On Language and Meaning: A Randomized Experiment

Daniel Houser
George Mason University

Yang Yang[1]
Sun Yat-sen University

December, 2018

Abstract

We develop a method for random assignment of language to participants in a controlled laboratory experiment. We use this approach to test the linguistic relativity hypothesis (also referred to as the Sapir-Whorf hypothesis). Linguistic relativity suggests that the structure of one's language can influence one's perceptions, interpretations and beliefs about the world around them. Although provocative, empirical evidence on this hypothesis has been elusive. A reason is that previous empirical studies typically rely on naturally occurring languages whose speakers differ in ways that correlate with language differences. Here we hypothesize that linguistic relativity can emerge when the same object resides in different semantic categories across different languages. To test this, we develop a novel extension of laboratory games within which languages emerge endogenously. We show, first, that one can control the semantic categories of an emergent language by varying the game's incentives. This enables random assignment of language. Advantaged by this randomization, our experiment finds support for the hypothesis that the meaning people attribute to the same object can vary according to its semantic categorization. Our methodological and substantive insights promise to be important in improving communication, cooperation and understanding among human societies.

Acknowledgements: We are grateful to Jian Song for assistance with the experiments, to Hong Deng and Yuxuan He for assistance with the data, and Kun Ma for assistance with programming. We received valuable feedback at early stages of this research from Xianjian Zhao and Jie Zheng. For helpful advice and comments we thank especially Joel Sobel, as well as workshop and seminar participants at Beijing Normal University, Central South University, Dongbei University of Finance and Economics, George Mason University, Harbin Institute of Technology, ShanghaiTech University, Shanghai University of Finance and Economics, and Tsinghua University. We are grateful to participants at the 2017 Camphor workshop in Dali, and the Brisbane 2018 ESA Asian meetings. Support for this research was provided by the Interdisciplinary Center for Economic Science at George Mason University, and research grants to the second author from the National Natural Science Foundation of China (#71703175) and the Guangdong Natural Science Foundation (#2017A030310392).

[1] Correspondent: yangyang8@mail.sysu.edu.cn

1. Introduction

We develop an approach for random assignment of language to participants in a laboratory experiment, and we use this approach to test the linguistic relativity hypothesis. Linguistic relativity (Whorf, 1940, 1956) posits that one's language impacts one's understanding of the world around them. If so, this may help to explain why people speaking different languages may also differ in culture: that is, in prevailing customs, beliefs and behaviors. Many studies have investigated this possibility using naturally occurring data, by investigating the impact of language on counterfactual construction (Bloom, 1979; Au, 1983), categorization decisions (Berlin and Kay, 1991; Davies and Corbett, 1997; Lindsey et al., 2002; Kay and Regier, 2003; Ji et al, 2004; Athanasopoulos, 2009), diversity of thought (Lucy, 1992) or perceptions and behaviors related to time (Boroditsky, 2001; Chen 2007; Chen 2013; Roberson et al, 2015).

The goal of empirical studies is to document a causal link between language and a specific outcome variable. Compelling conclusions, however, have remained elusive. An important reason is endogeneity: languages sharing common features have a common ancestor language, thus cultural similarity across groups with similar languages could be due to a common ancestor culture with characteristics inherited by the descendent cultures. As a result, the causal effect of language on behavior in natural environments cannot be identified. Indeed, Roberson et al (2015) argue that the linguistic relativity hypothesis cannot be informed by large-scale cross-cultural correlational studies. In light of this, they argue that tightly-controlled laboratory experiments would be especially valuable. This paper takes a step in that direction.

A laboratory analysis requires a formal statement of linguistic relativity around which to develop specific hypotheses and subsequent experiment designs. We provide a rigorous formulation of linguistic relativity showing that it implies and is implied by the existence of at least one object that (i) resides in different semantic categories in different languages; and (ii) has different interpretations in different languages. Here, by "semantic category", we mean a grouping of objects whose meanings are interrelated, and who play a role in determining each other's meaning. Semantic categories are fluid and can differ among languages (Pinker, 1999). Consequently, a particular object might be grouped with different objects in different semantic categories in different languages, and this could create different meanings for the same object across languages.

An ideal laboratory test of linguistic relativity would involve random assignment of participants to languages with different semantic categories. An important contribution of this paper is that we develop a method to do this. In particular, we build from the "emergence-of-language" literature, and particularly the design reported by Selten and Warglien (2007).[2] Like their experiment, our participants play a communication game in pairs. Each pair is asked to label (emoji) objects using different pre-determined (fruit) symbols. If their labels match, they earn money, and if not, they do not.

We find, as do Selten and Warglien (2007), Hong and Zhao (2017) and Hong et al (2017), that languages do emerge in this game, and some of these languages are "compositional".[3] In addition, we provide a methodological contribution to this literature by showing that the incentive structure of the game can be manipulated to affect in predictable ways the nature of the emergent language. In particular, by varying the costs of the fruit symbols, we can promote the endogenous emergence of two different languages with two different semantic categories.[4] Importantly, our procedure helps to ensure that a "candidate" emoji falls into different semantic categories between the two emergent languages.

Moreover, we contribute by providing a direct laboratory test of linguistic relativity. We find that the candidate emoji – a red-faced smiling emoji – is more likely to be interpreted as meaning "happy" in treatments where the emergent language is more likely to place it in a semantic category with other smiling emojis, and less likely when the emergent language tends to group it instead with non-smiling but red-faced emojis. Consequently, our data offer rigorous evidence supporting linguistic relativity.

The remainder of the paper is organized as follows. Section 2 is the literature review, and section 3 develops our theory of linguistic relativity. We detail our experimental design in

[2] Many have investigated the emergence of language and meaning, or the impact of meaning in economic contexts. See, e.g., Chan et al (2011), Cremer et al (2007), Devetag (2005), Franke (2014, 2016), Galantucci and Garrod (2010, 2011), and Weber and Camerer (2003).

