed that participants in the vocal-only condition were less accurate than participants in the multimodal or gesture-only condition, the latter reaching the ceiling level. Within the multimodal pairs, interlocutors who used only vocalizations to describe the stimuli were less accurate than pairs who converged on using gestures.

5.1. Communicative success

The findings of the present thesis have several implications about the role of modality in the success of communication when interlocutors are deprived of an existing language system and need to communicate about novel entities. First, the present results are consistent with previous findings of similar communication game experiments by showing that gestural communication has an advantage over vocal when grounding new systems (Fay et al. 2014; Macuch Silva et al., 2020). However, in previous work, participants hardly converge on a shared system in the vocalization-only condition, while here we found that, by the end of the game, pairs in the vocal-only condition reached an average accuracy of 73% – a success well above chance level. The higher accuracy demonstrated by vocalization-only participants in this study could be due to the differences in our methods and stimuli compared to earlier studies, which are arguably more ecologically valid. In the present experiment, the target items were unfamiliar to participants and had no conventionalized labels (as opposed to studies such as Fay et al. (2013; 2014), which use existing referents like "sleeping" and "tree"). Moreover, the items differed in semantic features that were matched for both vocal and gestural signaling (e.g., shape, size can be iconically expressed in both vocal languages and gesture – see Perlman & Cain, 2014). Moreover, participants communicated face-to-face, meaning that additional information, such as that conveyed by facial and other visual cues, was available to them. Facial signals are shown to contribute to the communication of relative information, such as communicative acts of asking and providing an answer (Nota, Trujillo, & Holler, 2021). Immersing participants in face-to-face interaction without computer interference arguably made our experiment closer to

real-world communicative situations, potentially facilitating vocalization success. Nevertheless, it is currently unknown whether the increased communicative success of vocal pairs in our experiment was due to the availability of face-to-face interaction. For purposes of the present thesis, I do not try to answer this question, but future work could investigate the role of facial cues in vocal communication when the use of an existing language is restricted.

As for multimodal communication, the finding that it was less successful than gestural communication, but more successful than vocal communication is also in line with previous studies (Fay et al., 2014; Macuch Silva et al., 2020). In addition, the current analysis replicates the findings of Macuch Silva et al. (2020), namely that pairs in the multimodal condition communicate about the stimuli mostly in one modality. In our experiment, 3 out of 6 multimodal pairs used only gestures, 2 used only vocalizations, and 1 used a combination of gestures and vocalizations (the latter pair was not included in the analysis as the audio recordings were not complete). All pairs converged on the one modality they started with from the onset of the experiment. Thus, it appears that when interlocutors had the choice to produce unimodal or multimodal signals, they utilized the flexibility of the multimodal condition to describe the stimuli in a modality they found to have better affordance for a given situation.

Why then multimodal pairs did not perform equally well as pairs communicating with gestures only? As the analysis results indicate, the overall accuracy of multimodal pairs was brought down by multimodal pairs that chose to communicate only in vocalizations, while multimodal pairs that used only gestures had similar accuracy levels to the pairs in the gesture-only condition. Moreover, it might be that the ability to choose either or both modalities could have initially confused multimodal participants, slightly delaying the increase of their accuracy until they converged on the strategy most for them.

5.2. Iconicity

The thesis' findings offer insights into the role of iconicity in emerging vocal communication systems. First, the current experiment echoes previous communication game studies, showing that interlocutors can create iconic vocalizations to successfully signal different meanings (e.g., Perlman et al., 2015). Second, the current findings replicates sound symbolism studies of existing spoken languages (discussed to a greater extent in section 3.4. *Measurements and Predictions*) by demonstrating that meanings can be iconically mapped through acoustic modulations of speech. In line with the previous findings and predictions of the present thesis, it was found that participants changed the duration of their vocalizations to refer to creatures of different sizes and speeds (longer signals for small or slow creatures), pitch level to refer to different sizes (higher pitch for small creatures), loudness to refer to size (louder signals for large creatures), syllable number to refer to speed (more syllables for faster creatures), and speech rate to refer to speed (faster speech rate for fast creatures). Additionally, the analysis showed several significant acoustic measure-semantic meaning mappings that, to the author's knowledge, have not been explored or found in previous iconicity or sound symbolism studies. Among these are the mappings between target speed and pitch (higher pitch for fast creatures), speed and loudness (louder vocalizations for fast items), movement type and vocalization duration (longer vocalizations for walking items), and syllable number and movement type (more syllables for walking items).

One prediction was not confirmed: differences in creature sizes did not affect the number of syllabic units in participant's vocalizations. Notably, the prediction was formulated based on iconicity findings in reduplication, which indicate speakers of some existing languages duplicating words or smaller morphosyntactic units to infer meanings of intensity and repetition, among others (Dingemanse et al., 2015). Reduplication stands at an interesting threshold between iconicity and compositionality: on one side, the repeating of a root word or its part intuitively mimics experienced intensity or repetition, at the same time, such

duplications of root words can eventually be systematically used to create new signals, making them grammaticalized morphemes. Reduplication can be found in our participants' languages, especially in vocalizations of Pairs 20 and 23, when participants referred to creatures of different speeds (see Figure 8; for an example of re-duplication, see Figure 11, language B). However, it appears that creature size did not evoke the connection between the intensity of size and the number of syllable repetitions. Instead, our participants mapped size on acoustic measurements, namely vocalization duration, pitch, and loudness. It is possible that participants found this tactic more intuitive for expressing largeness and smallness.

Overall, vocalization duration appeared to be the most iconically expressive feature, used to refer to most meanings (size, movement type, and speed). Harmonics-to-noise levels did not differ significantly for any of the meanings. All other acoustic features mapped one or two semantic features. Noteworthy, it is possible that some of the considered acoustic measures are connected and do not necessarily show that certain meanings are purposely or subconsciously associated with and mapped on a specific acoustic feature. For instance, vocalization duration, number of syllables, and speech can be interrelated. That is, an increased number of syllables might result in longer vocalizations; simultaneously, a faster speech rate might make vocalizations shorter even with a high syllable count. The connection between duration, speech rate, and syllable number can suggest an explanation of why variations in vocalization duration described so many meanings. It is plausible that the significant differences in vocalization duration stemmed from the significant differences in speech rate and syllable number, or that participants employed multiple acoustic tactics simultaneously to enhance the iconic descriptions of semantic features. A larger data sample could potentially help disentangle the connection between these or other acoustic measurements.

