THE UNIVERSITY *of* EDINBURGH

MSc Evolution of Language and Cognition

Predicting compositionality in semantic space

by

Sigurd Farstad Iversen

2023

Declaration

This thesis is submitted in partial fulfilment of the requirements for the degree of MSc Evolution of Language and Cognition. I declare that this thesis was composed by myself, that the work contained therein is my own, except where explicitly stated otherwise in the text, and that it has not been submitted, in whole or in part, for any other degree or professional qualification.

Sigurd Farstad Iversen

Word Count: 7987 words

This thesis was conducted under the supervision of Prof. Kenny Smith and Elizabeth Pankratz.

Abstract

Compositionality is undeniably a fundamental feature of language, and arguably, uniquely human. For the last couple decades, compositionality has fruitfully been explored as a product of cultural evolution, and has been seen as solving the opposing selection pressures of expressivity and learnability (e.g., Kirby et al. 2008; Kirby et al. 2015). Further, popular perspectives have echoed that these pressures reflect specific processes of learning and use (Tamariz and Kirby, 2016) which are not necessary conditions for the emergence of compositionality. By contrast, a largely unexplored and undertheorised factor which affects the emergence of compositionality is the shape of the meaning space. In this thesis, I attempt to map out part of these spatial variables in artificial language learning, i.e., the structure of the meaning space, and how it relates to the emergence of compositionality. I investigate two previously proposed ways of structuring the (meaning) space: Smith et al. (2003)'s *structure* and Reeder et al. (2013)'s *overlap*. The former minimises the semantic difference between objects in the meaning space while the latter maximises equal distribution across it. Through an artificial language learning experiment, I put the two proposals against one another to see how the two structural configurations differ in the degree to which they allow for generalisation within a meaning space. The results, though inconclusive, indicate that generalisation and learning in general is easier in overlap spaces where semantic values are equally spread across. This could suggest that overlap facilitates the emergence of semantic combinatoriality, a prerequisite for compositionality, which in turn could arise as a consequence of specific parameters, e.g., number of semantic values and their relative degree of exposure. Along with these specific parameters, I suggest that future research should focus on how mechanisms other than generalisation contribute to different kinds of compositionality.

Acknowledgement

I have a lot to be thankful for. First of all, for my supervisor Prof. Kenny Smith who has been far more generous in his engagement and feedback than I could have expected from anyone, who has been a great teacher, and who was and still is an inspiration for my academic interest. My second supervisor, Elizabeth Pankratz, has been immensely helpful, especially in my lack of statistical competence. I am privileged to know her as a great linguist, and an even better friend, *chaya t'not*. I am also grateful for getting to know Prof. Jenny Culbertson and Dr. Shira Tal in their research, and for Frank Mollica and Prof. Simon Kirby for some of the best lectures of my life. To Iris, Leo, Megan, Serra, Salman, Fruzsina, Maya, Tanishq, Isaac, Alice, Reiss, Ashley, and Nadine, thank you for making my year. To those who came here, my family, and friends, Bård, Aurora, Halvor, Leo, Håkon, and Frode, *fy k'i he samnæ'åkkå*. Most of all I thank Urtei, who makes my life better in every way. *Bú er betra, þótt lítit sé...*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Part of what defines us as humans is the emergence of language. Defining language, however, is far less trivial. While some would mention aspects like metaphorical speech (e.g., Ellison and Reinöhl 2022), however, if any trait defines language, it is a structural one. One suggestion that has gained some ground is *compositionality* (e.g., Zuberbühler 2020, but see Hauser et al. 2002 for *merge*). This is the idea that the meaning of an expression derives from the meaning of its constituents, i.e., meaningful individual words in meaningful syntactic constructions. While a lot of human communication tends to be highly compositional, there cannot be 'turtles all the way down', and at the bottom, the constituents that make up a compositional expression must be *holistic*, i.e., an arbitrary constituent-to-meaning mapping.(Griffiths and Kalish, 2007, p.465). Dictionaries and theoretical views on *the lexicon* (e.g., Jackendoff 2003) are great examples of the holistic parts that make up language. Further, the vast majority of animal communication seems to be holistic (Zuberbühler, 2020), begging the question of why human language is compositional at all. While it may not always have been so (e.g., Tallerman 2007; Verhoef et al. 2012; Smith 2008), current theoretical explanations (e.g., Tamariz and Kirby 2016) suggest that compositional traits accumulate in response to two pressures: one to be learnable; and one to be expressive. As I will argue at the end of this chapter (see Section1.3), the latter assumes not just a sufficient number of meanings, but meanings structured in a particular way so as to make compositionality useful. If like vervet monkeys, the entire language consists of three alarm calls, each for a different predator,

having holistic calls might prove easier to learn. In this thesis, I will talk a lot about sets of linguistic units (e.g., words or meanings) as constituting *spaces*. I will particularly focus on how structural relationships between units within such spaces contribute to the emergence of compositionality. Like most of the research I refer to, my general framework is one of cultural evolution (see, e.g., Mesoudi et al. 2006) where compositionality is seen as an adaption to cultural pressures. Using an artificial language learning paradigm (see Chapter 2 and 3), I test two predictions of how spatial structure affects the emergence of compositionality. Subsequently, I discuss the implications of the results (see Section 4). To start off, I will elaborate on what consititues *compositionality*.

1.1 Compositionality

Consider the following definition:

> *"[X is compositional if t]he meaning of [X] is a function of the meanings of its parts and of the way they are syntactically combined."* (Partee, 1984, p.281)

In other words, we grasp the meaning of 'house pig', i.e., a pig normally kept in the house: (a) because we understand the words 'house' and 'pig'; and (b) because the way they combine (c.f., 'pig house'). Though similar notions have been around for over 1500 years (Pagin and Westerståhl, 2010), the definition above has its roots in philosophy, most often attributed to Gottlob Frege's *principle of compositionality* (Szabó n.d.; Garavaso 2018). Frege himself reportedly stated that "the structure of the sentence serves as an image of the structure of the thought" (Garavaso, 2018, p.106). In contrast to transformational ideas of semantics as*interpretive*, i.e., by definition deriving from a syntactic basis, Frege suggests that syntax is constrained by structures of thought, akin to the infamous generative semanticists (Katz, 1970). In this sense, compositionality is a purely semantic phenomenon, where *meanings* are compositional if derived systematically combined semantic units. Though modern accounts enters from similarly semantic angles (Pagin and Westerståhl, 2010), directly assessing a semantic system inside the mind is not currently possible. Hence, we require an assumption like Frege's that externalised language reflects internalised structural properties (see *competence* in Chomsky 1986).

Consequently, I will treat compositionality as either internalised (*i-compositionality*) or externalised (*e-compositionality*), the latter assuming an internalised equivalent (see Smith 2003 for similar terminology).

Further, I suggest that compositionality *can* be understood as *semantic combinatoriality*. Take for instance birdsong, broadly considered to display *combinatoriality* (Suzuki, 2014), i.e., systematic combination of acoustic notes (Zuidema and de Boer 2018; c.f., human phonology). In the debate on whether birds display *compositionality*, both sides agree that externalised syntax, i.e., the systematic combination of meaningful units, is needed (e.g., Suzuki et al. 2019; Bolhuis et al. 2018), the disagreement primarily concentrated in what is considered sufficient evidence. This is crucial because while compositionality, i.e., semantic combinatoriality, could theoretically exist without an external compositional language (Piantadosi et al., 2016), we rely on the use of a second combinatorial system of meaningful forms, i.e., syntax. Hence, though semantic combinatoriality is enough in an of itself *in theory*, a corresponding syntactic combinatoriality is necessary *in practice*. The presence of these two systems could be thought of as a kind of *duality of patterning* (Hockett 1960; de Boer et al. 2012). However, in the traditional sense, this consists of one combinatorial system of meaningful units (e.g., words, phrases, clauses), built on another combinatorial system of meaningless units (e.g., prosody, phonology, cherology or other; see, e.g., Edwards 2014). The latter system, however, is not necessary for compositionality (e.g., *compositional views* in Arbib 2012). Rather, the duality of patterning practically necessary to confirm compositionality is a regular mapping between a combinatorial semantic space and a combinatorial syntactic space. This mapping is not always one-to-one such that syntactically combinatorial idioms are often semantically holistic in natural language (e.g., 'kick the bucket' = DIE). Conversely, the the reverse is probably possible, where holistic forms are semantically combinatorial (e.g., 'bachelor' = *UNMARRIED* + *MAN*). However, this is difficult to assess. I will now turn to the contemporary discussion of how compositionality evolves.

1.2 Current debate

Emerging initially from a number of computational and mathematical models and simulations (Hurford 2000; Kirby 2000*a*; Kirby 2000*b*; Kirby 2002; Zuidema 2002; Brighton 2002; Smith et al. 2003; Kirby et al. 2004; Brighton et al. 2005; Kirby et al. 2007; Griffiths and Kalish 2007), and later spearheaded by two main studies with participants (Kirby et al. 2008; Kirby et al. 2015), supported by further findings (e.g., Theisen-White et al. 2011; Beckner et al. 2017; Guo et al. 2019), compositionality has become known as a linguistic property that emerges from "twin pressures of *expressivity* and *learnability*"(Spike et al., 2016, p.1, emphasis added)'. I offer the following definitions of these two abstract pressures:

> **Expressivity:**
>
> Linguistic properties that allow speakers to distinguish between meanings (of a certain number) accumulate over time.
>
> **Learnability:**
>
> Linguistic properties that are less cognitively straining accumulate over time.

Notice that the pressures essentially only point out that there is *some process* by which languages become expressive and learnable. Though this follows naturally from ideas of adaption in cultural evolution (Mesoudi and Thornton, 2018), the specific claims in theoretical discussions vary a lot. Some claims directly link compositionality to the abstract pressures, talking about languages being "under pressure to be simultaneously informative (so as to support effective communication) and simple (so as to minimize cognitive load)" (Kemp et al., 2018, p.111). However, often, the claims can narrow so that the emergence of compositionality is linked to the specific implementations of models and experiments. These claims shift the causal weight somewhat, so that compositionality arises from 'transmission', 'learning and communication' Tamariz and Kirby (2016), "repeated episodes of learning and production" (Smith et al., 2013, p.1348) or 'learning and use'(Smith 2022; Smith 2018).

The latter phrase can be understood in both a broad and a narrow sense. As mentioned in Smith (2022), *learning* favors simpler solutions and communicative *use* favors encod-

ing a useful set of distinctions, echoing the broad sense of the abstract pressures of learnability and expressivity respectively (see upper schematic in Figure 1.1). However, learning and use also corresponds to distinct experimental phases in Kirby et al. (2008) and Kirby et al. (2015), making them easily associated with a more narrow procedural sense. In Kirby et al. (2008), participants were asked to learn labels for images varying in shape (3), colour (3), and movement (3), i.e., 3x3x3 meaning space (see Figures 1.1 and 1.3). The participants were organised in diffusion chains so that the labels produced by one become the labels to be learned by another, effectively creating cultural *chains* of *generations* of participants. In the learning phase, participants were only given a random subset of the complete 3x3x3 space, while in the production phase, they had to produce labels for the entire space. Thus, a learnabiity pressure was simulated by it being impossible for them not to learn every image-label combination. Further, in their second experiment they filtered out any homonyms so that every distinct image always had a distinct label, simulating an expressivity pressure. Kirby et al. (2015) furthered this work with some modifications, mainly that instead of manually filtering out homonyms, they paired participants up in a communication game simulating a more natural pressure to have distinct labels (see Figures 1.1 and 1.3). Both found that compositionality emerged when both pressures where active [1]. Crucially, however, 'learning and use' are easily separable in these experiments. For 'learning', both implementations restricted learnability to a subset of meanings, and for 'use', both required distinct labels to be transmitted from one generation to another, ensuring expressivity. This narrower sense of 'learning and use', associated with a specific kind of transmission most likely goes back to Hurford's idea of the *arena of use* (1990; 2000) citing Anderson (1973, as illustrated in Figure 1.2. In the following section, I will show how this narrow sense can be somewhat problematic.