[3] Compositional grammars, including all human grammars, have the advantage that they allow speakers to describe objects never before seen. This differs, for example, from simple codes which only allow one to describe known objects. A large literature highlights and explores the unique advantages of compositional grammars. See e.g., Bresnan (1982), Steinert-Threlkeld (2016) or Szabo (2013).

[4] Rubinstein (1996) was the first to provide a formal economic analysis of the way incentives ("evolutionary forces") of a language environment can impact the emergence and development of language. More broadly, work on economics and language includes Lipman (2003), Marschak (1965), and Rubinstein (2000).

section 4, and sections 5 and 6 present procedures and results. Section 7 offers a concluding discussion.

2. Literature Review

Many have suggested that language structure can influence thought (for a review see, e.g., Kay and Regier, 2003). Loosely speaking, this is thought to occur due to the way language organizes and categorizes an otherwise unstructured reality (Saussure, 1916). Clearly, for this to be the case it is necessary that people speaking a common language coordinate on this categorization and organization. This point was made forcefully by Whorf (1940), who wrote:

> We cut nature up, organize it into concepts, and ascribe significances as we do, largely because we are parties to an agreement to organize it in this way — an agreement that holds throughout our speech community and is codified in the patterns of our language. The agreement is, of course, an implicit and unstated one, but its terms are absolutely obligatory; we cannot talk at all except by subscribing to the organization and classification of data which the agreement decrees (Whorf, 1940).

Despite a century's work on this topic, Whorf's conjecture remains controversial.[5] Linguistic theorists have argued for the presence of pan-human universal language (Chomsky, 1956; Pinker, 1994) that leaves no room for language to impact thinking. Empirically, tests of linguistic relativity have been many, with Chen (2013) offering evidence that language structure could systematically impact decision-making. Roberts et al (2015) reanalyzes the data from Chen (2013), casting some doubt on the earlier results and ultimately concluding that the validity of linguistic relativity remains an open question. As noted above, these authors suggest that further evidence on linguistic relativity would seem to require controlled laboratory tests of the relationship between language and meaning.

[5] There are many early discussions of the connection between language and thought. Vygotsky (1934) provides an early treatment connected to developmental psychology. Montague (1970, 1973) offers an influential formal characterization of grammar emphasizing the importance of categorization and organization of concepts.

While they do not consider linguistic relativity, a small but influential literature investigates the emergence of language in controlled laboratory experiments.[6] These papers test specific hypotheses related to the optimality and efficiency properties that languages should hold. The typical framework includes a finite set of objects (or states) to be labeled (Selten and Warglien, 2007). Participants engage in a game, the outcome of which can be a labeling (or coding) that forms a language (e.g. Selten and Warglien, 2007; Hong et al, 2017; Hong and Zhao, 2017). In the current paper we too employ these types of games, drawing particularly heavily from Selten and Warglien (2007). Our paper is the first to use these games to inform the linguistic relativity hypothesis by testing whether speakers of different languages assign different meanings to the same object.[7]

3. Formal Statement of Linguistic Relativity

The purpose of this section is to develop a formal statement of linguistic relativity, from which we can derive precise hypotheses and an experiment design appropriate for their testing. Towards this end, denote a language space by *M*, so that *M* consists of a meaningful vocabulary; and a shared understanding among users of *M* that it is appropriate to interpret statements $m \in M$ according to their focal meaning. Let Ω denote the universe of objects $\omega \in \Omega$ that can be uniquely identified by statements in *M*. Let Ξ denote the universe of meaning, and $\xi^M : \Omega \to \Xi,$ so that $\xi^M(\omega) \in \Xi$ denotes the meaning of object $\omega.$ The linguistic relativity hypothesis is that the function ξ^M varies with the underlying language *M*. The reason is that semantic categories can vary across languages, and this can impact the meaning.

[6] A substantial amount of research in artificial intelligence has investigated “emergent language” using simulation techniques, see, e.g., Foerster et al, (2016), Lazaridou et al. (2017), Mordatch and Abbeel (2017), Cao et al, (2018), Resnik et al, (2018).

[7] Hong and Zhao (2017) asked subjects to describe a novel object using a single “word”. They were first to show that the single “word” people use to do this can systematically vary according to the emergent language they “speak”. While interesting, this finding does not inform linguistic relativity. The reason is that they provide no evidence that the speakers of different languages assign different meanings to the same object. For example, when a person speaking a language that distinguishes colors describes a basketball as “brown” and another shape-distinguishing language speaker describes a basketball as “round”, they are simply referring to the more salient feature of the object, rather than understanding the basketball to be either only “brown” or only “round”.

To see this, let X^M denote the set of semantic categories of language *M*, with element $\chi^M \in \mathrm{X}^M$. Thus, χ^M represents a set of messages *m* that describe objects with similar meaning. Let $\Omega(\chi^M)$ denote the set of objects described by messages in χ^M, and let $\Xi^M(\Omega(\chi^M))$ denote the set of meanings of the objects described by messages in χ^M.

Linguistic relativity posit that the meaning of an object is influenced by its semantic category. The reason is that objects within the same semantic category participate in each other's definition, and thus have more similar meaning than do objects between semantic categories. Because semantic categories are fluid and can vary across languages, this can lead to the same object falling into different semantic categories in different languages, and consequently holding different meanings in different languages.

Thus, to demonstrate linguistic relativity it is necessary and sufficient to demonstrate the existence of some object $\omega \in \Omega$ and two languages *M* and *M'* with distinct semantic categories $\chi^M \neq \chi^{M'}$ such that $\omega \in \Omega(\chi^M)$ and $\omega \in \Omega(\chi^{M'})$ but $\xi^M(\omega) \neq \xi^{M'}(\omega)$.