Notably, although previous research has demonstrated robust findings regarding the mappings between the semantic domain of shape and phonetic properties of vocalizations, the current analysis showed that target shape did not sufficiently account for the models as a fixed effect and was integrated as a random effect. Several factors might explain the lack of effect of target shape in the current study. First, in contrast to previous work, the stimuli used here were less familiar than real-world entities, i.e., they were less likely to have pre-established labels participants could use. Unlike with other meanings such as different speeds and sizes, participants did not possess readily available tactics for mapping target shape through iconic modulations of their vocalizations. At the same time, shape (i.e., different creature types) is the most obvious feature differentiating the stimuli, which poses a need to refer to it in a way that is obvious and stable. Thus, the initial unfamiliarity with the stimuli and the simultaneous need to signal creature shape might have influenced participants to opt for a more uniform tactic than acoustic modulations. That is, mapping shapes through compositional units and re-using the established units for all targets of the same shape. This is further addressed below in the subsequent section discussing compositional structure.

Another possible explanation for the lack of effect of target shape may be participants focusing on stimuli features that were not included in the analysis. In other words, "target shape" could have been decomposed into sub-features that were not investigated here. Supporting this idea, 3 participants in the vocal-only condition and 1 participant in the multimodal condition named color as the distinguishing feature between the target items, as different creature shapes had varying colors (see Figure 1). One of these participants reported producing "higher sounds" for "brighter colors" and "lower sounds" for "darker colors". Moreover, there might be more abstract meanings participants came to associate with the target items that the current analysis did not take into account. In connection to our stimuli, Kilpatrick and colleagues (2023) showed a systematic relationship between certain phonemes in names for Pokémon creatures and the level of creatures' friendliness across six spoken languages. Notably, older work on sound symbolism has shown that shape can evoke affective

associations: rounder shapes (maluma/bouba) are judged as friendlier compared to sharp shapes (takete/kiki) (Lindauer, 1990). Moreover, people prefer rounded shapes to shapes with sharp angles, as the latter appear to be indicators of threat (Bar & Neta, 2006). Consequently, the perceived friendliness of stimuli might have affected acoustic changes of participants' vocalizations in our experiment, as some creatures were potentially less friendly (e.g., spider-looking creature) than more bubbly-looking creatures (e.g., dinosaur-looking creature). Follow-up work could investigate possible interactions between physical and abstract features of referents with more detail.

5.3. Compositional structure

The present study provides evidence for the emergence of compositionality in vocal communication systems created in a context as small as pair interaction. The analysis showed that varying levels of compositionality emerged in all languages over the course of the experiment, and that compositionality emerged even in the absence of new learners (as argued in Kirby et al., 2015). This finding was already supported by Nölle et al., (2018) and Raviv et al. (2019), who showed that generational transmission is not necessary for the emergence of compositional structure when there is a need to interact with multiple interlocutors and/or communicate about an expanding meaning space. The present thesis is in line with previous work by showing that a large and expanding meaning space already presents enough learnability pressure to give rise to more compositional languages within the limits of dyad communication. Specifically, the present experiment had a large meaning space, with 4 semantic domains and 32 possible meaning combinations (compared to 2 semantic domains and 23 meaning combinations in Raviv et al., 2019 and 2 domains and 12 possible combinations in Kirby et al., 2015) that kept expanding over time (participants did not see all possible creatures in one trial). It is possible that the need to refer to an increasing number of new items over time pressured pairs to re-use some linguistic components that were previously agreed upon and recombine them to refer to the same feature (e.g., fast speed) instead of creating new motivated signals for each new referent from scratch. Similarly, the relatively large meaning space used in this experiment could have introduced enough complexity to the task so that participants could not rely only on iconic expressions for successful communication, as doing so would mean negotiating new signals for each meaning and memorizing all unique labels for a large number of meaning combinations. Therefore, the need to differentiate many items throughout communication likely introduced a pressure for converging on some compositional structure. In this sense, the present findings of compositionality in novel vocal communication systems support Motamedi et al.'s (2019) findings on gestural systems, which show that the pressure to create compositionality can be present in both generational transmission and repeated interaction between the same interlocutors. In addition, the analysis demonstrated that the increase in compositionality across participants' languages predicts an increase in communicative success. This is in connection with the findings of Raviv, Meyer, and Lev-Ari (2020), who showed that higher structure in participants' languages boosted communicative success and convergence in groups of four and eight interlocutors. The current experiment demonstrates that interacting dyads can also increase their communication accuracy not only through motivated signs but also through developing compositional structure.

Another important point of discussion is the potential tradeoff between compositionality and iconicity. As discussed in greater detail in section 1.3., iconically motivated signals can help ground meanings; simultaneously, cognitive restrains introduce pressures for interlocutors to re-use established signals instead of creating new highly iconic signals for each novel entity in the expanding meaning space. Eventually, re-composition should become the primary strategy for creating new labels. The current analysis showed that pairs 20 and 23 were

consistently highly iconic across meanings and acoustic measures (see Figures 5-9 in section 4.2), while their compositional structure scores fell in the middle, reaching 0.4 and 0.44, respectively (see Appendix C for the total compositional structure scores of all pairs). Meanwhile, pairs that seem less apparent or consistent in their modulations of acoustic measurements that could suggest iconic mappings (such as Pairs 13, 15, and 16), showed different compositionality levels. For instance, Pairs 13 and 15 had the lowest total compositional structure scores among all pairs (0.08 and 0.17, respectively), while Pair 16 reached one of the highest structure scores (0.6). Notably, the pairs that were not particularly iconic and had the lowest structure scores (Pairs 13 and 15) also demonstrated the lowest accuracy rates among all pairs, with Pair 13 reaching around 23% accuracy and Pair 15 – 24% accuracy (see Figure 4). The most compositional pair, namely Pair 6, demonstrated considerable iconicity in certain meaning-acoustic measure mappings, while simultaneously displaying no discernible trend in other mappings. Despite this apparent absence of a uniform trend, we can highlight several interesting observations: 1) even the most iconic pairs developed a moderate level of compositionality; 2) least accurate pairs struggled to effectively employ either iconicity, compositionality, or a combination of both strategies to increase the communicative success; 3) the most iconic pairs did not show the lowest levels of compositional structure, and the most structural pairs were not necessarily the least iconic. These implications suggest that while iconicity and compositionality seem to come at an expanse of one another to a varying degree, both strategies can co-exist in participants' languages.