[1] They measured compositionality based on Brighton et al. (2005) looking at the degree to which the Hamming distance (1950) between two meanings mirrored the Levenstein distance (1966) between the equivalent label strings.

1.3 Problems

1.3.1 Sufficient but not necessary conditions

This high-level theoretical discourse, surrounding the significance of these two pressures, seems to suggest that the necessary conditions for the evolution of compositionality are the abstract twin pressures (learnability and expressivity), or more narrowly, specific processes of learning and use. owever, it would be a mistake to link the learnability pressure to its specific implementation, i.e., cultural transmission in the form of chains of participants required to produce more expressions than they are initially trained on. Rather, it is important to keep claims about the specific experimental implementation separate from claims about the linguistic processes it represents. Most theoretical discussions are good at pointing this separation, arguing, e.g., that "[i]n the real world, learning and use are, of course, not as clearly separated" (Smith, 2022, p.183). However, sometimes the narrow view prevails. For instance, Raviv et al. (2019) argue that Kirby et al. (2015) claim “that compositional languages can emerge *only* when languages are transmitted across multiple generations"(Raviv et al., 2019, p.151, emphasis added). They then go on to demonstrate how compositionality can emerge ‘in a single generation’ with a continuously expanding meaning space, i.e., introducing new meanings as the experiment progresses, thus simulating a learnability pressure without Kirby et al. (2015)’s diffusion chains (see also Selten and Warglien 2007; Winters et al. 2018). Raviv et al. (2019)’s mistake, or the mistake they claim Kirby et al. (2015) made, is the assumption that the pressures leading to compositionality requires a specific kind of transmission. Conversely, associating compositionality with specific pressures also causes the reverse, i.e., ignoring the causal contribution of other crucial aspects of the experimental design. In this sense, I argue that the structure of the meaning space is largely unexplored.

1.3.2 Exhaustive meaning spaces

Consider the meaning spaces in Figure 1.3. All these meaning spaces are taken from experiments supporting the idea that compositionality emerges under pressures of learning (or learnability) and use (or expressivity). However, most tend to have meaning spaces

bigger than 10 object-label pairs, with 2-3 *dimensions* (e.g., shape, colour) with 2-3 *values* each (square, blue). While this uniformity partially comes from artificial language learning being a fairly young (see Folia et al. 2010 for an overview), it largely arises from practical considerations, e.g., geometric stimuli being easier to generate than complex visual scenes. Yet, one common denominator I argue is particularly underrated is the exhaustive manner with which every possible combination of values in different dimensions fill the space. As seen in Figure 1.4, the meanings are structured in thematic groups. However, there are no explicit semantic relationships between a meaning from one group, and a meaning from another. However, in studies related to the emergence of compositionality, values and dimensions combine *exhaustively* such that the meaning space is maximally filled (see Figure 1.3). This matters for two reasons. First, the underlying semantic combinatoriality in the meaning space, is a prerequisite for compositionality (see Section 1.1). For instance, it would be impossible to make a compositional language out of the meaning space in Figure 1.4 unless the meaning space itself was restructured into something properly combinatorial. Second, and more importantly, natural languages do not look like this. Rather, they are irregular (e.g., Kirby 2001; Smith et al. 2023) and varied in where and how much compositionality there is.

1.4 Nonexhaustive meaning spaces

1.4.1 Natural complexity

Consider the following meaning space:

	Sheep	Dog
Happiness	“happy sheep"	“happy dog"
Fluffiness	“fluffy sheep"	“fluffy dog"

Table 1.1: Example of real compositional space

In Table 1.1, the meaning space, and the real English form space, are both compositional, i.e. there is both i- and e-compositionality. Now, imagine if I decided to talk about the largest and smallest breeds of various species, temporarily constructing a semantic space like Table 1.2:

	Sheep	Dog
Largest breed	Suffolk	Mastiff
Smallest breed	Ouessant	Chihuahua

Table 1.2: Example of a nonreal semantic space

However, the forms remain holistic. If put into an artificial language learning experiment, it might be that participants would develop a compositional system, but the real world meaning space in which these animal names exist, is far more complex. For instance, Suffolk sheep are thought of as far more than the largest sheep breed. Their coat pattern, wool texture, and lack of horns is far more distinct to any sheep farmer. Similarly, English Mastiffs have incredibly characteristic faces, Ouessants are known for their regional origin, and Chihuahuas are famous for their temperament and accessory function. Thus, a more natural meaning space, unlike the exhaustive meaning spaces in Figure 1.3, would look more like the following:

	Sheep	Dog	Rhinoceros
Four horns	Manx Loaghtan	???	???
Two horns	Suffolk	???	White rhino
One horn	???	???	Indian rhino
Zero horns	Adal	Chihuhua	???

Table 1.3: Slightly more realistic semantic space

In Table 1.3, though the values of the two dimensions do combine, they do not do so exhaustively. All compositional systems are dependent on a holistic base similar to a lexicon and so it is not surprising that names for sheep breeds will often be holistic. However, given that the degree to which a meaning space is filled is a gradient, at some point between Table 1.3 and the meaning spaces in Figure 1.3, compositionality should emerge. I will now turn to two studies that are relevant to answering this question.

1.4.2 Smith et al. (2003)

An early exploration of these structural traits in meaning spaces can be found in a study by Smith et al. (2003) which offers the following reflection:

> *"[I]n compositional languages the structure of signals reflects the structure of*

> *underlying meanings, and if the underlying meanings are unstructured then compositionality is immediately ruled out. Structured semantic representations therefore form a necessary, but not sufficient, condition for the cultural evolution of compositional language."* (Smith et al., 2003, p.545)

The *structure* they refer to here was implemented as a variable of a meaning space where the *average intermeaning Hamming distance* (henceforth AIHD) is minimised. For them, given a meaning space of 5x5x5 (three dimensions and five values), of which only a subset of meanings are used, the maximally structured meaning space forms a solid block, while the maximally unstructured resembles more of a randomly scattered cloud of meanings (see Figure 1.5). They present data with a thousand independent runs of an iterated learning model allowed to progress until a stable state using the spaces in Figure 1.5. Overall they find that though compositional languages are rare, high compositionality only occurs when the meaning space is *structured*. Further, they suggest that this arises as a consequence of the higher proportion of shared values in the structured spaces as compared to unstructured ones, which follows naturally from minimising the AIHD. However, their structured meaning spaces also include fewer values overall, i.e., less semantic complexity in general and not just that the average pair of meanings share more values. This is clear if we consider spaces (a) and (b) in Figure 1.5, where while (b)'s configuration needs nothing beyond a 2x2x3 space, (a) almost includes every value in the 5x5x5 space. To test whether compositionality is really ruled out in unstructured spaces, one would have to compare two spaces that are equal in how many values they display overall, but different in their AIHD. Further, another issue is the iterated learning paradigm. One could simply convert Smith et al. (2003)'s model to a very comparable iterated learning experiment with participants. However, though e-compositionality could evolve, i.e., the syntactic appearance of semantic combinatoriality, it is unclear whether a participant at any given time would develop an internalised compositional understanding of the meaning space. Rather, examining i-compositionality, i.e., real semantic combinatoriality, would only require a paradigm that simply trains participants on compositional forms from subsets of a shared compositional meaning space and then tests their ability to generalise to the rest of the meanings. The following study, happens to have such a

paradigm.

1.4.3 Reeder et al. (2013)

Reeder et al. (2013) published their study a decade after Smith et al. (2003), with the goal of seeing whether participants could form syntactic categories based purely on (meaningless) distributional information. Through a number of experiments, they tweak variables like Smith et al. (2003)'s *density*, but crucially, they also vary the structure of the space under the term *overlap*. The idea of 'overlap' comes from the observation that language users are sensitive to distributional information, suggesting that (syntactic) categories can be formed by "generaliz[ing] across [all] lexical items [in a space] (indicating that gaps [in shared distributional information] are likely to be accidental)"(Reeder et al. 2013, p.43). To explain overlap, they utilise a number neat schematics (see Figure 1.6). Through a series of experiments, they trained participants on different subsets of their syntax space, consisting of strings of letters functioning as (meaningless) words. Each string consisted of three words, where each word position could take a number of unrelated forms (three for Experiments 1-4). Consider Figure 1.7. They found that when participants are trained on a subset with complete overlap, where every word type is seen with every other word type, as seen in Figures 1.6 and 1.7a, they are more able to generalise to the entire space. However, when the training subset deviates from complete overlap, generalisation decreases[2] . I will now consider how these studies relate to one another.

1.5 Synthesis

Though Reeder et al. (2013)'s paradigm does not directly address the issue of how compositonality emerges, their overlap measure does target a core ability responsible from compositional understanding, i.e., the ability to generalise (e.g., baroni2020). Furthermore, they do this in an artificial language learning paradigm very similar, especially in its space design, to previous work on the emergence of compositionality (see Figure 1.3). Conversely, Smith et al. (2003) is not an artificial language learning experiment at all, but

[2]Generalisation was measured by grammaticality judgements, i.e., how acceptible they found a set of test strings.

a computation model. However, in common, both provide a measure of how to ideally structure a space for the desired combinatorial system to emerge. I offer the following definitions based on the authors' descriptions:

> **Structure:**
>
> *"[organising the subset of semantic units in a space] in such a way as to minimize the average intermeaning Hamming distance"* (Smith et al., 2003, p.549)

> **Overlap:**
>
> *Organising a subset of (semantic) units in a space in such a way as to disallow for subcategories larger than one unit to form.*

Interestingly, though not obvious from their definitions, it so happens that these measures are fairly at odds with one another. Consider Figure 1.8. Compare the cubes in Figure 1.5 from Smith et al. (2003)'s conditions, to my illustration of Reeder et al. (2013)'s condition in Figure 1.7, and Figure 1.8 which includes Smith et al. (2003)'s structure adapted to Reeder et al. (2013)'s space. First, since all the spaces consist of three dimensions, each schematic could be seen as a three-dimensional cube. Second, unlike Smith et al. (2003)'s spaces in Figure 1.5, the space conditions in Figure 1.8 includes the same number of features overall while still remaining either maximally structured or having complete overlap. Further, compared to the structured space (AIHD = 1.72), the complete overlap space is very unstructured, if not perfectly randomised to scatter over the entire space (AIHD = 2.22, however, c.f. partial overlap AIHD = 2.25). Conversely, a structured space like that in Figure 1.8 is far from complete overlap, with an overlap pattern best represented in Figure 1.9 (c.f., Figure 1.6). Though Smith et al. (2003) provide useful suggestions for the underlying mechanisms that give rise to compositionality (e.g., average shared values through AIHD), Reeder et al. (2013)'s discussion of overlap is somewhat more mysterious. They focus a lot on language users' perception of gaps, i.e. that the language users *perceive* the fact that X-words are not seen with specific A- and B-words, i.e., gaps. By contrast, in the complete overlap condition, where each X-word is seen at least once with every kind of A- and B-word, there are no gaps (see Figure 1.6). Thus, it seems that overlap does not make generalisation easier; rather, it makes subcategori-

sation impossible, so that if a generalisation occurs, it occurs across the entire space. Though both have been applied to facilitated the correct spread of *some* combinatorial system, this is from quite distinct angles, making them not necessarily mutually exclusive. However, as indicated by the discussions in Smith et al. (2003) and Reeder et al. (2013), each measure seems to represent the other's theorised opposite (either an unstructured space or a partial overlap space). Hence, one of the theories could have the wrong idea of exactly how their measure works (see Section4.4). The following chapter details the resulting experiment where I put the measures up against one another.