That is, linguistic relativity implies and is implied by the existence of at least one object with meaning that varies according to its semantic categorization.

4. Testing Linguistic Relativity

In view of the above, it is clear what one requires in order to test linguistic relativity: at least one object that falls into different semantic categories in different languages. If, further, the meaning of this object differs between these two languages, then this is evidence supporting linguistic relativity. Unfortunately, such tests have proven difficult to construct using naturally occurring data. For example, a natural approach (e.g., Chen, 2013) identifies an object (e.g., the future) whose description lies in different semantic categories in different languages, and then attempts to determine whether the meaning of this object differs across these languages. One approach to doing this is to determine whether future-oriented behaviors differ between people who speak different languages, but as pointed out by Chen et al, 2015, any discovered differences may owe to factors beyond language differences. An alternative approach has been to focus on perception, say of colors (Berlin and Kay, 1991; Davies and Corbett, 1997; Lindsey et al., 2002; Kay and Regier, 2003; Athanasopoulos, 2009), which are often described differently across different languages. The argument is that if people who describe colors differently also perceive colors differently, this is evidence in favor of linguistic relativity. The concern with this argument is endogeneity: the descriptions of colors may differ because different groups of people are sensitive to colors differently, perhaps for example due to systematic genetic differences across populations.

Our experiment circumvents difficulties found with naturally occurring data by developing a test of linguistic relativity within a controlled laboratory environment. Our experiment design enables the endogenous emergence of different languages, characterized by different semantic structures, such that the same object is in different semantic groups in different languages. Once that is accomplished, we must assess whether the interpretation of this object differs between users of the different languages.

4.1. Overview of Experiment Design to Test Linguistic Relativity

Emergent Language: We build from an experiment design first suggest by Selten and Warglien (2007) to generate emergent language in a laboratory environment. The design, described in detail below, involves a coordination game. It turns out that language, even compositional languages, emerge reliably using this game (see, e.g., Selten and Warglien, 2007; Hong *et al*, 2017[8]).

Obtaining Different Languages with Different Semantic Groups: In our experiment it is costly to use symbols to describe objects. Different symbols have different costs. Our design includes 12 objects to be described, and the symbols cost structure creates an economic incentive, in one case, to organize the objects into four semantic groups of three objects each, while in the other case to organize the objects into three groups of four objects each. Our paper is the first to show that semantic groups can be systematically determined by varying the incentives of the language environment. The ability to do this is crucial, as it enables random assignment of language to groups.

The Objects: Our experiment includes 12 different emojis. Participants create strings of up to six fruit symbols to describe uniquely these emojis. Our emoji set includes four with smiles, four with frowning faces, and the remaining four with round open mouths. Among those, one smiling emoji has a red face, one frowning emoji has a red face, and one round

[8] Hong *et al* (2017) use a modified version of Selten and Warglien's (2007) coordination game in their experiment.

open mouth emoji has a red face. By varying the costs associated with fruit, it turns out that we can provide an economic incentive to create three semantic categories of four emojis each (based on the shape of the mouth) or four semantic categories of three emojis each (where the red-faced emojis form their own unique semantic category).

Linguistic Relativity Hypothesis: In the context of our design, the linguistic relativity hypothesis is that users of different languages with different semantic categories will attribute different meanings to at least one red-faced emoji. Our design enables clean inferences in this regard, for the following reasons. First, because people are randomized into languages, we control for the possibility that differences in meaning of an object across languages is due to differences in the people assessing meaning. Second, we investigate the meaning of emojis, which are likely more flexibly interpreted than physical or abstract objects (e.g., sun or fire; or a square or circle). Third, we are able to determine whether the semantic categorization of the emoji in the "fruit" language impacts its meaning in relation to an existing idea expressed in the English language. The ability to assess meaning in relation to an existing standard is a substantial advantage of our design, as it provides increased power to detect linguistic relativity.

Finally, it is important to note that our design also solves a potential endogeneity problem: if meaning differs across languages it may be because initial interpretations of the emojis differed, and these different interpretations caused different languages to emerge. Our design controlled this possibility: different initial interpretations are randomized across treatments (language environments), thus the systematic differences in the emergent language across environments can arise only due to the incentives of the environment.

5. Experiment Design

We use emojis and fruits as objects and symbols in the communication games detailed below. In doing this we differ from other studies on emergent artificial languages in the lab (Blume et al, 1998; Selten and Warglien, 2007; Hong and Zhao, 2017; Hong et al, 2017). In those studies, the objects are geometric shapes and the symbols are either English letters or signs (e.g. !, @, #, $, %, ^). While studies using shapes have proven value, in our case the use of emojis improves the power of our design. In particular, geometric objects have unique objective interpretations, while the meaning ascribed to emojis can be relatively more flexible. Consequently, we expect our design with emojis to have improved power to detect linguistic relativity effects.

5.1 Treatments *3-3* and *4-2*

Our experiment includes two incentive structures, which we denote by ***3-3** and **4-2***, both building from the experiment design introduced by Selten and Warglien (2007). Within each incentive structure people play a communication/coordination game, as described below.