It is important to acknowledge that the compositional structure observed in the current analysis might be downplayed or not fully captured. In this study, compositionality scores were calculated based on orthographic annotations of participants' vocalizations. However, compositional structure can be more complex and extend beyond morphosyntactic elements to encompass other linguistic levels. For example, it is possible that participants modulated pitch or loudness among other acoustic features in a compositional manner, such as systematically raising their pitch every time a target creature was small. In other words, some of the iconic mappings to acoustic measurements could be employed systematically, thus demonstrating the presence of compositionality at a phonetical level – a thing that was not investigated here. In other words, our measure of compositionality is likely downplayed in the present thesis, and if we were to use a more sensitive/wholesome measure we may have detected compositional structure emerging even earlier or to a greater extent across participants' languages. A good direction for further work would be to consider a novel approach to measuring compositionality. For instance, clustering could be utilized to group vocalization segments based on their acoustic features (e.g., pitch, loudness, duration, etc.). The resulting clusters of segments may potentially indicate a systematic re-use of the segment to express a given meaning. Then, these clusters could be incorporated as pair-wise comparisons alongside Levenshtein's distances used in the present analysis.

Chapter 6. Conclusions

The present thesis addresses two key questions concerning the creation of artificial vocal communication systems: the level of iconicity in the vocal modality and the conditions needed for compositional structure to emerge. The experiment was carried out in an interactive Virtual Reality environment with participants communicating face-to-face about a novel set of stimuli differing in semantic features, namely shape, size, movement, and speed. I measured iconicity as correlations between semantic and acoustic features and compositionality as the pair-wise correlation between meanings and orthographic annotations and analyzed their presence in the vocal systems of participants. The analysis showed that participants mapped meanings with modulations in the acoustic features of their vocalizations, namely duration, pitch, loudness, number of syllables, and speech rate. Simultaneously, all pairs developed some level of compositional structure, with compositionality increasing over repeated interaction and growing meaning space despite no introduction of new learners. These findings indicate that iconicity and compositionality can co-exist in participants' vocal communication systems. Noteworthy, the analysis may not reflect all existing compositionality in participants' languages: only orthographic annotations were used to calculate compositionality scores, and more compositionality may be found on the acoustic level.

Ultimately, the thesis shows that participants relied on both iconic mappings and compositional structure to refer to novel entities. The findings also highlight the need for follow-up studies developing better measures for compositional structure that account for compositionality in other linguistic levels apart from morphosyntax.

References

Arbib, M. A., Liebal, K., & Pika, S. (2008). Primate vocalization, gesture, and the evolution of human language. *Current anthropology*, 49(6), 1053-1076. DOI: 10.1086/593015

Aryani, A., Conrad, M., Schmidtke, D., & Jacobs, A. (2018). Why 'piss' is ruder than 'pee'? The role of sound in affective meaning making. *PLoS One* 13:e0198430. DOI: 10.1371/journal.pone.0198430

Bar, M., & Neta, M. (2006). Humans prefer curved visual objects. *Psychological science,* 17(8), 645-648. DOI: 10.1111/j.1467-9280.2006.01759.x

Bates, D. M. (2010). lme4: *Mixed-effects modeling with R*. See https://cran.r-project.org/web/packages/lme4/index.html

Blake, B. J. (2017). Sound symbolism in English: Weighing the evidence. *Australian Journal of Linguistics*, 37(3), 286-313. DOI: 10.1080/07268602.2017.1298394

Blasi, D. E., Wichmann, S., Hammarström, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound–meaning association biases evidenced across thousands of languages. *Proc. Natl. Acad. Sci. U.S.A*. 113, 10818–10823. DOI: 10.1073/pnas.1605782113

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Bohn, M., Call, J., & Tomasello, M. (2019). Natural reference: A phylo-and ontogenetic perspective on the comprehension of iconic gestures and vocalizations. *Developmental science*, 22(2), e12757. DOI: 10.1111/desc.12757

Bosker, H. R., (2022a). *Segments into trials* [Praat script]

Bosker, H. R., (2022b). *Annotation* [Praat script]

Botha, R. (2006). Pidgin languages as a putative window on language evolution. *Language & Communication*, 26(1), 1-14. DOI: 10.1016/j.langcom.2005.07.001

Cartmill, E. A., Beilock, S., & Goldin-Meadow, S. (2012). A word in the hand: action, gesture and mental representation in humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585), 129-143. DOI: 10.1098/rstb.2011.0162

Cäsar, C., Zuberbühler, K., Young, R. J., & Byrne, R. W. (2013). Titi monkey call sequences vary with predator location and type. *Biology letters*, 9(5), 20130535. DOI: 10.1098/rsbl.2013.0535

Christiansen, M. H., & Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in cognitive sciences*, 7(7), 300-307. DOI: 10.1016/S1364-6613(03)00136-0

Corballis, Michael C. (2012). The origins of language in manual gestures. In Maggie Tallerman & Kathleen R. Gibson (eds.), *The Oxford Handbook of Language Evolution*, 382–386. Oxford: Oxford University Press. DOI:10.1093/oxfordhb/9780199541119.013.0041

Dediu, Dan & Bart de Boer. (2016). Language evolution needs its own journal. *Journal of Language Evolution* 1(1). 1–6. DOI:10.1093/jole/lzv001.

Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Lang. Linguist. Compass* 6, 654–672. DOI: 10.1002/lnc3.361

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends Cogn. Sci*. 19, 603–615. DOI: 10.1016/j.tics.2015.07.013

Ekström, A. G., Nirme, J., & Gärdenfors, P. (2022). Motion iconicity in prosody. *Frontiers in Communication*, 7, 994162. DOI: 10.3389/fcomm.2022.994162

Fay, N., & Ellison, T. M. (2013). The Cultural Evolution of Human Communication Systems in Diferent Sized Populations: Usability Trumps Learnability. *PLoS ONE*, 8(8), e71781. DOI: 10.1371/journal.pone.0071781

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

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a Communication System from Scratch: Gesture Beats Vocalization Hands Down. *Frontiers in Psychology*, 5: 354. doi: 10.3389/fpsyg.2014.00354

Feyereisen, P., & De Lannoy, J. D. (1991). *Gestures and speech: Psychological investigations.* Cambridge University Press.

Fitch, W. T. (2010). The evolution of language. *Cambridge University Press*.