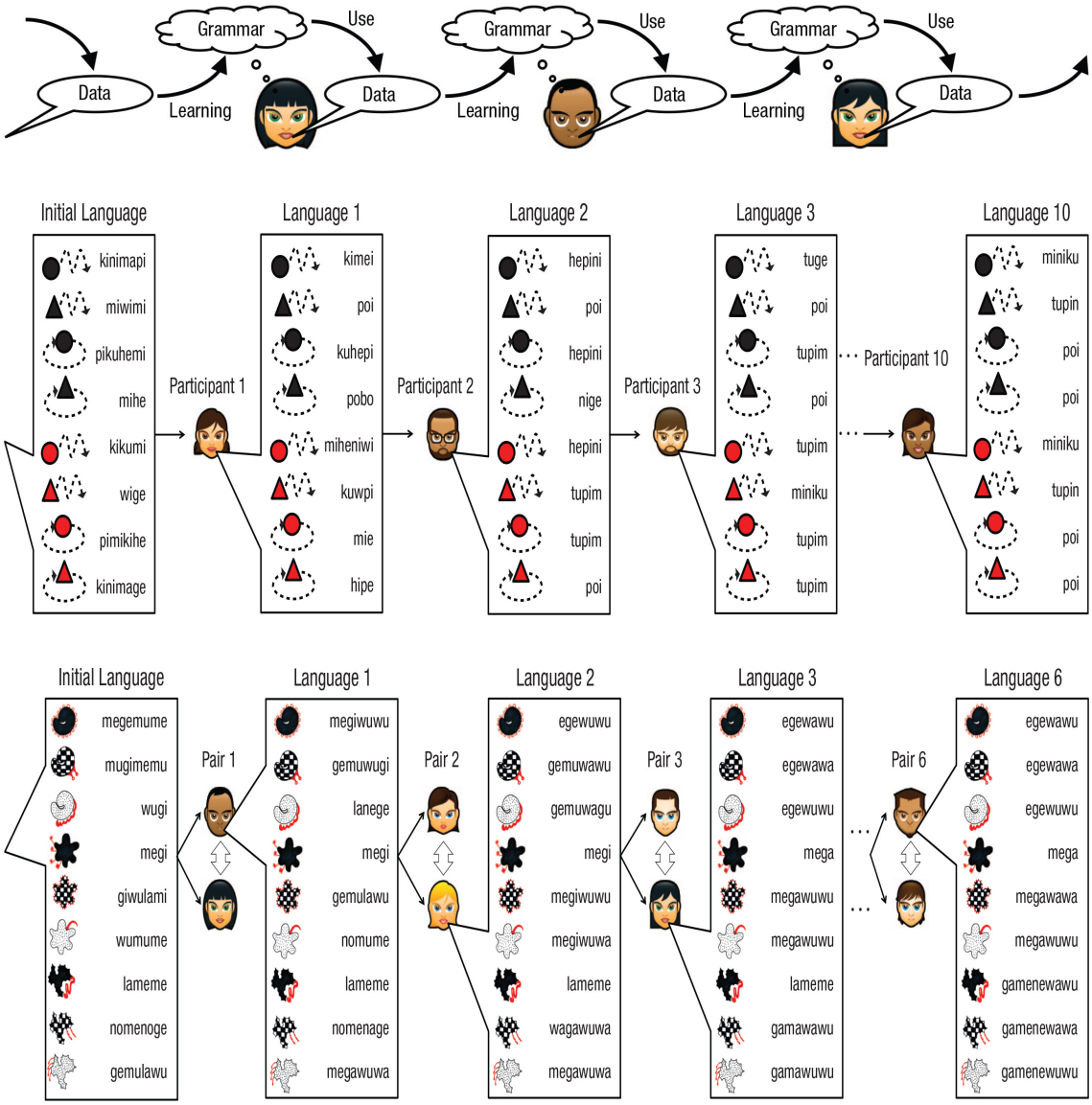
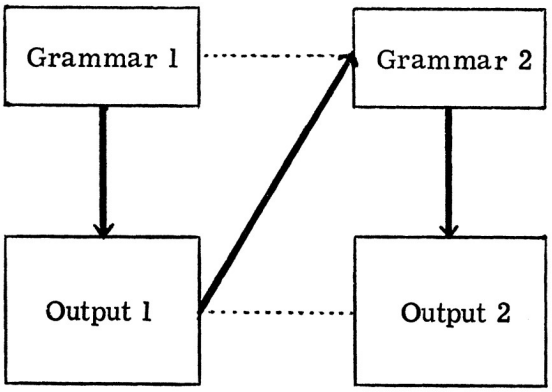
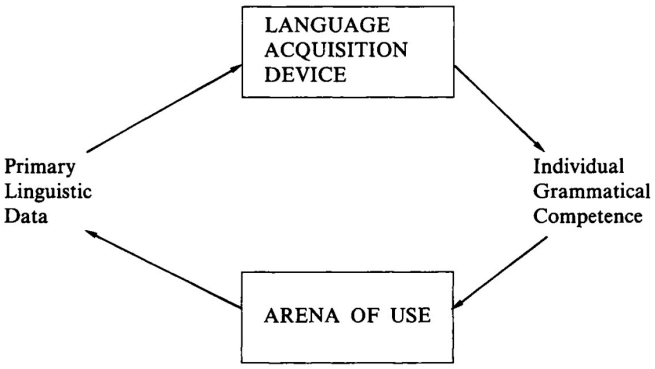


Figure 1.1: Illustrations of learning and use from Smith (2022). The upper schematic is a more abstract representation. The middle schematic is taken from Kirby et al. (2008), Experiment 1, Chain 4, where participants were asked to learn labels for colored moving shapes. The lower schematic is taken from Kirby et al. (2015) Chain Aa. In this experiment, participants learned labels for patterned shapes, and was paired up in a communication game. In both Kirby et al. (2008) and Kirby et al. (2015), the production of one generation was given as learning input for the next.

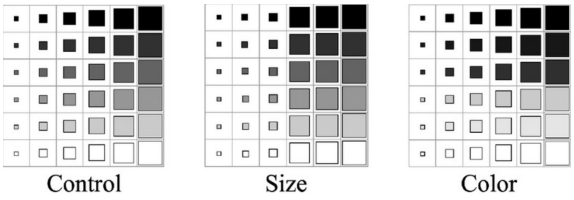


Andersen (1973)

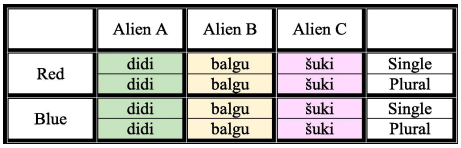


Hurford (1990)

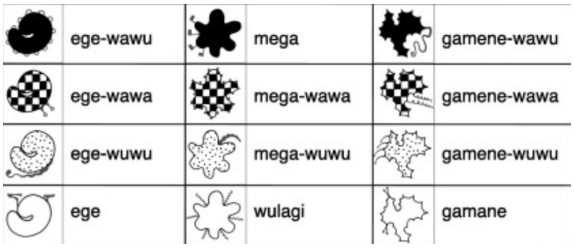
Figure 1.2: Early schematics of processes of learning and use in language. Hurford (1990) cite's Andersen (1973) as a source for his schematic. Both schematics implies a simple form of transmission.



Perfors and Navarro (2014)



Raviv et al. (2019)

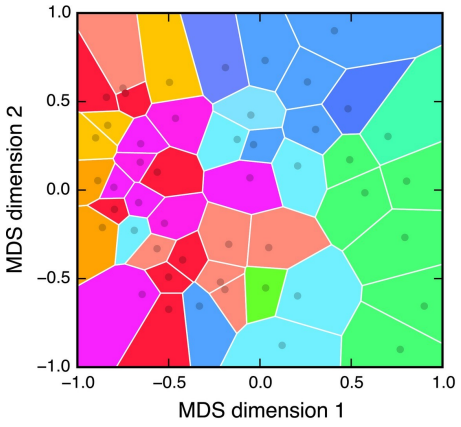


Kirby et al. (2015)

Kirby et al. (2008)

	'red'	'green'	'blue'	
'berry'	shen-to	shen-ta	shen-to	'1'
	shen-tra	shen-tro	shen-tra	'2'
	shen-trio	shen-trio	shen-trio	'3'
'key'	div-tro	div-tro	div-tro	'1'
	dev-tro	dev-tro	dev-etrio	'2'
	dev-stra	div-stra	dev-stra	'3'
'phone'	lolni-tro	lolni-tro	lolni-to	'1'
	lolne-stra	lolni-tro	lolne-stro	'2'
	lolni-tra	lolni-stra	lolni-stra	'3'

Beckner et al. (2017)



Carr et al. (2017)

Figure 1.3: Meaning spaces from various studies on compositionality. In all of these spaces, every value of every dimension combine *exhaustively*, i.e., every possible imaginable meaning within the space could be relevant to the participant.

Places	People	Entertainment	Objects	Abstract
Art Gallery	Arnold Schwarzenegger	Cartoon	Computer Monitor	Homesick
Parliament	Brad Pitt	Drama	Microwave	Loud
Museum	*Hugh Grant*	*Sci-Fi*	*Refrigerator*	Poverty
Theatre	Russell Crowe	Soap Opera	Television	*Sadness*

Figure 1.4: Schematic of meanings in Fay et al. (2010).

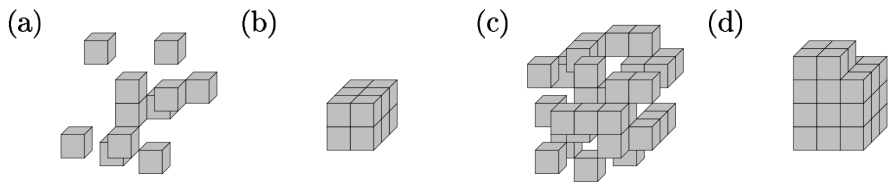


Figure 1.5: Meaning space variation according to 'structure' and 'density' in Smith et al. (2001). 'Structured' spaces, (b) and (d), follows the minimization of the AIHD while 'dense' spaces, (a) and (c), refers to the number of meanings in the space.

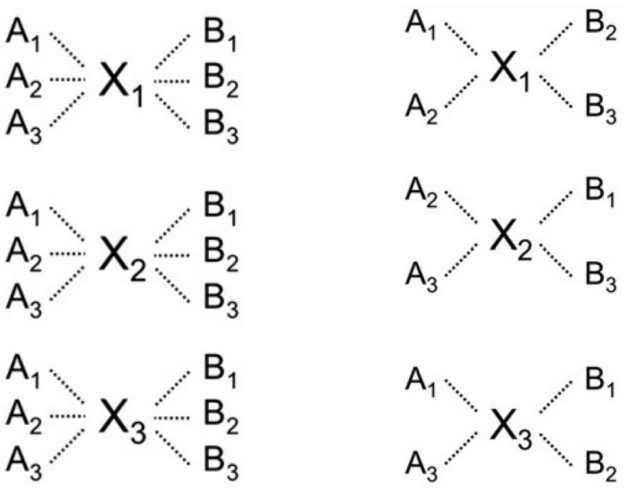
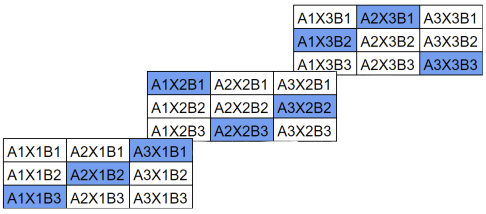
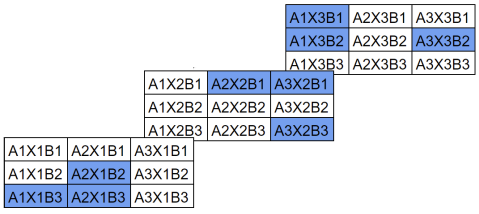


Figure 1.6: Degrees of overlap between X-words in Reeder et al. (2013)'s 3x3x3 syntax space in Experiments 1-4. A, X, and B, refer to the first, second, and third word position respectively. In complete overlap, every word in the X-position is seen with every word in the A position and every word in the B-position. In partial overlap, each X-word is specifically never seen with one word in each position.