Communication Game

The communication game involves two players, one *sender* and one *receiver*. They both see the same list of *n* objects, $o_1, o_2, \ldots, o_n$, on their own screens, with order randomized and different for different subjects. For each object on the list, both the sender and the receiver need to compose an expression using m symbols, $f_1, f_2, \ldots, f_m$. Each player can choose any symbol(s) from the repertoire for each expression. The symbols may appear in any order and each may appear any number of times. It is not allowed to use the same expression to describe different objects or to leave a blank expression for any object.[9] When both players have submitted their expressions for all the objects, one object is randomly selected and the players' expressions for that object are compared. If the expressions perfectly match, the communication is successful and each of the subjects earns s EC, otherwise the communication fails and the subjects earn 0 EC.[10]

Regardless of the success or failure of the communication, the sender has to pay for the fruits used in her own expression for the selected object. For instance, denote the cost for f_i as c_i, then for an expression with *t* fruits, "$f_{k_1} f_{k_2} \ldots f_{k_t}$", where $k_1, k_2, \ldots, k_{t,} k \in \{1, 2, \ldots, m\}$, a sender needs to pay $\sum_{j=1}^{t} c_{k_j}$ experimental coins. Receivers do not need to pay for symbols used in their expressions. Following the design of Selten and Warglien (2007), the role

[9] When a subject submits his/her expressions for all the objects in a round, if there is any identical expression used for different objects, or blank expression for any object, the program will pop up an error message to the subject and asks for a revision of the expressions until no violations of these rules are detected.

[10] Our game is a coordination game. The key difference between it and a standard coordination game is that the strategy space is infinite, in the sense that for each emoji the expression may in principle consist of an infinite number of fruits, which evidently leaves coordination more difficult *ex ante*.

assignment is only realized at the end of each round. Therefore, when subjects make decisions regarding their expressions, both have 50% chance to be the sender, thus both should be sensitive to costs.

At the end of each round, each subject receives the following information: the randomly selected object (emoji) and both subjects' expressions for that object within the pair, his/her randomly assigned role (sender/receiver), his/her current round's payoff, and the accumulated payoff from the first to the current round.

Treatments 3-3 and 4-2: Details

Participants play two 60-round communication games with emojis (e.g.. 🙂, 😯) as objects, and fruits (e.g.. 🍎, 🍇) as symbols. The two main treatments are denoted *3-3* and *4-2.* The full list of the objects (emojis) and the fruit repertoire with the costs are listed in Table 1.

For the first 10 rounds, subjects only need to form expressions for 2 emojis using 2 fruits, each with cost 1 EC, i.e., $n = 2, m = 2, c_1 = c_2 = 1$. From 11th round to 30th round, four more emojis are introduced into the object set to be described, while the size of the repertoire remains as before (i.e., $n = 6, m = 2, c_1 = c_2 = 1$). From the 31st to the 60th round, six new emojis are added to the set, and we also make four additional fruits available for the expressions (i.e. $n = 12, m = 6$). Some new fruits cost 1 EC, and some cost 5 EC per use, as shown in Table 1. Beginning with few objects and symbols leaves it easier for subjects to create a fruit language. With the two distinct cost structures in the two treatments, we expect different language structures to emerge in a precise way that we detail further below.

The two incentive conditions are exactly the same except that the cost of 🍇 is 5 EC in treatment *3-3* and 1 EC in Treatment *4-2*. This means Treatment *3-3* includes three relatively lower-cost fruits (🍎, 🍌, 🍓) and three with higher costs (🍇, 🍉, 🍒) in the final 30 rounds, as compared to four lower-cost (🍎, 🍌, 🍓, 🍇) and two higher-cost fruits (🍉, 🍒) in Treatment *4-2*.

Table 1. The emojis objects and the fruits repertoire (cost) in each round of the communication games in Treatment *3-3* and *4-2*

Round	Objects Treatment 3-3 & 4-2	Repertoire Treatment 3-3	 Treatment 4-2
1-10	🙂 😯	🍎 🍌 (1)	
11-30	🙂 😯 ☺️ 😮 😀 😲	🍎 🍌 (1)	
31-60	🙂 😯 ☺️ 😮 😀 😲 ☹️ 🙁 😦 😳 😊 😣	🍎 🍌 🍓 (1) 🍇 🍉 🍒 (5)	🍎 🍌 🍓 🍇 (1) 🍉 🍒 (5)

Note: The number in the parentheses denote the cost per use of each fruit. E.g. '🍎🍌 (1)' means each use of 🍎 or 🍌 costs one EC.

In each treatment, we implemented the 60-round communication games twice. In the first part, subjects are randomly paired and play the 60 rounds with the same partner. Part 2 follows immediately. Subjects are randomly rematched to play another 60-round communication game as in Part 1, but this time with a different partner. Subjects are endowed with 250EC for both parts, and are informed that their initial 250 EC plus the ECs they earn during the two parts of the experiment will be converted to US dollars at a rate of 1 EC= $0.03. Earnings were bounded below by zero (though as a practical matter this never occurred).

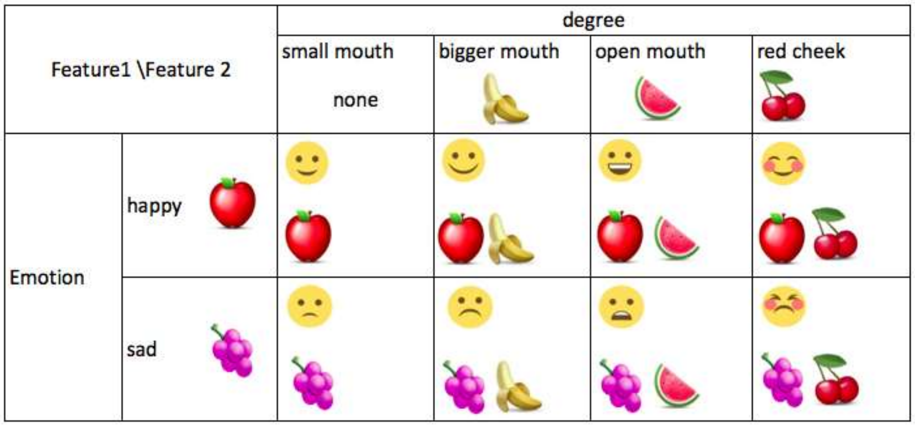
Language Predictions

In each part, we say a pair of subjects achieved a *common code* if they matched expressions for all 12 emojis in the final round of the communication game. Common code can often be understood as a language, and our predictions regarding the nature of the emergent language require two assumptions. The first is cost efficiency, meaning that subjects prefer to use a language with lower average costs per emoji. Selten and Warglien (2007) report cost efficiency is important factor for communication success in their experiment.