Fröhlich, M., Sievers, C., Townsend, S. W., Gruber, T., & van Schaik, C. P. (2019). Multimodal communication and language origins: integrating gestures and vocalizations. *Biological Reviews*, 94(5), 1809-1829. DOI: 10.1111/brv.12535

Gentilucci, M., & Corballis, M. C. (2006). From manual gesture to speech: A gradual transition. *Neuroscience & Biobehavioral Reviews*, 30(7), 949-960. DOI: 10.1016/j.neubiorev.2006.02.004

Goldin-Meadow, S. (2017). What the hands can tell us about language emergence. *Psychonomic Bulletin & Review*, 24, 213-218. DOI: 10.3758/s13423-016-1074-x

Haynie, H., Bowern, C., & LaPalombara, H. (2014). Sound symbolism in the languages of Australia. *PLoS One*, 9(4), e92852. DOI: 10.1371/journal.pone.0092852

Hirst, D. J., & de Looze, C. (2021). Measuring Speech. Fundamental frequency and pitch. In Knight, R., A. & Setter, J. (Eds.), *The Cambridge Handbook of Phonetics*. Cambridge University Press. DOI: 10.1017/9781108644198

Hunsicker, D., & Goldin-Meadow, S. (2012). Hierarchical structure in a self-created communication system: Building nominal constituents in homesign. *Language*, 88(4), 732. DOI: 10.1353/lan.2012.0092

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological science*, 16(5), 367-371. DOI: 10.1111/j.0956-7976.2005.01542.x

Kawahara, S. (2021). Phonetic bases of sound symbolism: a review. *PsyArXiv*. DOI: 10, 31234.

Kelly, S. D., Özyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological science*, 21(2), 260-267. DOI: 10.1177/0956797609357327

Kendon, A. (2014). The 'poly-modalic' nature of utterances and its implication. In D. Dor, C. Knight, & D. Lewis (Eds.), *The social origins of language* (pp. 67–76). Oxford: OUP.

Kendon, A. (2017). Reflections on the “gesture-first” hypothesis of language origins. *Psychonomic bulletin & review*, 24, 163-170. DOI: 10.3758/s13423-016-1117-3

Kilpatrick, A., Ćwiek, A., Lewis, E., & Kawahara, S. (2023). A cross-linguistic, sound symbolic relationship between labial consonants, voiced plosives, and Pokémon friendship. *Frontiers in Psychology*, 14, 1113143. DOI: 10.3389/fpsyg.2023.1113143

Kirby, S. (2002). Natural language from artificial life. *Artificial life*, 8(2), 185-215. DOI: 10.1162/106454602320184248

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681-10686. DOI: 10.1073/pnas.0707835105

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28, 108-114. DOI: 10.1016/j.conb.2014.07.014

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

Kisler, T., Reichel, U. D., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, Volume 45, September 2017, pages 326–347. DOI: 10.1016/j.csl.2017.01.005

Lacey, S., Jamal, Y., List, S. M., McCormick, K., Sathian, K., & Nygaard, L. C. (2020). Stimulus parameters underlying sound-symbolic mapping of auditory pseudowords to visual shapes. *Cognitive Science*, 44(9), e12883. DOI: 10.1111/cogs.12883

Levinson, S. C. (1983). Pragmatics. *Cambridge university press*.

Levinson, S. C., & Holler, J. (2014). The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130302. DOI: 10.1098/rstb.2013.0302

Levréro, F., Touitou, S., Frédet, J., Nairaud, B., Guéry, J. P., & Lemasson, A. (2019). Social bonding drives vocal exchanges in bonobos. *Scientific reports*, 9(1), 711. DOI: 10.1038/s41598-018-36024-9

Lindauer, M. S. (1990). The meanings of the physiognomic stimuli taketa and maluma. *Bulletin of the Psychonomic Society*, 28(1), 47-50. DOI: 10.3758/BF03337645

Lowe, M. L., & Haws, K. L. (2017). Sounds big: The effects of acoustic pitch on product perceptions. *Journal of Marketing Research*, *54*(2), 331-346. DOI: 10.1509/jmr.14.0300

MacNeilage, P. F. (2010). The origin of speech (No. 10). *Oxford University Press*.

Macuch Silva, V., Holler, J., Ozyurek, A., & Roberts, S. G. (2020). Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society open science*, 7(1), 182056. DOI: 10.1098/rsos.182056

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, 192, 103964. DOI: 10.1016/j.cognition.2019.05.001

Nölle, J., & Spranger, M. (2022). From the field into the lab: Causal approaches to the evolution of spatial language. *Linguistics Vanguard*, 8(s1), 191-203. DOI: 10.1515/lingvan-2020-0007

Nölle, J., Hartmann, S., & Tinits, P. (2020a). Language evolution research in the year 2020: A survey of new directions. *Language Dynamics and Change*, 10(1), 3-26. DOI: 10.1163/22105832-bja10005

Nölle, J., Kirby, S., Culbertson, J., & Smith, K. (2020b). Does Environment Shape Spatial Language? A Virtual Reality Experiment. *LANGUAGE* of, 321. DOI: 10.17617/2.3190925

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition*, 181, 93-104. DOI: 10.1016/j.cognition.2018.08.014

Noordzij, M. L., Newman-Norlund, S. E., De Ruiter, J. P., Hagoort, P., Levinson, S. C., & Toni, I. (2009). Brain mechanisms underlying human communication. *Frontiers in Human Neuroscience*, 3, 478. DOI: 10.3389/neuro.09.014.2009

Nota, N., Trujillo, J. P., & Holler, J. (2021). Facial signals and social actions in multimodal face-to-face interaction. *Brain Sciences*, 11(8), 1017. DOI: 10.3390/brainsci11081017

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive science*, 33(1), 127-146. DOI: 10.1111/j.1551-6709.2008.01007.x

Ohala, J. J., Hinton, L., & Nichols, J. (1997). Sound symbolism. In *Proc. 4th Seoul International Conference on Linguistics [SICOL]* (pp. 98-103).