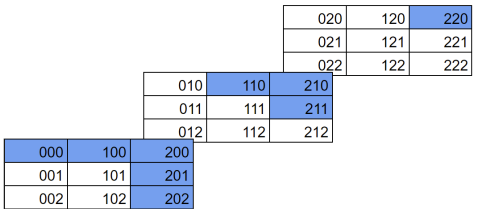


(a)
Complete overlap

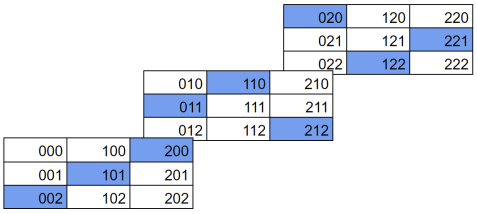


(b)
Partial overlap

Figure 1.7: Training subsets a 3x3x3 space converted from Table 2 in Reeder et al. (2013)

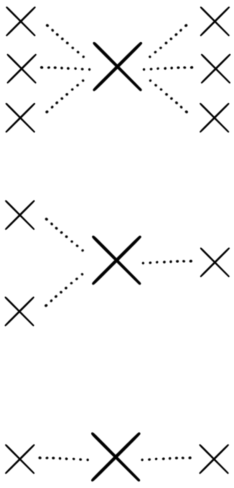


Structured space

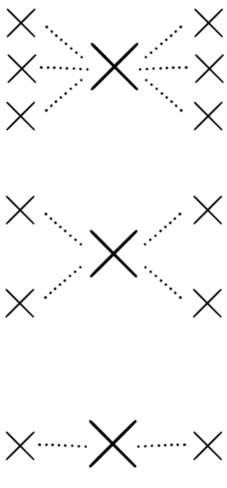


Overlap space

Figure 1.8: Experiment conditions, each based on subsets according to each measures in a 3x3x3 space. Semantic annotation converted directly from Reeder et al. (2013). See more on the semantic annotation in Figure **??**.



two of the three dimensions



one of the three dimensions

Figure 1.9: Contextual overlap pattern by feature in the structure condition. Here, a word type (X) of any given position will be seen with, either: every kind of the two other words around it (= 3/3, see the upper X-words); 1/3-2/3 of each word type (depending on the dimension, see middle X-words); or 1/3 of each word type (see lower X-words).

Chapter 2

Methods

To explore the effects Smith et al. (2003)'s *structure* and Reeder et al. (2013)'s *overlap* have on the development of language users' compositional understanding (i-compositionality), I designed an artificial language learning experiment where participants are exposed to perfectly compositional languages. They are asked "to learn a language aliens use to describe each other" consisting of image-label pairs, and are initially trained on subsets corresponding to the *structure* and *overlap* measures respectively, and later asked to produce labels for the entire space of images. I built the experiment using the jsPsych library of Javascript for designing behavioural experiments in a web browser (de Leeuw, 2016). Further, the initial scaffold of the experiment was based on a template by Smith (n.d.).

2.1 Participants

Participants were obtained through the online crowdsourcing platform Prolific, and completed the experiment on average after 20-35 minutes, for which each was paid 5.5£, aimed towards current minimum wage (https://www.gov.uk/national-minimum-wage-rates). Data from 52 participants were collected in total, 26 in the *overlap* condition and 26 in the *structure* condition. All participants self-reported to be over 18 in addition to having English as their native first language. Prolific also initially filtered for profiles that aligned with these criteria.

2.2 Materials

In the experiment, the artificial language consisted of two parts: (a) a set of objects; and (b) a set of individual labels mapped one-to-one to each object.

2.2.1 Semantic space

I obtained a set of objects from Jia Loy filtering out 27 illustrations of aliens varying in specific aspects. Consider the conditions in Figure 2.1:

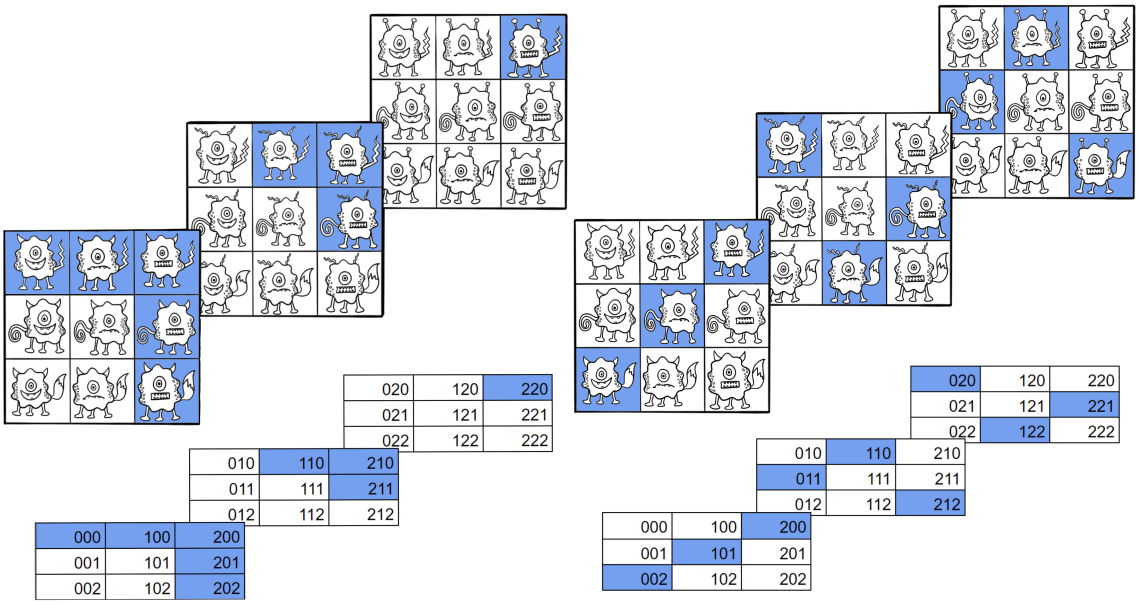


Figure 2.1: Meaning spaces, one with aliens, and one wiyth corresponding semantic annotation from Figure 1.8. The first number refers to mouth-values, the second headgear-values, and the third tail-values. The mapping was arbitrarily chosen.

As seen in the figure above, the aliens share similarities in body shape, eyes, feet, and skin. However, their mouths, headgear, and tails (i.e., the dimensions) vary systematically along the dimensions (see, e.g.,*smile*, *frown*, and *surprise*, in the vertical mouth dimension in Figure 2.1). There are three dimensions with three values each, i.e., 3x3x3 = 27 different aliens, directly applicable to the design in Reeder et al. (2013)[3]. The number of dimensions and values of the two conditions (see Figure 1.8) are taken from Reeder et al. (2013) as it was easier to apply structure, i.e., minimised AIHD, to a 3x3x3 space than to

[3]Initially, I planned to include a syntactically non-encoded fourth distractor dimension (c.f., Reeder et al. 2013's Experiment 5). However, due to priorities in time, this was not implemented.

apply overlap to Smith et al. (2003)'s 5x5x5 space. Notice that in contrast to Smith et al. (2003)'s structured space, my structured space also includes every kind of value along the three dimensions so that its effect can be distinguished from the effect of the specifically structured configuration of the space. This will also allow participants to generalise to the entire space (in theory).

2.2.2 "Label" space

To have labels be both parseable and alien, I utilised a syllable transition matrix (Pankratz, 2023), which counted the number of transitions in an English corpora between different syllables. Consider Table 2.1

	Dimension A	Dimension B	Dimension C
Value 1	za	xa	ja
Value 2	zi	xi	ji
Value 3	zu	xu	ju

Table 2.1: Syllables selected for general low attestation in English and specifically for low frequency of transitions.

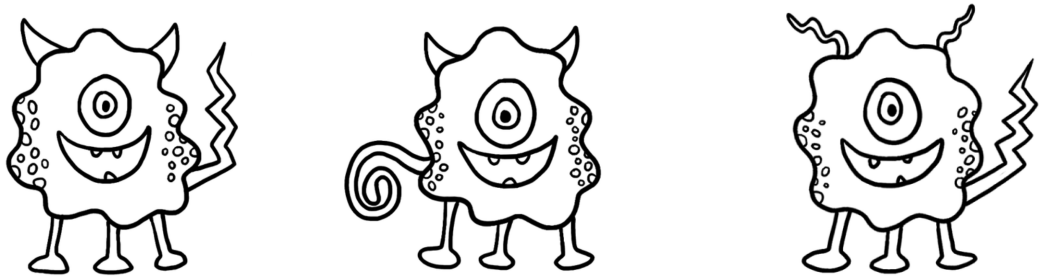
In Table 2.1, there are 9 syllables that can be combined into 27 trisyllabic words. Each of the syllables were found to be very infrequent but not absent from English, which was also the case for the syllable transitions ('xaja' and 'xiji' attested once). Additionally, the consonants were selected for sharing phonological aspects in spoken English (fricatives or affricates), and the vowels ('a', 'i', 'u'), and syllable pattern (CVCVCV) followed common typological systems (Gordon, 2016, Chapter 3-4). The below is the entire label space (27):

zaxaja, zaxaji, zaxaju, zaxija, zaxiji, zaxiju, zaxuja, zaxuji, zaxuju,
zixaja, zixaji, zixaju, zixija, zixiji, zixiju, zixuja, zixuji, zixuju,
zuxaja, zuxaji, zuxaju, zuxija, zuxiji, zuxiju, zuxuja, zuxuji, zuxuju

Now, I will turn to how these were mapped onto the aliens.

2.2.3 Pairing aliens to labels

To avoid potential concerns of a predetermined object-label mapping for all participants (e.g., Cuskley et al. 2017), the mapping was automatically randomised on two separate levels. First, each semantic dimension randomly corresponded to a syllable position characterised by an onset consonant (*z-*, *-x-*, and then *-j-*), yielding 6 different dimension mappings. Additionally, each value randomly corresponded to a vowel within a syllable position, yielding 6 different mappings. Thus, this totals 6x6x6x6x6x6 = 46656 potential languages (see examples in Figure 2.2). I will now turn to the procedure of the experiment as a whole.



000	001	010
-	-	-
TAIL-HEADGEAR-MOUTH	-	-
zixaja	zaxaja	**zi**xija
-	-	-
MOUTH-HEADGEAR-TAIL	-	-
zuxu**ji**	zuxuja	zuxa**ji**
-	-	-
HEADGEAR-TAIL-MOUTH	-	-
zu**xa**ji	zuxuji	za**xa**ji

Figure 2.2: Alien-label pairings with examples of different orders of semantic dimensions. The 'spike-tail'-value is underlined and in bold. The specific language is randomly generated.

2.3 Procedure

Having accepted the task listed on Prolific, the participants were be redirected to the experiment via a link. After some initial instructions, the participants started Round 1. Each round consisted of three blocks of the same learning trials: (1) *observation*; (2) *picture-selection*; and (3), *training production*. In every block, the participant interacted with the 9

image-label pairs from their specific condition (see Figure 2.1). In the observation block, participants were shown images of the alien with corresponding labels (see Figure 2.3). This was followed by a forced choice picture-selection block, where participants shown a label had to click on one of two randomly ordered aliens, one being correct and the other a foil selected randomly from the rest of the 26 aliens (see Figure 2.4). Then, in the training production block, participants were shown an alien and were asked to construct the correct label using 9 buttons, one for each syllable type. The syllable buttons were initially randomised for each participant but retained one order throughout the experiment (see Figure 2.5). Further, for every picture-selection and training production trials, participants were feedback ('CORRECT' or 'INCORRECT'), followed by exposure to the correct object-label pair, similar to observation trails. The learning phase consisted of 4 of these three-trial type rounds[4]. After the learning phase, a short testing phase ensued, where the participants were given production trials for the entire language (27). Here they were not given any feedback. Among other data, the experiment saved whether or not the picture chosen or label produced was correct (see Section 3.2.1), as well as which labels were produced exactly (see 3.2.2).

2.4 Hypothesis

Since it was unclear to me how exactly these two measures would affect the emergence of compositionality (see Section 1.5), I ended up with the following open hypothesis:

Hypothesis:

There will be a statistically significant difference between the conditions with regard to the ability to generalise.

[4] I initially ran a pilot where 9 participant had 3 of these three-trial type rounds. However, it seemed that most here struggled to learn anything at all (see Chapters 3 and 4), and hence I increased the round number.