The second assumption relates to the structure of the emergent language. We expect, based on previous literature, emergent languages to have a "compositional" structure, which means the language uses specific fruits in specific orders to describe specific features of emojis. Tables 2a and 2b provide examples of compositional languages.

Table 2. Two Examples of Compositional Language

2(a)



Feature1 \Feature 2		degree			
		small mouth none	bigger mouth 🍌	open mouth 🍉	red cheek 🍒
Emotion	happy 🍎	🙂 🍎	🙂 🍎🍌	😀 🍎🍉	😊 🍎🍒
	sad 🍇	🙁 🍇	☹️ 🍇🍌	😦 🍇🍉	😣 🍇🍒

2(b)

Feature1 \Feature 2		degree			
		weakest single fruit	weak twice repetition	strong triple repetition	strongest four-time repetition
Emotion	happy 🍎	🙂 🍎	☺ 🍎🍎	😀 🍎🍎🍎	😊 🍎🍎🍎🍎
	sad 🍇	🙁 🍇	☹ 🍇🍇	😧 🍇🍇🍇	😣 🍇🍇🍇🍇

Participants are of course allowed to develop any coding structure. It has been shown, however, that as compared to other language structures, compositional language has many advantages, including learning efficiency (learnable with minimal number of examples, Blume, 2004), and the ability to describe objects that have never been seen before (Blume, 2000), based on which Blume (2000) further argues that through an evolutionary process such optimal structure should better survive and thus be observed often in natural environments. Consistent with his view, experimental studies find compositional languages do emerge frequently (Selten and Warglien, 2007; Hong et al, 2017; Hong and Zhao 2017)[11].

Like all coordination games, anything on which people coordinate is an equilibrium. Based on the above two assumptions, however, it is easily verified that, given the cost structures, there is a unique cost-efficient compositional language in treatment 3-3, and two possible cost-efficient languages in 4-2, though these languages can of course be expressed in many different ways (by label switching, such as apples for bananas, for example). We predict the cost-efficient equilibria, examples of which are shown in Table 3, to emerge in our game.[12]

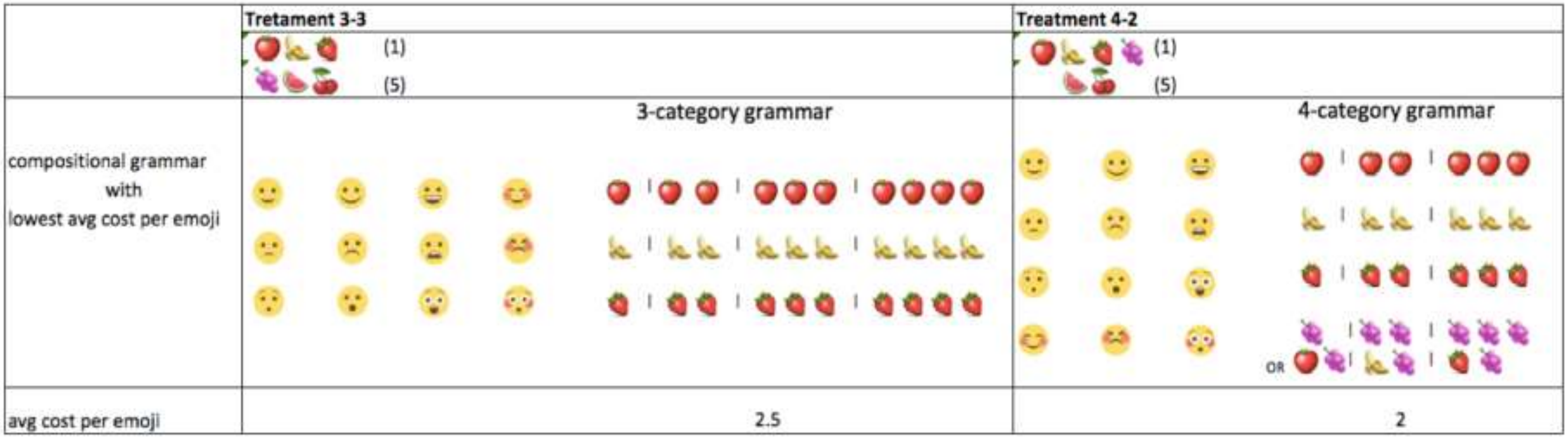
[11] Another type of grammar, "positional compositionality", may further lower expression costs. We do not consider these as they are difficult for participants to create, are not observed in our experiment, and have not ever been observed in previous related experimental studies (Selten and Warglien, 2007; Kirby et al, 2008; Cornish et al, 2010; Hong et al, 2017).

[12] This prediction is supported by previous experiments using similar games (Selten and Warglien, 2007; Hong et al, 2017; Hong and Zhao, 2017), and more generally experimental work with coordination games (e.g., Van Huyck et al, 1991) as well as theoretical work on coordination games including Crawford and Haller (1990); and Crawford and Sobel (1982).

In treatment 3-3, where there are only three lower-cost fruits avaialable, the most cost-efficient language uses one type of the cheap fruit only (🍎, 🍌, 🍓), for each of the four emojis with a common feature. For instance, as shown in the example in Table 3, using different numbers of apples to describe a series emojis with a smiling face, using bananas for frowning faces, and strawberries for emojis with round-open mouths. Note that the emojis described by the same type of fruit are natually semantically linked under such a language, and are natually separated into three semantic categories. Hence, we call this language a 3-semantic-category language, or 3-category language.