Parise, C. V., & Pavani, F. (2011). Evidence of sound symbolism in simple vocalizations. *Experimental Brain Research*, 214, 373-380. DOI: 10.1007/s00221-011-2836-3

Peeters, D. (2019). Virtual reality: A game-changing method for the language sciences. *Psychonomic bulletin & review*, 26, 894-900. DOI: 10.3758/s13423-019-01571-3

Perlman, M. (2017). Debunking two myths against vocal origins of language: language is iconic and multimodal to the core. *Interaction Studies*, 18(3), 376-401. DOI: https://doi.org/10.1075/is.18.3.05per

Perlman, M., & Cain, A. A. (2014). Iconicity in vocalization, comparisons with gesture, and implications for theories on the evolution of language. *Gesture*, 14(3), 320-350. DOI: 10.1075/gest.14.3.03per

Perlman, M., Clark, N., & Johansson Falck, M. (2015a). Iconic prosody in story reading. *Cognitive Science*, 39(6), 1348-1368. DOI: 10.1111/cogs.12190

Perlman, M., Dale, R., & Lupyan, G. (2015b). Iconicity can ground the creation of vocal symbols. *Royal Society open science*, 2(8), 150152. DOI: 10.1098/rsos.150152

Perlman, M., Little, H., Thompson, B., & Thompson, R. L. (2018). Iconicity in Signed and Spoken Vocabulary: A Comparison Between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in Psychology*, 9, 1433. DOI: 10.3389/fpsyg.2018.01433

Perlman, M., Paul, J., & Lupyan, G. (2022). Vocal communication of magnitude across language, age, and auditory experience. *Journal of Experimental Psychology: General*, 151(4), 885. DOI: 10.1037/xge0001103

Perniss, P. (2018). Why We Should Study Multimodal Language. *Frontiers in Psychology*, 9. DOI: 10.3389/fpsyg.2018.01109

Perniss, P., & Vigliocco, G. (2014). The bridge of iconicity: from a world of experience to the experience of language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651), 20130300. DOI: 10.1098/rstb.2013.0300

Perniss, P., Thompson, R. L., & Vigliocco, G. (2010). Iconicity as a general property of language: evidence from spoken and signed languages. *Frontiers in psychology*, 1, 227. DOI: 10.3389/fpsyg.2010.00227

Petitto, L. A., & Marentette, P. F. (1991). Babbling in the manual mode: Evidence for the ontogeny of language. *Science*, 251(5000), 1493-1496. DOI: 10.1126/science.2006424

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*, 13(4), 707-727. DOI: 10.1017/S0140525X00081061

Pisanski, K., & Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America*, 129(4), 2201-2212. DOI: 10.1121/1.3552866

Pollick, A. S., & De Waal, F. B. (2007). Ape gestures and language evolution. Proceedings of the *National Academy of Sciences*, 104(19), 8184-8189. DOI: 10.1073/pnas.0702624104

R Core Team (2023). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org.

Ratcliffe, J. M., & Nydam, M. L. (2008). Multimodal warning signals for a multiple predator world. *Nature*, 455(7209), 96-99. DOI: 10.1038/nature07087

Raviv, L., Meyer, A., & Lev-Ari, S. (2019a). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, 286(1907), 20191262. DOI: 10.1098/rspb.2019.1262

Raviv, L., Meyer, A., & Lev-Ari, S. (2019b). Compositional structure can emerge without generational transmission. *Cognition*, 182, 151-164. DOI: 10.1016/j.cognition.2018.09.010

Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The role of social network structure in the emergence of linguistic structure. Cognitive Science, 44(8), e12876. DOI: 10.1111/cogs.12876

Rigaill, L., Higham, J. P., Lee, P. C., Blin, A., & Garcia, C. (2013). Multimodal Sexual Signaling and Mating Behavior in Olive Baboons (P apio anubis). *American journal of primatology*, 75(7), 774-787. DOI: 10.1002/ajp.22154

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. *Cognition*, 141, 52-66. DOI: 10.1016/j.cognition.2015.04.001

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in cognitive sciences*, 14(9), 411-417. DOI: 10.1016/j.tics.2010.06.006

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. (2009). Signalling signalhood and the emergence of communication. *Cognition*, 113(2), 226-233. DOI: 10.1016/j.cognition.2009.08.009

Senghas, A., Kita, S., & Ozyurek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, 305(5691), 1779-1782. DOI: 10.1126/science.1100199

Shintel, H., & Nusbaum, H. C. (2007). The sound of motion in spoken language: visual information conveyed by acoustic properties of speech. *Cognition* 105, 681–690. DOI: 10.1016/j.cognition.2006.11.005

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language*, 55(2), 167-177. DOI: 10.1016/j.jml.2006.03.002

Sidhu, D. M., Westbury, C., Hollis, G., & Pexman, P. M. (2021). Sound symbolism shapes the English language: the maluma/takete effect in English nouns. *Psychon. Bull. Rev.* 28, 1390–1398. DOI: 10.3758/s13423-021-01883-3

Slocombe, K. E., Waller, B. M., & Liebal, K. (2011). The language void: the need for multimodality in primate communication research. *Animal Behaviour*, 81(5), 919-924. DOI: 10.1016/j.anbehav.2011.02.002

Taremaa, P., Kiik, J., Toots, L. K., & Veismann, A. (2022). Speed as a dimension of manner in Estonian frog stories. *Nordic Journal of Linguistics*, 1-30. DOI: 10.1017/S0332586522000245

Teixeira, J. P., & Fernandes, P. O. (2014). Jitter, shimmer and HNR classification within gender, tones and vowels in healthy voices. *Procedia Technology*, 16, 1228–1237. DOI: 10.1016/j.protcy.2014.10.138

Tomasello, M. (2010). Origins of human communication. *MIT press*.

Verhoef, T., Kirby, S., & De Boer, B. (2016). Iconicity and the emergence of combinatorial structure in language. *Cognitive science*, 40(8), 1969-1994. DOI: 10.1111/cogs.12326

Winter, B., Sóskuthy, M., Perlman, M., & Dingemanse, M. (2022). Trilled/r/is associated with roughness, linking sound and touch across spoken languages. *Scientific Reports*, 12(1), 1035. DOI: 10.1038/s41598-021-04311-7

Zeileis, A. & Hothorn, T. (2002). "Diagnostic Checking in Regression Relationships." *R News*, 2(3), 7–10. https://CRAN.R-project.org/doc/Rnews.