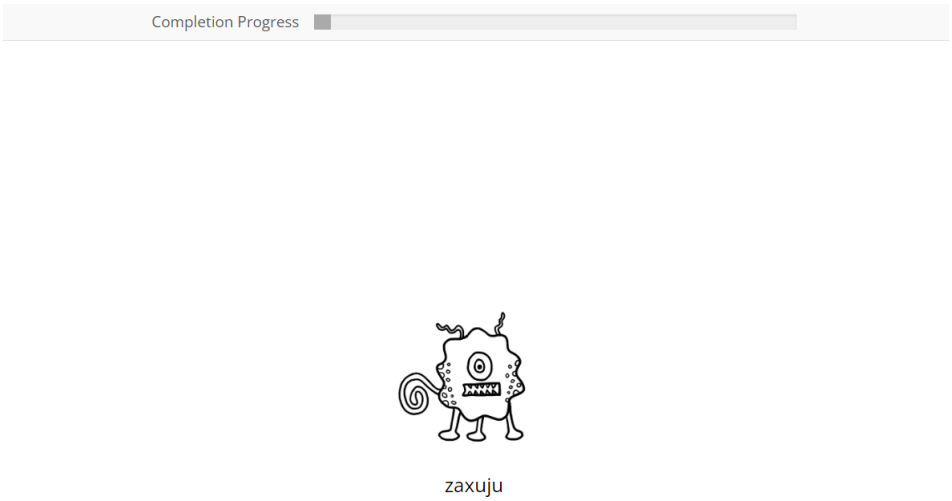


Figure 2.3: Screenshot from observation trial.

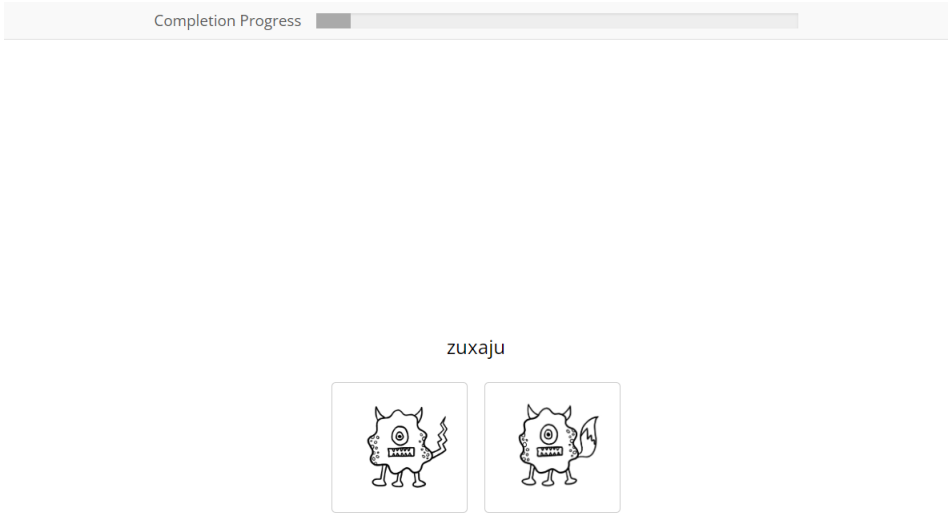


Figure 2.4: Screenshot from picture-selection trial.

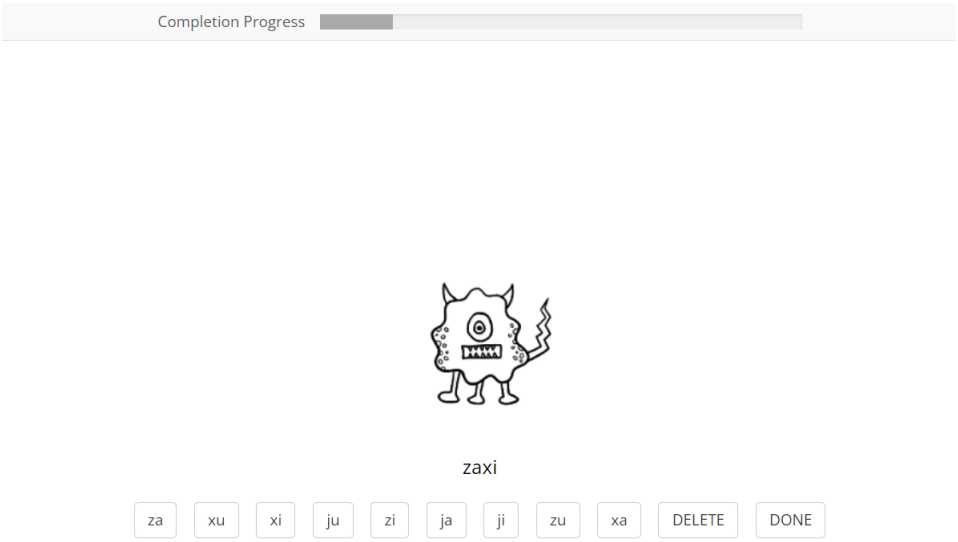


Figure 2.5: Screenshot from production trial.

Chapter 3

Results

To analyse the data, I utilise R (2023). I will mostly consider the data from the testing trials, but to get an initial idea of what the data looks like, and to see how much the participants learned in general, I will first consider the data from the learning phase.

3.1 Learning trials

The first measure I consider is *accuracy*, i.e., whether the label produced was exactly the correct label or not. As seen in picture-selection trials in Figure 3.1, participants in both conditions seem to get better at choosing the correct picture.

As shown in Figure 3.2, in the left plot, participants' mean accuracy increases over rounds. Yet, by the fourth round the mean production accuracy is still below 50% (see Chapter 4 for more on chance levels). However, it is interesting that participants in the overlap condition seem somewhat more proficient here.

In the plot to the right, I consider a second measure, i.e., edit distance between the labels the participants produced and the correct labels. Specifically, I use the Levenstein distance (1966) provided by the *stringdist* package (van der Loo, 2014) to account for the few times participants produced labels of different lengths. However, otherwise, Hamming (1950), Damerau-Levenstein (Brill and Moore, 2000), and Optimal String Alignment (Yujian and Bo, 2007) provided virtually identical numbers. Overall, for each round means Levenstein distance decrease indicating that participants produce labels generally closer

to the correct ones, with participants in the overlap condition appearing marginally mpre proficient. I will now turn to the testing trials.

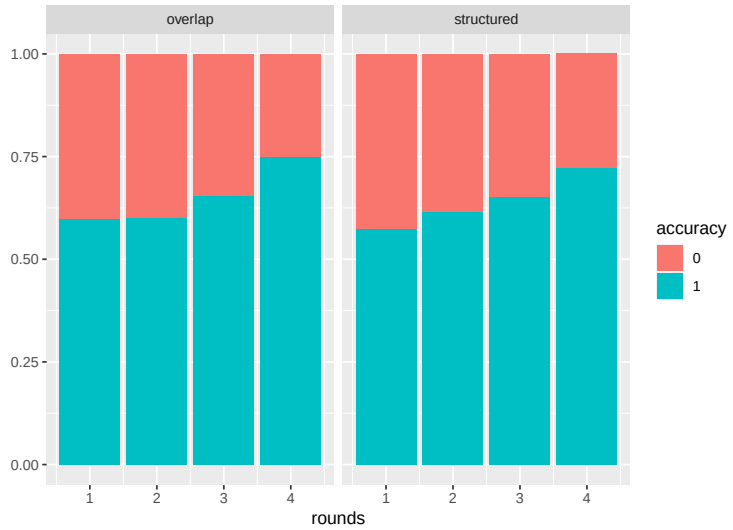
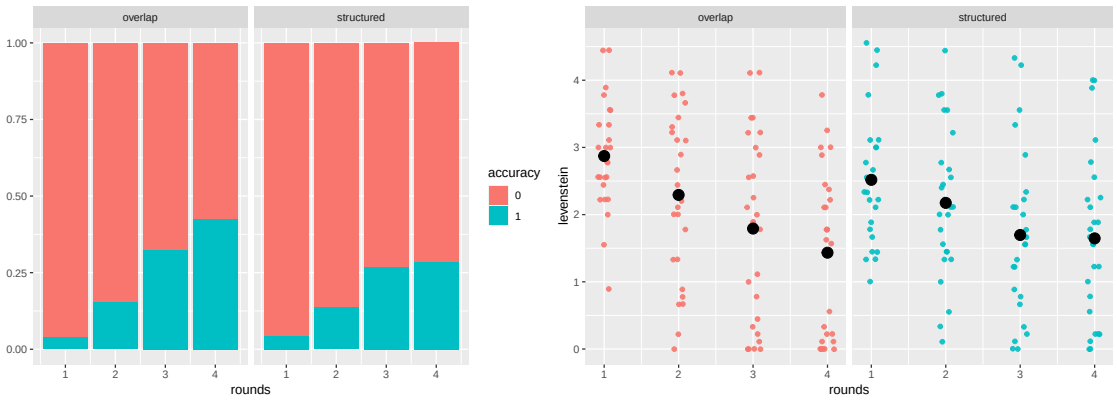


Figure 3.1: Amount of accurate picture selection choices for each round by condition.



Accuracy Levenstein distance

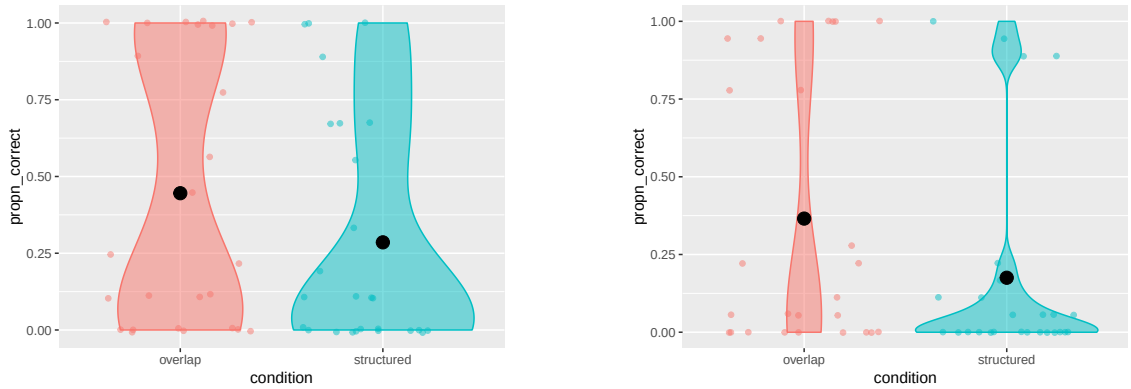
Figure 3.2: Means in training production. In the plot to the left, overall accuracy propotions for each round are shown by condition. In the plot to the right, the small dots show each participant's mean Levenstein distance from the correct label while the larger black dot show their overall mean.

3.2 Testing trials

3.2.1 Accuracy

First of all, what became immediately clear was that participants did not learn very well generally. As seen in Figure 3.3, generalising to novel aliens was more difficult and participants in the overlap condition did better overall, not because the did slightly better

all the time, but because a greater number of participants produced most of their labels accurately, while people in the structure condition failed completely more often.



Familiar aliens Novel aliens

Figure 3.3: Mean accuracy in each condition. Coloured dots represent each participant's mean while the black dots indicate the mean per condition.

Further, I fitted a linear model using a Bernoulli distribution (i.e., binomial, or logistic regression) to our measure 'accuracy', that is, whether the label produced was exactly the correct label (accuracy = 1) or not (accuracy = 0). As predictors, I included: 'condition', i.e., whether the participant that produced the label was trained on the overlap subset or the structure subset; and 'familiarity', i.e., whether the alien was familiar or novel. I included an interaction between condition and familiarity. The predictors were coded with treatment contrasts and the first level mentioned above was set as the reference level. To account for individual variance, I included random intercepts by participant [5]. According to the model (see Table 3.1), at baseline, i.e., when the participant labels a familiar alien having been trained on an overlap space, the log-odds of producing the correct label is –0.13, which is equivalent to a probability of 46.9%(SE = 0.92, z = -0.14, $p > 0.05$). The log-odds difference between the intercept and labelling a novel alien is -1.17, which when added to the baseline yields –1.30, equivalent to a probability of 21.4% of generalising correctly when a participant in the overlap condition labelled a novel alien (SE = 0.32, z = -3.70, $p < 0.05$). The log-odds difference between the intercept and the structure condition is –2.29, which when added to the baseline yields –2.41. This means that the probability of producing the correct label for familiar aliens in the structure condition is

[5]This is the maximal random effect structure possible, because the between-subjects design did not permit me to also include by-participant random slopes.