For treatment 4-2, where there are four inexpensive fruits, the cheapest compositional language to express the 12 emojis uses a cheap fruit and its repetition for the emojis with the above common features (smiling faces, frowning faces, and round-open mouths) without the red cheeks, but using the fourth cheap fruit only (or combining the fourth fruit with the other three cheap fruits) for the red emojis. For such a language, the expressions are semantically categorized into four classifications: the first three each involve one type of the low-cost fruit only, and the fourth involves a fourth low-cost fruit. Hence, we call this language a 4-category language.

Table 3. Predicted (Cost-Efficient-Compositional) Emergent Languages in Treatment 3-3 and 4-2.



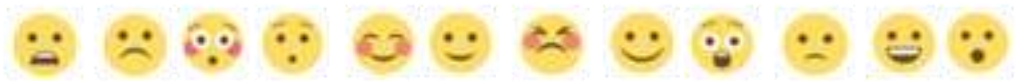
	Tretament 3-3	Treatment 4-2
	🍎🍌🍓 (1) 🍇🍉🍒 (5)	🍎🍌🍓🍇 (1) 🍉🍒 (5)
compositional grammar with lowest avg cost per emoji	3-category grammar 🙂 ☺️ 😀 😊 🍎 | 🍎🍎 | 🍎🍎🍎 | 🍎🍎🍎🍎 🙁 ☹️ 😧 😣 🍌 | 🍌🍌 | 🍌🍌🍌 | 🍌🍌🍌🍌 😯 😮 😲 😳 🍓 | 🍓🍓 | 🍓🍓🍓 | 🍓🍓🍓🍓	4-category grammar 🙂 ☺️ 😀 🍎 | 🍎🍎 | 🍎🍎🍎 🙁 ☹️ 😧 🍌 | 🍌🍌 | 🍌🍌🍌 😯 😮 😲 🍓 | 🍓🍓 | 🍓🍓🍓 😊 😣 😳 🍇 | 🍇🍇 | 🍇🍇🍇 OR 🍎🍇 | 🍌🍇 | 🍓🍇
avg cost per emoji	2.5	2

Note : The costs of the fruit sysmbols are displayed in the second row for each treatment.

Meaning Elicitation Tasks

Following both parts of the 60-round communication games, subjects in both Treatments *3-3* and *4-2* were asked to complete a meaning elicitation task. For this task, each subject remained paired the same partner as in Part 2 of the communication game. Each was asked to answer three questions.

Question 1: Among the 12 emojis you have seen in our experiment, which three emojis best characterize “happy”?



The emojis are the same used in the communication game. The order of the emojis shown on the question sheet for every subject and all the three questions is exactly as shown above. This task was incentivized: if two paired subjects selected exactly the same three emojis for the question, each wins a $2 reward. The incentive structure makes this task, again, a coordination game, only that the strategy spaces here is in emojis.

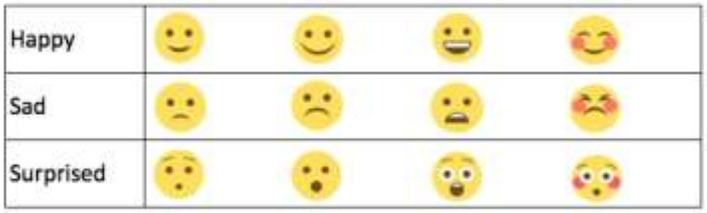
Question 2 and 3 are exactly the same as Question 1, except that “happy” is replaced by “sad” and “surprised” respectively.

Notice that this is the first time that the terms “happy”, “sad” and “surprised” appeared in the experiment. Thus, the only factors that could have an influence on the subjects’ answers to these questions are: 1) their pre-existing interpretations to the listed emojis—how closely they relate each emoji to the indicated emotion; and 2), their experience in developing and communicating in their fruit language. The former would be the same across treatments, thus any between-treatment differences in the categorization task can be attributed to their use of different languages. This forms the basis for our test of linguistic relativity.

5.2 Baseline Treatment: Identifying a Candidate Emoji

As detailed in the model above, linguistic relativity requires (i) that there exists at least one object that lies in two different semantic categories in two languages; and (ii) that the object holds a different interpretation between these two languages. The purpose of this baseline treatment is to establish candidate emojis that might allow a powerful test of linguistic relativity.

Table 4. Emojis and Emotions



Happy	🙂	☺	😀	😊
Sad	😕	☹	😧	😣
Surprised	😯	😮	😲	😳

Specifically, for each of the three emotions happy, sad and surprised our experiment includes three related non-red emojis and one related red emoji, as shown in Table 4[13]. Our interest is in knowing how strongly subjects relate each red emoji from each group to each of the three emotions. We assess this by measuring how often each red emoji is selected as one of the three emojis that best characterize each emotion in the categorization task.

To test linguistic relativity we focus exclusively on red-faced emojis that were categorized as a particular emotion at a rate insignificantly different from 75%. This is the rate the red-faced emoji would be chosen if people felt that all four emojis represented the emotion equally well. As we will see, it turns out this results in a single candidate: the red-faced smiling emoji.

The reason we focus exclusively on red-faced emojis is that we design our experiment so that red-faced emojis are likely to be semantically categorized either with the emotion to which they are most closely associated, or alternatively with other red-faced emojis, depending on the treatment. We expect the yellow-faced emojis to be grouped according to the emotion they express in both treatments. Hence, we expect any variation in meaning to occur only for the red-faced emojis.