Appendices

A. Praat scripts

1. Praat script for extracting vocalization duration, F0, intensity, harmonics-to-noise ratio, and number of syllables

```
########################################################################
## Extract total duration, F0, intensity, harmonicity, and number of
syllables
## (c) Kotryna Motiekaityte, (2023)
## Template for directory creation and checking (c) Hans Rutger Bosker,
(2022)
## This script iterates through every pair of audio and TextGrid annotation
files;
## it assumes that audio and TextGrid file pairs have the same name;
## and are located in the same directory.
## The script extracts F0, duration, intensity, and harmonicity from .wav
files;
## and the number of intervals from the syllabic annotation tier in .TextGrid
files;
## then saves the extracted information into a .csv file.
########################################################################

########################################################################
## Specify input and output directories
########################################################################

dir_in$ = "input directory"


dir_out$ = "output directory"

## Checking whether input directory exists
## If input directory ends with a backflash, the script removes it

if right$(dir_in$,1)="/"
    dir_in$ = left$(dir_in$,length(dir_in$)-1)
elsif right$(dir_in$,1)="\"
    dir_in$ = left$(dir_in$,length(dir_in$)-1)
endif

########################################################################
##  SCRIPT
########################################################################

## Create the output file and write the first line (header).
outPath$ = "'dir_out$'" + "/" + "accoustic_measurements.csv"

sep$ = tab$

header$ = "File" + sep$
    ...+ "Duration (s)" + sep$
    ...+ "F0 (Hz)" + sep$
    ...+ "Intensity (dB)" + sep$
    ...+ "Harmonicity (dB)" + sep$
    ...+ "Number of syllables"

writeFileLine: outPath$, header$
```

```
## Create a list of file pairs
strings = Create Strings as file list: "fileList", dir_in$ + "/*.wav"
numberOfFiles = Get number of strings

## Iterating through the list of files
for i from 1 to numberOfFiles
    selectObject: strings
    fileName$ = Get string: i

    ## Open WAV files
    Read from file: dir_in$ + "/" + fileName$
    soundName$ = selected$: "Sound"

    ## Total duration
    select Sound 'soundName$'
    dur$ = Get total duration

    ## F0
    select Sound 'soundName$'
    object_f0 = To Pitch: 0, 75, 600
    f0$ = Get mean: 0, 0, "Hertz"
    removeObject: object_f0

    ## Intensity
    select Sound 'soundName$'
    object_intens = To Intensity: 100, 0, "yes"
    intensivity$ = Get mean: 0, 0, "dB"
    removeObject: object_intens

    ## Harmonics-to=noise ratio
    select Sound 'soundName$'
    object_harm = To Harmonicity (cc): 0.01, 75, 0.1, 1
    harmonicity$ = Get mean: 0, 0
    removeObject: object_harm

    ## Open TextGrid files
    Read from file: dir_in$ + "/" + soundName$ + ".TextGrid"
    tgName$ = selected$: "TextGrid"

    ## Get number of syllables
    select TextGrid 'tgName$'
    tierNumber = 6
    no_syllables$ = Get number of intervals: tierNumber

    ## Append output to the .csv file
    dataRow$ = fileName$ + sep$
        ...+ dur$ + sep$
        ...+ f0$ + sep$
        ...+ intensivity$ + sep$
        ...+ harmonicity$ + sep$
        ...+ no_syllables$

    appendFileLine: outPath$, dataRow$

endfor

writeInfoLine: "Done! Check the spreadsheet."

## Cleaning up
select all
Remove
```

```
#############################################################################
## End of script
#############################################################################
```

2. Praat script for extracting orthographic annotations

```
#############################################################################
## Extract orthographic annotations
## (c) Kotryna Motiekaityte, (2023)
## Template for directory creation and checking (c) Hans Rutger Bosker,
(2022)
## This script iterates through all TextGrid files with annotations;
## The script extracts orthographic annotations (for my files: located in
tier 1);
## then saves the extracted information into a .csv file.
#############################################################################

#############################################################################
## Specify input and output directories
#############################################################################

dir_in$ = "input directory"

dir_out$ = "output directory"

## Checking whether input directory exists
## If input directory ends with a backflash, the script removes it

if right$(dir_in$,1)="/"
    dir_in$ = left$(dir_in$,length(dir_in$)-1)
elsif right$(dir_in$,1)="\"
    dir_in$ = left$(dir_in$,length(dir_in$)-1)
endif

#############################################################################
##  SCRIPT
#############################################################################

## Create the output file and write the first line (header).
outPath$ = "'dir_out$'" + "/" + "annotations.csv"

sep$ = tab$

header$ = "File" + sep$
    ...+ "Annotation"

writeFileLine: outPath$, header$

## Create a list of file pairs
strings = Create Strings as file list: "fileList", dir_in$ + "/*.TextGrid"
numberOfFiles = Get number of strings

## Iterating through the list of files
for i from 1 to numberOfFiles
    selectObject: strings
    fileName$ = Get string: i

    ## Open tg files
    Read from file: dir_in$ + "/" + fileName$
```

```
    tgName$ = selected$: "TextGrid"

    ## Get annotations
    select TextGrid 'tgName$'
    tierNumber = 1
    numberOfIntervals = Get number of intervals: tierNumber

    for intervalNumber from 1 to numberOfIntervals
       annotation$ = Get label of interval: 1, intervalNumber
    endfor

    ## Append output to the .csv file
    dataRow$ = fileName$ + sep$
        ...+ annotation$

    appendFileLine: outPath$, dataRow$

endfor

writeInfoLine: "Done! Check the spreadsheet."

## Cleaning up
select all
Remove

################################################################################
## End of script
################################################################################
```

B. Models for statistical analysis

Communicative success

Model 1: Communicative success

Accuracy ~ centered.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape)

	Estimate	Std. Error	t-value	p-value
(Intercept)	0.590308369	0.089441109	6.59996702	2.39508E-06
Total trial number	0.002933877	0.000767682	3.82173549	0.000140781
Condition	-0.063099167	0.170061455	-0.3710374	0.715264537
Target size	0.047773411	0.040583634	1.17715952	0.239414772
Target movement	-0.052164866	0.040585667	-1.2853027	0.198987129
Target speed	0.011834179	0.024867081	0.4758974	0.634252473
Total trial nr X Condition	-0.000532224	0.001565017	-0.3400754	0.733871851
Total trial nr X Target size	0.00116167	0.001095345	1.06055106	0.289152556
Condition X Target size	-0.020736067	0.081177395	-0.2554414	0.798435259
Total trial nr X Target movement	0.000548197	0.001094539	0.50084755	0.616589725
Condition X Target movement	0.008804988	0.08118383	0.10845741	0.91365484
Target size X Target movement	-0.001523456	0.057397957	-0.026542	0.978830386
Total trial nr X Condition X Target size	-0.001427218	0.002202807	-0.6479086	0.517194254
Total trial nr X Condition X Target movement	0.001258082	0.002212171	0.56870946	0.569682293
Total trial nr X Target movement X Target size	-0.001468864	0.001555494	-0.944307	0.345243129
Condition X Target size X Target movement	-0.005296423	0.114810352	-0.0461319	0.963214399
Total trial nr X Condition X Target size X Target movement	0.003349574	0.003111046	1.07667134	0.281889356