0.08% (SE = 1.36, z = -1.68, p $>$ 0.05). To figure out the the probability of producing a correct label for novel aliens for participants in the structured condition, I added all the estimates together, yielding -4.15 in log-odds, equivalent to 0.02% (SE = 0.37, z = -1.22, p $>$ 0.05). Only one difference had a probability of less than 5% of being a fluke, i.e. had a significant p-value, namely the difference between the labelling familiar and novel aliens in the overlap condition. The predictions of the model are shown in Figure 3.4.

Table 3.1: Summary of the accuracy model's estimates

	Estimate	Std. Error	P-value
Intercept: overlap+familiar	-0.13	0.92	0.89
Condition: structure+familiar	-2.29	1.36	0.09
Familiarity: overlap+novel	-1.17	0.32	<0.01
Interaction: structured*novel	-0.57	0.47	0.22

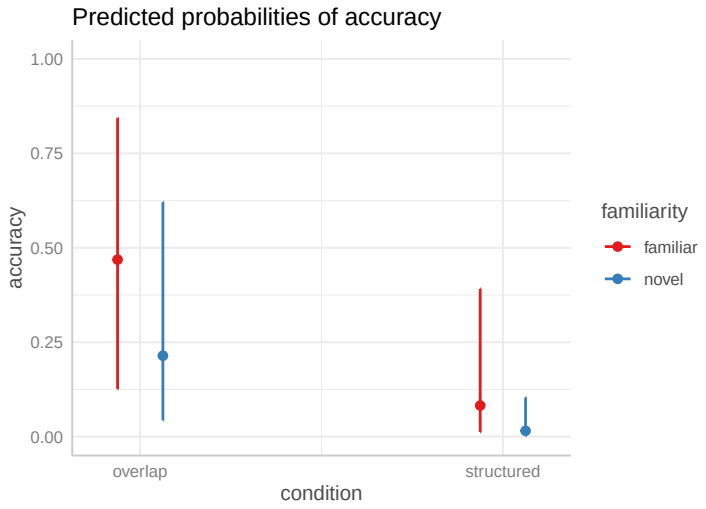
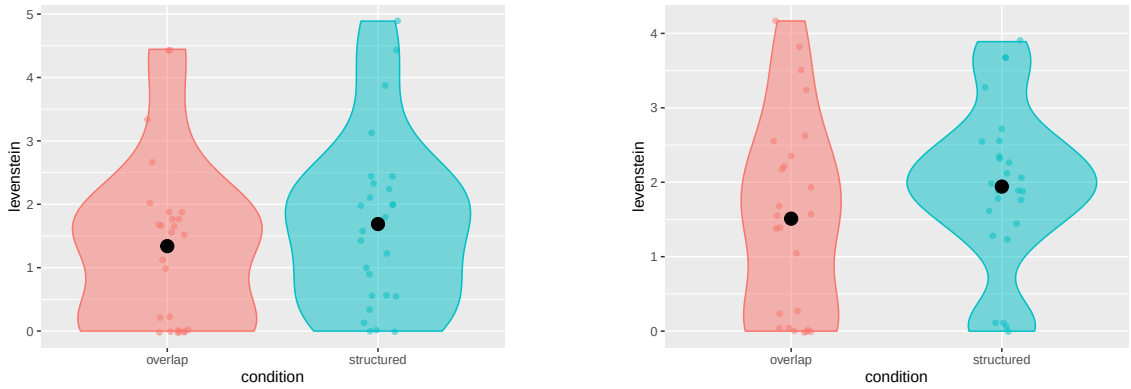


Figure 3.4: Predictions of the mean accuracy model with 95% confidence intervals.

3.2.2 String metrics

In Figure 3.5, two plots show Levenstein distances between the correct labels and the labels produced. Similar to accuracy (see Figure 3.3), participants' mean cluster more on longer distances in the structured condition and with novel aliens. This suggests that participants are somewhat better with familiar aliens and when they are trained on an overlap

space. I fitted a linear model with a Gaussian distribution to the log of the Levenstein distances between the correct labels and labels the participants produced. I used the log since the edit distance contains positive-only values (by definition), and therefore cannot be modelled with a Gaussian distribution. The 0-length edit distances were replaced with a reasonably low 0.1-length so as to transform properly. Again, as predictors, we included 'condition' and 'familiarity', random intercepts by participant, predictors being coded with treatment contrasts where the first level mentioned here was set as the reference level. P-values were obtained with the lmerTest package Kuznetsova et al. (2017) using the Satterthwaite's approximation of degrees of freedom. According to the model (see Table 3.2), when in the overlap condition labelling a familiar alien, the baseline mean log of the Levenstein distance is –0.61, equivalent to an edit distance of 0.54 (SE = 0.25, df = 52.87, t = –2.49, $p < 0.05$). The difference between the intercept and labelling a novel alien is 0.24, which added to the baseline, gives us -0.38, equivalent to an edit distance of 0.69 (SE = 0.06, df = 1337.02, t = 3.87, $p < 0.05$). The difference between the intercept and the structure condition is 0.47, which added to the baseline, gives us -0.15, equivalent to an average edit distance of 0.86 (SE = 0.25, df = 52.90, t = 1.34, $p > 0.05$) for a participant trained on a structured space labelling a familiar alien. To figure out the average edit distance for participants in the structured condition labelling novel aliens, I added all the estimates together, yielding 0.20 in log-odds, equivalent to 1.23 letters off from the correct label (SE = 0.09, z = 1337.02, t-value = 1.31, $p > 0.05$). These results show that though participants in the overlap condition generally seem to do produce labels closer to the correct label than that of their structure-trained counterparts, the difference is not significant. Again, the only difference that had a probability of less than 5% of being a fluke, i.e. had a significant p-value under 0.05, was between labelling familiar and novel aliens in the overlap condition. The predictions of the model are shown in Figure 3.4. Now, I will turn to the implications of these results.



Familiar aliens Novel aliens

Figure 3.5: Levenstein distances plotted by condition. The coloured dots represent each participant's mean while the black dots show mean per condition.

Table 3.2: Summary of the string metrics model's estimates

	Estimate	**Std. Err.**	**P-value**
Intercept: overlap+familiar	-0.61	0.25	0.02
Condition: structure+familiar	0.47	0.35	0.19
Familiarity: overlap+novel	0.24	0.06	<0.01
Interaction: structured*novel	0.11	0.09	0.19

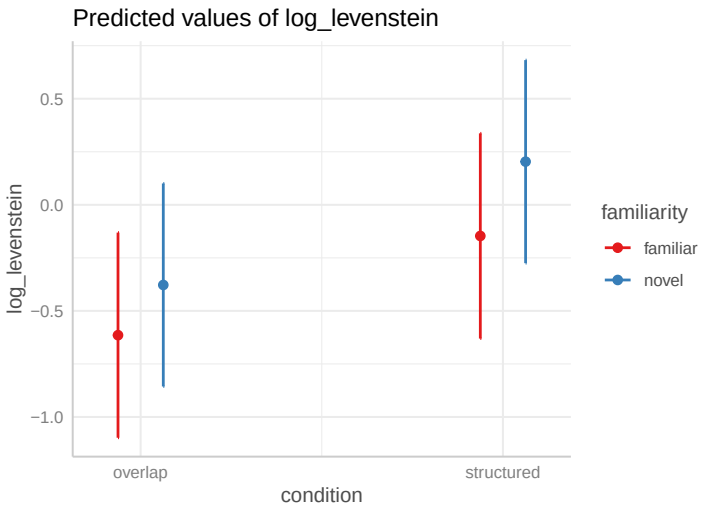


Figure 3.6: Predictions of the mean Levenstein distance with 95% confidence intervals.

Chapter 4

Discussion

4.1 Significant differences

Only two significant p-values were found, both showing a significant difference between labelling familiar aliens and labelling novel aliens in the overlap condition. While participants in the overlap condition seem to do better overall, no significant differences were found *between* the conditions. In other words, being trained on an overlap space did not produce a statistically significant improvement in learning from the structure condition to the overlap condition. Rather, familiarity happened to become a significant predictor *within* the overlap condition. Thus, my hypothesis is unambiguously wrong (see Section 2.4) [6]. However, there are some interesting nonsignificant differences.

4.2 Nonsignificant differences

Often, p-values are overrated in their statistical power (Wagenmakers 2007; Gigerenzer 2004), and because of this, other indicators are often ignored. In our case, the estimates happen to be fairly informative. Consider the testing phase where the participants had to generalise to novel aliens. While in picture selection a random pick would be correct 50% of the time (see Section 2.3), producing the correct label is a bit more complex. Since

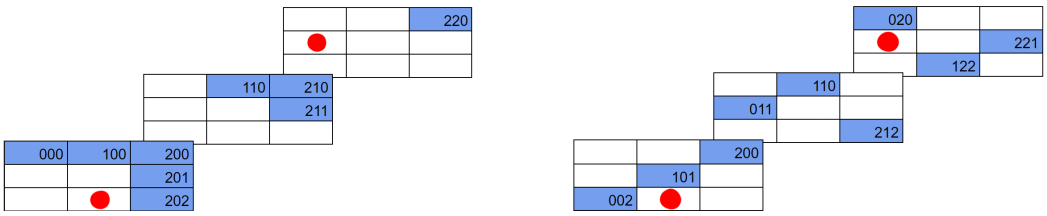
[6] One might point out that I could have stumbled upon the *multiple comparisons problem* (Gelman et al., 2012) raising the probability that a false positive could be found. In my case, applying the Bonferroni Correction (Gelman et al., 2012) would produce a 0.025 threshold (c.f., 0.05), and hence not change much.

there was no limit to the produced label's length, if the participant learned nothing (which is unlikely) their alternatives would be infinite, and hence, chance would be incalculably small. However, if at the very least participants noticed that all labels were three syllable long, potential productions would number 9x9x9 = 729 (chance = 1/729 = 0.1%). Further, they could notice the constant consonant order (z-x-j-), thus making potential productions 3x3x3 = 27 in total (chance = 1/27 = 3.7%). With this in mind, its interesting to note that while participants trained on overlap spaces had a 21.4% chance of being completely correct, being on average 0.69 letters off, participants trained on structured spaces only had a 0.02% chance of being accurate, on average 1.23 letters off. It is particularly striking that they performed as if they neither learned the length nor the consonant pattern. Now, since I failed to reject the null hypothesis through the significance tests, these differences do not mean much on their own, and have at least a 5% chance of being a fluke. However, in the case that this is representative, I now turn to the implications these results have for the emergence of compositionality, though comfortable with due uncertainty (Vasishth and Gelman, 2021).

4.3 Implications for the emergence of compositionality

4.3.1 Generalisation as Sudoku

If the nonsignificant indications of the condition differences represent real relationships in natural language, what would it mean for how the (meaning) space predicts the emergence of compositionality? Consider Figure 4.1 below, and the following idea I will call the *Sudoku*-hypothesis.



Structured space Overlap space

Figure 4.1: Ideal learned meaning spaces for participants in each condition. The red spots exemplifies potential differences in generalisation difficulty (the upper = 021, the lower = 102). See Figure 2.1 for semantic space reference.