The reason we use 75% as the cut-off rate for our candidate emojis is that, as noted, our language manipulation is hypothesized to result in red emojis categorized according to their suggested "emotion" at about the same rate as the baseline in treatment 3-3 (where the four similar emotions should be grouped together), but at a lower rate in treatment 4-2 (where the red faces are grouped together). If participants do not generally agree, however, that the red-faced emoji could mean the emotion we suggest as well as the other yellow-face emojis, then there is

[13] The connection between the emojis and emotions shown in Table 4 is plausible, and also consistent with 97% (265 out of 274) of subjects' answers in the baseline categorization task.

evidently less room for the 4-2 treatment to cause change.[14] That is, the power of our design to detect linguistic relativity is positively related to the frequency with which the red-faced emoji is chosen in the baseline treatment.

A summary of the structure of all the three treatments of our experiment can be found in Table 5.

Table 5. Summary of Treatments

	Treatments		
	3-3	*4-2*	*Baseline*
Part 1: Communication games (60 rounds)	✓	✓	-
Part 2: communication games (60 rounds)	✓	✓	-
Categorization Test	✓	✓	✓

5.3. Specific Hypotheses

Hypothesis 1: The rate at which language emerges will be the same in the 3-3 and 4-2 treatments.

This hypothesis asserts that changing the incentive structure of the environment will not change the rate at which language emerges. The incentives should impact the nature of emergent language, as we detail in Hypotheses 2 below.

Hypothesis 2a: An exact 3-category language emerges at least as frequently as 4-category language in treatment 3-3.

Hypothesis 2b: An exact 4-category language emerges at least as frequently as 3-category language in treatment 4-2.

[14] Optimizing the power of the design is important. The reason is that participant behavior is noisy. We will see below that languages emerge according to our predictions, but imperfectly.

Hypothesis 2c: Exact 3-category language emerges more frequently in the 3-3 than 4-2 treatment, while the reverse is true for exact 4-category language.

Hypothesis 2d: The emergent exact 3-category language and the emergent exact 4-category language will be as predicted in Table 3 above.

Note that Hypotheses 2a-2c make one-sided predictions. Note further that these hypotheses are knife-edge, focusing on 3- or 4-category languages that emerge perfectly. We can broaden these hypotheses as follows.

Hypothesis 3a: In treatment 3-3, the emergent languages will be "closer" to 3-category than 4-category.

Hypothesis 3b: In treatment 4-2, the emergent languages will be "closer" to 4- than 3-category.

Hypothesis 3c: Language "close" to 3-category emerges more frequently in the 3-3 than 4-2 treatment, while the reverse is true for language close to 4-category.

We explain the measurement of "closer" below. Note that these are also one-sided predictions.

Hypothesis 4 (Linguistic Relativity): In view of the results of our Baseline treatment, we hypothesize that the smiling red-faced emoji will be categorized as "happy" more often in the 3-3 than 4-2 treatment.

Note that Hypothesis 4 is also a one-sided hypothesis.

5.4 Procedures

Experiments were conducted at the ICES lab of George Mason University in 2017. The subjects were recruited from the ICES subject pool consisting of undergraduate and graduate students from all backgrounds at George Mason University.

Subjects are randomly assigned to different treatments. A total of 72, 64 and 92 subjects individuals participated in our Treatments *4-2*, *3-3* and *Baseline*, respectively.[15] The

[15] Four of the 72 subjects who participated in Treatment *3-3* could not proceed to Part 2 after completing Part 1 due to technical difficulties, leaving only 68 observations from this treatment.

communication games were implemented using software coded in Python, HTML, JavaScript and CSS.

Each subject's earnings include a $5 show-up fee at the beginning of the experiment and their performance-based earnings from the communication games and the categorization task. Each session on average takes 150 minutes for treatment 3-3 and 4-2, and 45 minutes for Baseline. The average earnings are $36.5 in Treatment *3-3* and *4-4*, and $7.8 in *Baseline*.

6. Results

6.1 Language Emergence in Communication Games

To test Hypothesis 1, following Selten and Warglien (2007), we evaluate whether, for each part of the communication game, each pair of subjects matched their expressions for all 12 emojis in the final round. With the observations from the two parts of communication games by 68 subjects in treatment 3-3 and 64 subjects in treatment 4-2, we obtain the following result:

Result 1. Consistent with Hypothesis 1, the rate of language emergence does not differ across treatments.

We confirm this by calculating between-treatment differences in: 1) the fraction of subjects who have reached common code in Part 1; 2) fraction of subjects with common code in Part 2; 3) fraction of subjects with common code in both Part 1 and Part2; and 4) the fraction of subjects with common code in either Part 1 or Part 2. With a 2-sided t-test, the between-treatment difference is insignificant by all the above four criteria (see Table 6).

In order to test the remaining hypotheses, we focus exclusively on participants who achieved a common code in either Part 1 or Part 2 of the communication games.

Table 6. Fraction of subjects with Common Code by Treatment

	Number of Subjects	Common Code in Part 1	Common Code in Part 2	Common Code in Parts 1 & 2	Common Code in Part 1 or 2
Treatment 3-3	68	42/68 (62%)	54/68 (79%)	32/68 (47%)	64/68 (94%)
Treatment 4-2	64	30/64 (47%)	46/64 (72%)	23/64 (36%)	53/64 (83%)
p-value (two-sided t-test)		0.09	0.32	0.20	0.06

As noted above, the 3-category and 4-category languages are the most cost-efficient compositional languages for Treatment 3-3 and 4-2, respectively. Our data reveal that subjects respond to these incentives, and in the predicted way by Hypotheses 2.

Result 2. Incentives influence emergent language in the way predicted by Hypotheses 2.

In treatment 3-3, we observe 16% of the common codes that emerged in either Part 1 or Part 2 to have an exact 3-category language, as compared to only 7% having an exact 4-category language. This difference is significant by a 1-sided t-test (p-value=0.035), and supports Hypothesis 2a.

In treatment 4-2, we observe an exact 4-category language to emerge with a frequency of 8% among all common codes, while an exact 3-category language never emerges. The difference is significant (p-value<0.001 by 1-sided t-test) and supports Hypothesis 2b.