Iconicity

Model 2: Iconic mappings on vocalization duration

Duration ~ centered.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape)

	Estimate	Std. Error	t-value	p-value
(Intercept)	1.662993599	0.516528729	3.219556833	0.005939762
Total trial number	0.003923412	0.001896273	2.069012438	0.038803731
Condition	-1.021169425	1.031214971	-0.990258533	0.338392075
Target size	0.264185255	0.100131554	2.638381667	0.008460891
Target movement	0.265322579	0.100136756	2.649602306	0.008186952
Target speed	0.341080208	0.061340792	5.560414132	3.46065E-08

Total trial nr X Condition	-0.001476303	0.003857994	-0.382660941	0.702053233
Total trial nr X Target size	0.002685861	0.002701537	0.994197698	0.320369313
Condition X Target size	-0.249422414	0.200689795	-1.242825595	0.214226085
Total trial nr X Target movement	0.000972972	0.002700097	0.360347222	0.718664275
Condition X Target movement	-0.284723682	0.200112491	-1.422818138	0.155103584
Target size X Target movement	-0.276647577	0.141527028	-1.954733176	0.050895423
Total trial nr X Condition X Target size	-0.002807952	0.005430445	-0.517075767	0.605218647
Total trial nr X Condition X Target movement	-0.002655378	0.005452399	-0.48701098	0.626358247
Total trial nr X Target movement X Target size	-0.001875289	0.003834875	-0.489009093	0.624943462
Condition X Target size X Target movement	0.240463964	0.283374058	0.848574375	0.396323091
Total trial nr X Condition X Target size X Target movement	0.004164711	0.007667208	0.543184829	0.587124701

Model 3: Iconic mappings on pitch (F0)
F0 ~ centered.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape)

	Estimate	Std. Error	t-value	p-value
(Intercept)	225.8464	14.8171284	15.2422516	4.5459E-12
Total trial number	-0.0260003	0.10559219	-0.2462334	0.80555416
Condition	-26.58007	27.3395318	-0.9722211	0.34551344
Target size	17.9731321	5.60885592	3.20442036	0.00139792
Target movement	-8.0818283	5.63192819	-1.435002	0.15161188
Target speed	-13.790943	3.43199252	-4.0183489	6.3187E-05
Total trial nr X Condition	-0.1204875	0.21346324	-0.5644413	0.57258555
Total trial nr X Target size	0.15571188	0.15123723	1.02958698	0.3034629
Condition X Target size	39.9431214	11.1192727	3.59224228	0.00034446
Total trial nr X Target movement	0.10577365	0.15169013	0.69730082	0.48578318

Condition X Target movement	-10.376345	11.0988737	-0.9349007	0.35007432
Target size X Target movement	-4.836796	7.95462741	-0.6080481	0.54329925
Total trial nr X Condition X Target size	0.43679506	0.30080189	1.45210208	0.14679955
Total trial nr X Condition X Target movement	-0.0547794	0.30233354	-0.1811885	0.85625781
Total trial nr X Target movement X Target size	-0.1020169	0.2157255	-0.4729013	0.636391
Condition X Target size X Target movement	-8.1323367	15.7141209	-0.5175178	0.60491388
Total trial nr X Condition X Target size X Target movement	0.15332594	0.42530051	0.36051201	0.7185435

Model 4: Iconic mappings on loudness (intensity)

Intensity ~ centered.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape)

	Estimate	Std. Error	t-value	p-value
(Intercept)	52.5966471	1.04860192	50.1588317	1.2063E-21
Total trial number	0.01387409	0.00822186	1.68746499	0.09182958
Condition	3.61344633	1.8534437	1.94958516	0.06830556
Target size	-1.215478	0.43413904	-2.7997437	0.00521398
Target movement	-0.4358632	0.4341615	-1.0039195	0.31566323
Target speed	-1.7816333	0.2659539	-6.6990305	3.5173E-11
Total trial nr X Condition	-0.0100668	0.01673148	-0.6016688	0.54753266
Total trial nr X Target size	-0.0081024	0.01171623	-0.6915496	0.48938269
Condition X Target size	1.24823371	0.87013089	1.43453557	0.15173592
Total trial nr X Target movement	0.00670273	0.01170739	0.57252182	0.56709874
Condition X Target movement	-0.7235001	0.86762279	-0.8338879	0.40454574
Target size X Target movement	0.31559084	0.61361554	0.51431363	0.60714775
Total trial nr X Condition X Target size	-0.0082735	0.02354811	-0.351343	0.72540595
Total trial nr X Condition X Target movement	-0.0154522	0.02364162	-0.6536021	0.51352023
Total trial nr X Target movement X Target size	-0.0081602	0.01662686	-0.4907857	0.62368699
Condition X Target size X Target movement	-0.0383917	1.22862154	-0.0312478	0.97507824
Total trial nr X Condition X Target size X Target movement	0.02468255	0.03324525	0.74243829	0.45799828

Model 5: Iconic mappings on harmonics-to-noise ratio

HNR ~ centered.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape)

	Estimate	Std. Error	t-value	p-value
(Intercept)	10.4665122	0.72479087	14.440734	1.9245E-09
Total trial number	0.01366249	0.00728795	1.87466839	0.06113315
Condition	-0.3546543	1.09819067	-0.3229442	0.7499296

Target size	0.13784016	0.38483569	0.35817925	0.72028568
Target movement	-0.0950658	0.38485557	-0.2470168	0.80494652
Target speed	0.1792533	0.23575064	0.7603513	0.44722603
Total trial nr X Condition	0.00040327	0.01483086	0.02719136	0.9783126
Total trial nr X Target size	-0.0074361	0.01038549	-0.7160134	0.47415222
Condition X Target size	0.71722901	0.77131294	0.92988069	0.35265999
Total trial nr X Target movement	0.00118039	0.01037724	0.11374806	0.90946061
Condition X Target movement	0.46548567	0.76909096	0.60524138	0.54515761
Target size X Target movement	-0.0624681	0.54393002	-0.1148458	0.90859066
Total trial nr X Condition X Target size	0.00369809	0.0208724	0.17717609	0.85940637
Total trial nr X Condition X Target movement	-0.0061678	0.02095627	-0.2943175	0.76857708
Total trial nr X Target movement X Target size	-0.0011884	0.01473793	-0.0806336	0.93574968
Condition X Target size X Target movement	0.07591129	1.08909197	0.06970145	0.94444539
Total trial nr X Condition X Target size X Target movement	0.02322891	0.02946861	0.78825959	0.43073391