Similar to a game of abstract three-dimensional Sudoku, in the testing phase, each participant must create labels for the aliens in the white spaces, e.g., the red spots. Consider for instance the lower red spots (102). In the overlap space, the participant could draw information from six other different objects-label pairs, three for each value: for the initial 1– value, 122, 110, 100; for the -0- value, 002, 101, 200; and for the –2, 002, 212, 122. However, in the structured condition, though they can also draw from six different object-label pairs (000, 100, 200, 201, 202, 110), five of them share the -0- value. Now, if I interpret Reeder et al. (2013) correctly in that complete overlap can prevent subcategorisation, then theoretically, participants in my structured condition could construct a separate two-dimensional meaning space for the aliens that share the -0- value. If so, since this two-dimensional space is more 'filled' relative to the larger three-dimensional space of the overlap participants, then participants in the structured condition could be better at generalising to this exact spot. Compare this to the upper red spot (021). For participant in the overlap condition, the task is identical. However, in the structured condition, there are only 3 different object-label pairs to draw from, one for each value: 000 shares 0–; 220 shares -2-; and 211 shares –1. Generalising would barely be plausible without already having generalised to the rest of the space, if not impossible if the participant has constructed separate two-dimensional subspace categories [7]. However, as seen in Figure 4.2, there is no immediate indication that there is any difference between those two novel objects in any condition.

4.3.2 Compositionality as a key to learning

Alternatively, my earlier interpretation of Reeder et al. (2013) might be mistaken such that, while doing this abstract three-dimensional Sudoku, forming any category is difficult without complete overlap. In this case, it could be that the equal spread of semantic values provided by the overlap measure makes it easier to categorise values individually. Conversely, the AIHD of the structured space skews this such that some semantic values are seen often while others rarely, thus potentially difficult to remember individually. This could produce a kind of 'blur' of semantic values such that it is difficult to categorise the

[7] The idea that the order of generalisations matter for what kind of compositionality is possible is reminiscent of Raviv et al. (2019)'s expanding meaning space.

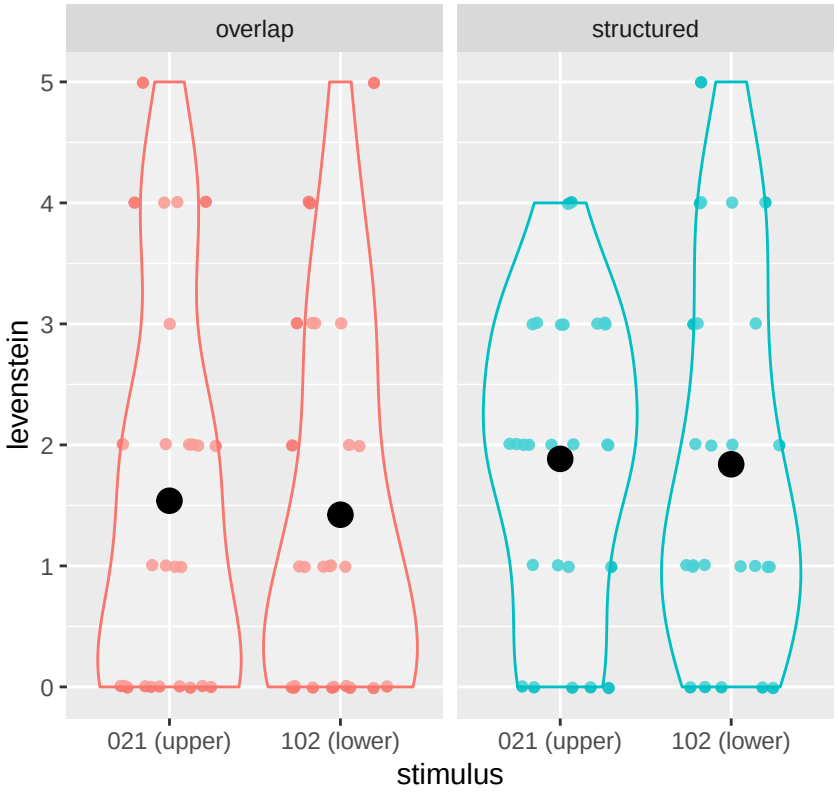


Figure 4.2: The smaller coloured dots represent each participant's edit distance from the correct label for the upper and lower red spots as shown in Figure 4.1. The larger black dot shows the overall mean.

values and dimensions correctly, thus obstructing the formation of a useful combinatorial space from which to generalise. Interestingly, this would not just prevent generalisation, but learning holistic object-label pairings would also be difficult since the semantic feature values that characterise the image would be 'blurred'. This is especially understandable considering the semantic space my participants had to structure, i.e., a bunch of similar, and potentially very confusable aliens. Further, this could partially explain why participants in the structured condition struggled to learn the labels for the 9 constantly reoccurring familiar aliens. If this is the case, overlap could facilitate semantic combinatoriality which in turn could facilitate learning (both compositional and noncompositional labels). I will now consider some potential improvements on experimental design and theory.

4.4 Potential improvements and future research

It could be that the experimental design prevented significant differences between the conditions from emerging[8]. For instance, for English speakers the labels might be too uniformly foreign, making their distinctiveness hard to parse (see Section 2.2.2). Similarly, each label being a synthetic three-morpheme string might not be ideal. Further, learning a grammar about different aliens' characteristics might be less natural than, e.g., scenes with agents, patients, and events (e.g., Tal and Arnon 2022). Regarding theory, there are also some interesting unknowns. First, given the nonsignificant p-values, it might be that the measures did not apply properly to this task. For instance, overlap might only facilitate generalisation in *syntactic* space (Reeder et al., 2013), or, unexpected effects could arise from using two combinatorial systems at once (see 1.1). Moreover, Smith et al. (2003)'s structure might not demonstrate the effect of AIHD, but rather the number of distinct values within a space and the relative exposure to each value (see Section 1.4.2). Lastly, while I have focused on generalisation as a measure of compositionality, there are probably many mechanisms beyond generalisation that contribute to both i-compositionality, and e-compositionality in natural language. These would all be important aspects to consider in future research on the emergence of compostionality.

4.5 Conclusion

In this thesis, I have explored part of a largely unexplored factor in the emergence of compositionality, i.e., the structure of the meaning space. While selection pressures like *expressivity* and *learnability* have been attested in artificial language learning through modifications of how the language is transmitted, little attention has been paid to the role of structure in the forms and meanings that shape the language. In my experiment, I draw on two quite different measures of how to induce combinatoriality in semantic or syntactic spaces. The first, Smith et al. (2003)'s *structure* has previously been demonstrated to increase compositionality in an iterated learning model. The second, Reeder et al. (2013)'s *overlap* has been used to maximise participants' ability to generalise across an

[8]Some might reject the functionality of artificial language learning (e.g., de Vries et al. 2008), however, the dominant view seems to support a useful relationship to natual language (Gómez and Gerken 2000; Bahlmann et al. 2008; Friederici et al. 2011; Misyak et al. 2010; Ettlinger et al. 2016).

entire space. Using the online crowdsourcing website Prolific, I recruited 26 participants for each measure condition, where each participant was trained on subsets of a fully regular compositional space. The results were inconclusive but indicated that generalisation and learning in general was easier when trained on an overlap space where semantic values are equally spread across a space. This could suggest that overlap facilitates the emergence of semantic combinatoriality a prerequisite for compositionality, which is in turn could arise as a consequence of specific parameters, e.g., number of semantic features and degree of exposure to each value. Along with these specific parameters, I suggest that future research should focus on how mechanisms other than generalisation contribute to different kinds of compositionality.

References

Andersen, H. (1973), 'Abductive and deductive change', *Language* **49**(4), 765–793.
URL: *http://www.jstor.org/stable/412063*

Arbib, M. A. (2012), 475 Compositionality and Beyond: Embodied Meaning in Language and Protolanguage, *in* 'The Oxford Handbook of Compositionality', Oxford University Press.
URL: *https://doi.org/10.1093/oxfordhb/9780199541072.013.0023*

Bahlmann, J., Schubotz, R. I. and Friederici, A. D. (2008), 'Hierarchical artificial grammar processing engages broca's area', *NeuroImage* **42**(2), 525–534.
URL: *https://www.sciencedirect.com/science/article/pii/S1053811908006046*

Beckner, C., Pierrehumbert, J. B. and Hay, J. (2017), 'The emergence of linguistic structure in an online iterated learning task', *Journal of Language Evolution* **2**(2), 160–176.
URL: *https://doi.org/10.1093/jole/lzx001*

Bolhuis, J. J., Beckers, G. J. L., Huybregts, M. A. C., Berwick, R. C. and Everaert, M. B. H. (2018), 'Meaningful syntactic structure in songbird vocalizations?', *PLOS Biology* **16**(6), 1–11.
URL: *https://doi.org/10.1371/journal.pbio.2005157*

Brighton, H. (2002), 'Compositional Syntax From Cultural Transmission', *Artificial Life* **8**(1), 25–54.
URL: *https://doi.org/10.1162/106454602753694756*

Brighton, H., Smith, K. and Kirby, S. (2005), 'Language as an evolutionary system',

Physics of Life Reviews **2**(3), 177–226.
URL: *https://www.sciencedirect.com/science/article/pii/S1571064505000229*

Brill, E. and Moore, R. C. (2000), An improved error model for noisy channel spelling correction, *in* 'Proceedings of the 38th Annual Meeting on Association for Computational Linguistics', ACL '00, Association for Computational Linguistics, USA, p. 286–293.
URL: *https://doi.org/10.3115/1075218.1075255*

Carr, J. W., Smith, K., Cornish, H. and Kirby, S. (2017), 'The cultural evolution of structured languages in an open-ended, continuous world', *Cognitive Science* **41**(4), 892–923.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12371*

Chomsky, N. (1986), *Knowledge of Language. Its Nature, Origin, and Use*, Convergence, Praeger, New York/Westport/London.

Cuskley, C., Simner, J. and Kirby, S. (2017), 'Phonological and orthographic influences in the bouba–kiki effect', *Psychological Research* **81**, 119–130.

de Boer, B., Sandler, W. and Kirby, S. (2012), 'New perspectives on duality of patterning: Introduction to the special issue', *Language and Cognition* **4**(4), 251–259.
URL: *https://doi.org/10.1515/langcog-2012-0014*

de Leeuw, J. (2016), 'jspsych'.
URL: *http://docs.jspsych.org/*

de Vries, M. H., Monaghan, P., Knecht, S. and Zwitserlood, P. (2008), 'Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures', *Cognition* **107**(2), 763–774.
URL: *https://www.sciencedirect.com/science/article/pii/S0010027707002533*

Edwards, T. (2014), Language Emergence in the Seattle DeafBlind Community, PhD thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-03-04.
URL: *https://www.proquest.com/dissertations-theses/language-emergence-seattle-deafblind-community/docview/1667450730/se-2*

Ellison, T. and Reinöhl, U. C. (2022), 'Compositionality, metaphor, and the evolution of language', *Int J Primatol* .

Ettlinger, M., Morgan-Short, K., Faretta-Stutenberg, M. and Wong, P. C. (2016), 'The relationship between artificial and second language learning', *Cognitive Science* **40**(4), 822–847.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12257*

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

Folia, V., Uddén, J., De Vries, M., Forkstam, C. and Petersson, K. M. (2010), 'Artificial language learning in adults and children', *Language Learning* **60**(s2), 188–220.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9922.2010.00606.x*

Friederici, A. D., Mueller, J. L. and Oberecker, R. (2011), 'Precursors to natural grammar learning: Preliminary evidence from 4-month-old infants', *PLOS ONE* **6**(3), 1–7.
URL: *https://doi.org/10.1371/journal.pone.0017920*

Garavaso, P. (2018), *Russell and Frege on the Power of Symbols and the Compositionality of Linguistic Expressions*, Springer International Publishing, Cham, pp. 93–114.
URL: *https://doi.org/10.1007/978-3-319-94364-0*$_4$

Gelman, A., Hill, J. and Yajima, M. (2012), 'Why we (usually) don't have to worry about multiple comparisons', *Journal of Research on Educational Effectiveness* **5**(2), 189–211.
URL: *https://doi.org/10.1080/19345747.2011.618213*

Gigerenzer, G. (2004), 'Mindless statistics', *The Journal of Socio-Economics* **33**(5), 587–606. Statistical Significance.
URL: *https://www.sciencedirect.com/science/article/pii/S1053535704000927*

Gordon, M. K. (2016), *Phonological typology / Matthew K. Gordon.*, Oxford surveys in phonology and phonetics ; 1, first edition. edn, Oxford University Press, Oxford, United Kingdom.

Griffiths, T. L. and Kalish, M. L. (2007), 'Language evolution by iterated learning with bayesian agents', *Cognitive Science* **31**(3), 441–480.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1080/15326900701326576*

Guo, S., Ren, Y., Havrylov, S., Frank, S., Titov, I. and Smith, K. (2019), 'The emergence of compositional languages for numeric concepts through iterated learning in neural agents'.

Gómez, R. L. and Gerken, L. (2000), 'Infant artificial language learning and language acquisition', *Trends in Cognitive Sciences* **4**(5), 178–186.
URL: *https://www.sciencedirect.com/science/article/pii/S1364661300014674*

Hamming, R. W. (1950), 'Error detecting and error correcting codes', *The Bell System Technical Journal* **29**(2), 147–160.

Hauser, M. D., Chomsky, N. and Fitch, W. T. (2002), 'The faculty of language: What is it, who has it, and how did it evolve', *Science* (298), 1569–1579.

Hockett, C. D. (1960), 'The origin of speech', *Scientific American* **203**(3), 88–96.

Hurford, J. R. (1990), *Nativist and Functional Explanations in Language Acquisition*, De Gruyter Mouton, Berlin, Boston, pp. 85–136.
URL: *https://doi.org/10.1515/9783110870374-007*

Hurford, J. R. (2000), *Social Transmission Favours Linguistic Generalisation*, Cambridge University Press, p. 324–352.

Jackendoff, R. (2003), 'Précis of foundations of language: Brain, meaning, grammar, evolution,', *Behavioral and Brain Sciences* **26**(6), 651–665.

Katz, J. J. (1970), 'Interpretative semantics vs. generative semantics', *Foundations of Language* **6**(2), 220–259.
URL: *http://www.jstor.org/stable/25000452*

Kemp, C., Xu, Y. and Regier, T. (2018), 'Semantic typology and efficient communication', *Annual Review of Linguistics* **4**(1), 109–128.
URL: *https://doi.org/10.1146/annurev-linguistics-011817-045406*

Kirby, S. (2000*a*), 'Function, selection, and innateness: The emergence of language universals', **38**(4), 817–826.
URL: *https://doi.org/10.1515/ling.2000.009*

Kirby, S. (2000*b*), *Syntax Without Natural Selection: How Compositionality Emerges from Vocabulary in a Population of Learners*, Cambridge University Press, p. 303–323.

Kirby, S. (2001), 'Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity', *IEEE Transactions on Evolutionary Computation* **5**(2), 102–110.

Kirby, S. (2002), *Learning, bottlenecks and the evolution of recursive syntax*, Cambridge University Press, p. 173–204.

Kirby, S., Cornish, H. and 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.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.0707835105*

Kirby, S., Dowman, M. and Griffiths, T. L. (2007), 'Innateness and culture in the evolution of language', *Proceedings of the National Academy of Sciences* **104**(12), 5241–5245.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.0608222104*

Kirby, S., Smith, K. and Brighton, H. (2004), 'From ug to universals: Linguistic adaptation through iterated learning', *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* **28**(3), 587–607.
URL: *https://www.jbe-platform.com/content/journals/10.1075/sl.28.3.09kir*

Kirby, S., Tamariz, M., Cornish, H. and Smith, K. (2015), 'Compression and communication in the cultural evolution of linguistic structure', *Cognition* **141**, 87–102.
URL: *https://www.sciencedirect.com/science/article/pii/S0010027715000815*

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B. (2017), 'lmerTest package: Tests in linear mixed effects models', *Journal of Statistical Software* **82**(13), 1–26.

Levenstein, V. I. (1966), 'Binary codes capable of correcting deletions, insertions, and reversals', *Soviet Physics Doklady* **10**(08), 707–710.

Mesoudi, A. and Thornton, A. (2018), 'What is cumulative cultural evolution?', *Proceedings of the Royal Society B: Biological Sciences* **285**(1880), 20180712.
URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2018.0712*

Mesoudi, A., Whiten, A. and Laland, K. N. (2006), 'Towards a unified science of cultural evolution', *Behavioral and Brain Sciences* **29**(4), 329–347.

Misyak, J., Christiansen, M. and Tomblin, J. B. (2010), 'On-line individual differences in statistical learning predict language processing', *Frontiers in Psychology* **1**.
URL: *https://www.frontiersin.org/articles/10.3389/fpsyg.2010.00031*

Pagin, P. and Westerståhl, D. (2010), 'Compositionality i: Definitions and variants', *Philosophy Compass* **5**(3), 250–264.
URL: *https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1747-9991.2009.00228.x*

Pankratz, E. (2023), 'English character bigram frequencies and transitional frequencies'.
URL: *https://github.com/elizabethpankratz/en-bigram-transitions*

Partee, B. H. (1984), 'Compositionality', *Varieties of Formal Semantics* **3**, 281–311.

Perfors, A. and Navarro, D. J. (2014), 'Language evolution can be shaped by the structure of the world', *Cognitive Science* **38**(4), 775–793.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12102*

Piantadosi, S. T., Tenenbaum, J. B. and Goodman, N. D. (2016), 'The logical primitives of thought: Empirical foundations for compositional cognitive models', *Psychological Review* **123**(4), 392–424.

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

Raviv, L., Meyer, A. and Lev-Ari, S. (2019), 'Compositional structure can emerge without generational transmission', *Cognition* **182**, 151–164.
URL: *https://www.sciencedirect.com/science/article/pii/S0010027718302464*

Reeder, P. A., Newport, E. L. and Aslin, R. N. (2013), 'From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes', *Cognitive Psychology* **66**(1), 30–54.
URL: *https://www.sciencedirect.com/science/article/pii/S001002851200076X*

Selten, R. and Warglien, M. (2007), 'The emergence of simple languages in an experimental coordination game', *Proceedings of the National Academy of Sciences* **104**(18), 7361–7366.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.0702077104*

Smith, K. (2003), The transmission of language : models of biological and cultural evolution, PhD thesis, Edinburgh.

Smith, K. (2008), 'Is a holistic protolanguage a plausible precursor to language?: A test case for a modern evolutionary linguistics', *Interaction Studies* **9**(1), 1–17.
URL: *https://www.jbe-platform.com/content/journals/10.1075/is.9.1.02smi*

Smith, K. (2018), 'The cognitive prerequisites for language: insights from iterated learning', *Current Opinion in Behavioral Sciences* **21**, 154–160. The Evolution of Language.
URL: *https://www.sciencedirect.com/science/article/pii/S235215461730178X*

Smith, K. (2022), 'How language learning and language use create linguistic structure', *Current Directions in Psychological Science* **31**(2), 177–186.
URL: *https://doi.org/10.1177/09637214211068127*

Smith, K. (n.d.), 'Week 9 practical'.
URL: *https://kennysmithed.github.io/oels2022/oels$_{p}ractical_{w}k9.html$*

Smith, K., Ashton, C. and Sims-Williams, H. (2023), The relationship between frequency and irregularity in the evolution of linguistic structure: An experimental study, *in* M. Goldwater, F. K. Anggoro, B. K. Hayes and D. C. Ong, eds, 'Proceedings of 45th Annual Meeting of the Cognitive Science Society', Cognitive Science Society, pp. 1348–1353.
URL: *https://escholarship.org/uc/item/1mz1q97f*

Smith, K., Brighton, H. and Kirby, S. (2003), 'Complex systems in language evolution: The cultural emergence of compositional structure', *Advances in Complex Systems* **06**(04), 537–558.
URL: *https://doi.org/10.1142/S0219525903001055*

Smith, K., Tamariz, M. and Kirby, S. (2013), Linguistic structure is an evolutionary trade-off between simplicity and expressivity, *in* M. Knauff , M. Pauen , N. Sebanz and I. Wachsmuth , eds, 'Proceedings of the 35th Annual Conference of the Cognitive Science Society', Cognitive Science Society, pp. 1348–1353. 35th Annual Conference of the Cognitive Science Society, CogSci 2013 ; Conference date: 31-07-2013 Through 03-08-2013.

Spike, M., Smith, K. and Kirby, S. (2016), Minimal pressures leading to duality of patterning, *in* S. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér and T. Verhoef, eds, 'The Evolution of Language: Proceedings of the 11th International Conference (EVOLANGX11)', Online at http://evolang.org/neworleans/papers/129.html.

Suzuki, T. N. (2014), 'Communication about predator type by a bird using discrete, graded and combinatorial variation in alarm calls', *Animal Behaviour* **87**, 59–65.
URL: *https://www.sciencedirect.com/science/article/pii/S0003347213004661*

Suzuki, T. N., Griesser, M. and Wheatcroft, D. (2019), 'Syntactic rules in avian vocal sequences as a window into the evolution of compositionality', *Animal Behaviour* **151**, 267–274.
URL: *https://www.sciencedirect.com/science/article/pii/S0003347219300107*

Szabó, Z. G. (n.d.), 'Compositionality'.
URL: *https://plato.stanford.edu/archives/fall2020/entries/compositionality/*

Tal, S. and Arnon, I. (2022), 'Redundancy can benefit learning: Evidence from word order and case marking', *Cognition* **224**, 105055.
URL: *https://www.sciencedirect.com/science/article/pii/S0010027722000439*

Tallerman, M. (2007), 'Did our ancestors speak a holistic protolanguage?', *Lingua*

117(3), 579–604. The Evolution of Language.
URL: *https://www.sciencedirect.com/science/article/pii/S0024384105000963*

Tamariz, M. and Kirby, S. (2016), 'The cultural evolution of language', *Current Opinion in Psychology* **8**, 37–43. Culture.
URL: *https://www.sciencedirect.com/science/article/pii/S2352250X15002225*

Theisen-White, C., Kirby, S. and Oberlander, J. (2011), Integrating the horizontal and vertical cultural transmission of novel communication systems, *in* 'Proceedings of the Annual Meeting of the Cognitive Science Society', Vol. 33, Institute of Electrical and Electronics Engineers Inc., pp. 956–961. doi 10.1109/CVPR.2017.690.
URL: *https://escholarship.org/uc/item/40t7578c*

van der Loo, M. (2014), 'The stringdist package for approximate string matching', *The R Journal* **6**, 111–122.
URL: *https://CRAN.R-project.org/package=stringdist*

Vasishth, S. and Gelman, A. (2021), 'How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis', *Linguistics* **59**(5), 1311–1342.
URL: *https://doi.org/10.1515/ling-2019-0051*

Verhoef, T., de Boer, B. and Kirby, S. (2012), *Holistic or synthetic protolanguage: evidence from iterated learning of whistled signals*, pp. 368–375.
URL: *https://www.worldscientific.com/doi/abs/10.1142/9789814401500*$_0$048

Wagenmakers, E. (2007), 'A practical solution to the pervasive problems ofp values', *Psychonomic Bulletin Review* **14**, 779–804. Statistical Significance.

Winters, J., Kirby, S. and Smith, K. (2018), 'Contextual predictability shapes signal autonomy', *Cognition* **176**, 15–30.
URL: *https://www.sciencedirect.com/science/article/pii/S0010027718300647*

Yujian, L. and Bo, L. (2007), 'A normalized levenshtein distance metric', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **29**(6), 1091–1095.

Zuberbühler, K. (2020), 'Syntax and compositionality in animal communication', *Philosophical Transactions of the Royal Society B: Biological Sciences*

375(1789), 20190062.

URL: *https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2019.0062*

Zuidema, W. (2002), How the poverty of the stimulus solves the poverty of the stimulus, *in* S. Becker, S. Thrun and K. Obermayer, eds, 'Advances in Neural Information Processing Systems', Vol. 15, MIT Press.

URL: *https://proceedings.neurips.cc/paper*$_{files/paper/2002/file/04ad5632029cbfbed8e136e5f6f7ddf}$*Paper.pdf*

Zuidema, W. and de Boer, B. (2018), 'The evolution of combinatorial structure in language', *Current Opinion in Behavioral Sciences* **21**, 138–144. The Evolution of Language.

URL: *https://www.sciencedirect.com/science/article/pii/S2352154617301298*