Model 6: Iconic mappings on the number of syllables per vocalization

SyllableNr ~ centered.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape)

	Estimate	Std. Error	t-value	p-value
(Intercept)	5.29464349	0.89965479	5.88519457	1.2113E-05
Total trial number	0.00354456	0.00663951	0.53385838	0.5935596
Condition	-1.076201	1.69192526	-0.6360807	0.53372333
Target size	0.3074866	0.35058564	0.87706558	0.38066423
Target movement	1.49545919	0.35060379	4.26538228	2.1872E-05
Target speed	-2.2985109	0.21476905	-10.702244	2.2907E-25
Total trial nr X Condition	0.00501489	0.01351125	0.37116445	0.7105946
Total trial nr X Target size	0.00864227	0.00946124	0.91343942	0.36123428
Condition X Target size	-0.3680068	0.70266749	-0.5237282	0.60058502
Total trial nr X Target movement	0.00435445	0.00945425	0.46058151	0.64520022
Condition X Target movement	-1.7752831	0.70064216	-2.5337942	0.01143708
Target size X Target movement	-0.1624559	0.49552055	-0.3278489	0.74309537
Total trial nr X Condition X Target size	-0.0137181	0.01901612	-0.7213943	0.47083761
Total trial nr X Condition X Target movement	-0.0065817	0.01909159	-0.344742	0.7303617
Total trial nr X Target movement X Target size	-0.0063902	0.01342697	-0.4759211	0.63423582
Condition X Target size X Target movement	-0.0785865	0.99216389	-0.0792072	0.93688384
Total trial nr X Condition X Target size X Target movement	0.01772678	0.02684696	0.66028999	0.50922164

Model 7: Iconic mappings on speech rate

SpeechRate ~ centered.TotalTrialNr * Condition * TargetSize * TargetMovement + TargetSpeed + (1|Participant) + (1|TargetItem) + (1|TargetShape)

	Estimate	Std. Error	t-value	p-value
(Intercept)	3.89731291	0.31845278	12.2382757	6.7492E-10
Total trial number	0.00095605	0.00264197	0.36186803	0.71752801
Condition	0.53497439	0.62855975	0.85111144	0.40702404
Target size	0.17524859	0.13950764	1.2561935	0.20934277
Target movement	0.21064422	0.13951487	1.50983346	0.13140565
Target speed	-1.0553495	0.0854626	-12.34867	1.1101E-32
Total trial nr X Condition	-0.001034	0.00537553	-0.1923467	0.84751013
Total trial nr X Target size	0.00054577	0.00376421	0.14498926	0.8847489
Condition X Target size	0.04389209	0.27960995	0.15697614	0.87529572
Total trial nr X Target movement	-0.0009443	0.00376189	-0.2510071	0.80186077
Condition X Target movement	-0.3046166	0.27880524	-1.0925786	0.27484519
Target size X Target movement	0.06818741	0.1971815	0.3458104	0.72955883
Total trial nr X Condition X Target size	-0.0042592	0.00756613	-0.5629287	0.57361104
Total trial nr X Condition X Target movement	0.00235684	0.00759665	0.31024721	0.75643839
Total trial nr X Target movement X Target size	-0.0032134	0.00534284	-0.6014379	0.54768629
Condition X Target size X Target movement	-0.4852478	0.39480909	-1.2290695	0.2193384
Total trial nr X Condition X Target size X Target movement	0.01005442	0.01068244	0.94120981	0.34682732

Compositional structure

Model 8: Compositional structure over time (raw scores)

CompositionalStructure (raw scores) ~ centered.BlockNr + (1|Participant)

	Estimate	Std. Error	t-value	p-value
(Intercept)	0.3700756	0.04795111	7.717770	0.000001336332
Block number	0.1964380	0.03794787	5.176522	0.000112707324

Model 9: Compositional structure over time (z-scores)

CompositionalStructure (z-scores) ~ centered.BlockNr + (1|Participant)

	Estimate	Std. Error	t-value	p-value
(Intercept)	8.6533924	1.1239987	7.69875658	0.000001376988
Block number	4.65861244	0.82769455	5.62841988	0.000048047526

Model 10: Compositional structure and communicative success (raw scores)

Accuracy ~ CompositionalStructure (raw scores) * centered.BlockNr + TargetSpeed + (1|Participant)

	Estimate	Std. Error	t-value	p-value
(Intercept)	0.24038104	0.07366966	3.26295832	0.00446562

Compositional structure	0.87418756	0.17064615	5.12280863	0.00003923511
Block number	-0.0681362	0.09781305	-0.6965958	0.50218744
Compositional structure X Block nr	0.27305121	0.22963938	1.18904345	0.26520728

Model 11: Compositional structure and communicative success (z-scores)
Accuracy ~ CompositionalStructure (z-scores) * centered.BlockNr + TargetSpeed + (1|Participant)

	Estimate	Std. Error	t-value	p-value
(Intercept)	0.24429885	0.07463416	3.27328471	0.00404972
Compositional structure	0.03703244	0.00741278	4.99575669	0.0000447258
Block number	-0.0611359	0.09032689	-0.6768297	0.51299521
Compositional structure X Block nr	0.01077649	0.00911692	1.18203189	0.26568289

Convergence

Model 12: Convergence over time
Convergence ~ centered.BlockNr + (1+ centered.BlockNr|Pair) + (1|TargetItem)

	Estimate	Std. Error	t-value	p-value
(Intercept)	0.45920333	0.09135133	5.0267831	0.00148885
Block number	0.23885169	0.05079398	4.70236249	0.00220171

C. Total averages of compositional structure scores for all pairs

Pair	Condition	Average compositional structure score (raw)	Average compositional structure score (z-score)
6	Vocalization-only	0.61782320	14.593238
11	Vocalization-only	0.25473650	5.815396
13	Vocalization-only	0.07653621	1.743450
15	Multimodal	0.17579329	4.398059
16	Vocalization-only	0.60487063	14.433549
20	Vocalization-only	0.38257318	9.156459
22	Multimodal	0.40460481	9.222756
23	Vocalization-only	0.44366714	9.864233