Hypothesis 2c compares between treatments. Comparing the 18% of common codes in Treatment 3-3 that are exact 3-category language to the zero in Treatment 4-2, the between treatment difference is significant (p-value<0.001 by 1-sided t-test). This supports Hypothesis 2c. The emergent rate of the 4-category language in Treatment 3-3 and 4-2 are respectively 7% and 8%, which is statistically insignificant but directionally consistent with Hypothesis 2c.

Finally, Hypothesis 2d is that the emergent languages will follow the predictions detailed in Table 3. To provide evidence on this, Appendix A1 and A2 detail all exact emergent languages we observed. It is clear from visual inspection that the exact languages support the predictions of Table 3. In particular, in the 3-category languages the red-faced emojis are grouped with the emojis related to the emotion to which they are most similar, while in 4-2 the red-faced emojis form their own category.

Result 3: Even when not exact, emergent languages are overall closer to 3-category in Treatment 3-3, and closer to 4-category language in Treatment 4-2.

Hypotheses 3 suggests that language is "closer" to 3-category in 3-3, and closer to 4-category in 4-2. To test this, we measure the "distance" of each common code from an exact 3-category language by determining the smallest number of fruit expressions that need to be changed in order to produce an exact-3-category language. Using the same procedure, we measure the distance of each common code from a 4-category language. We then calculate the difference between these distances, with a negative difference indicating that a language is closer to a 3-category language, and a positive difference indicting it is closer to a 4-category language.

The histogram of the distance differences for each treatment is shown in Figure 1. It is apparent that a greater number of positive differences appear in Treatment 4-2 than in Treatment 3-3, and a Chi-square test shows the two distributions are significantly different (p-value<0.001). This supports Hypotheses 3.

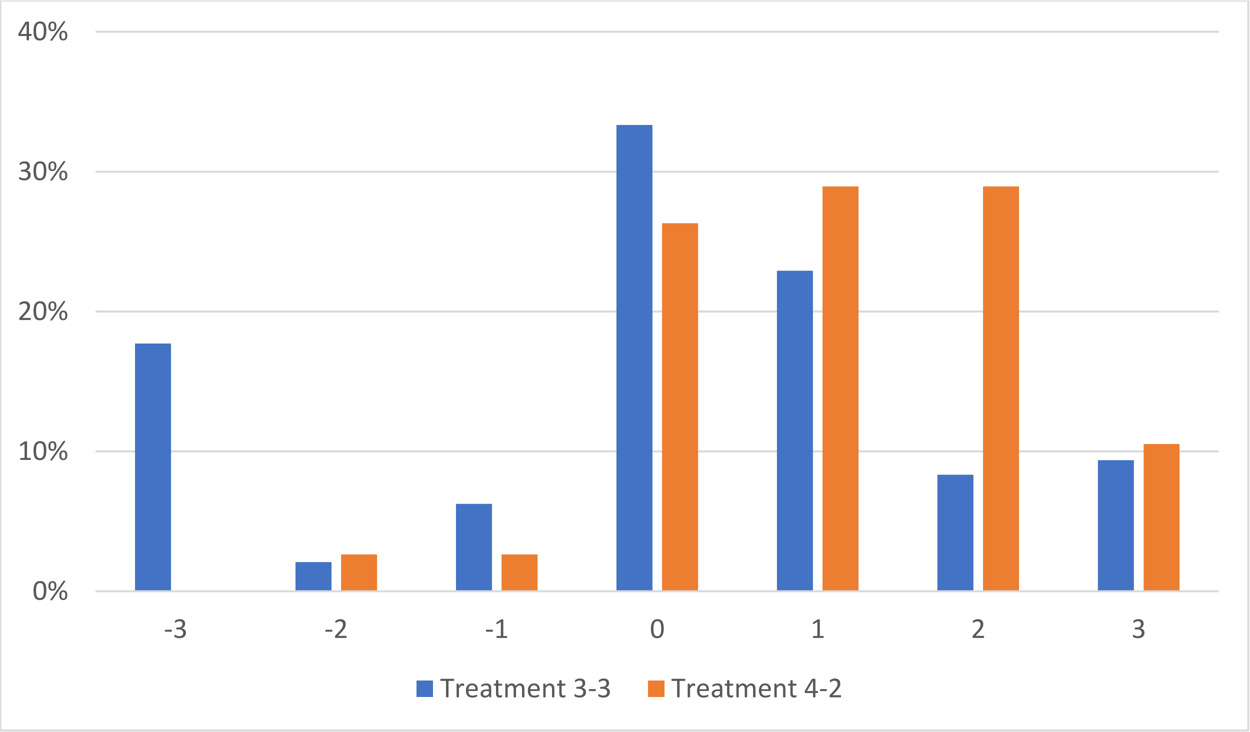


Figure 1. Distance of Common Code from a 3-category Language - Distance of Common Code from a 4-category language. (The positive domain indicates a language being closer to the 4-category than then 3-category, with negative domain the opposite.)

Yet another way to detail differences in the emergent language across treatments is to compare the fraction of common codes that are *clearly closer* to the exact 3-category language than to the exact 4-category language, and the fraction that are clearly closer to the exact 4-category language. We say a common code is clearly closer to the exact 3(4)-category language if its distance from the 3 (4)-category language is at least 2 less than that from the exact 4(3)-category language. The results, which can be derived from Figure 1, are shown in Figure 2. In Treatment 3-3, we observe that 20% of the common code are clearly closer to the exact 3-category language than 4-category language, which is higher than the fraction of the common code that is clearly closer to the 4-category language (18%). In treatment 4-2, we observe the opposite, that 39% of the common code is clearly closer to the 4-category language while only 3% is clearly closer to the 3-category language.

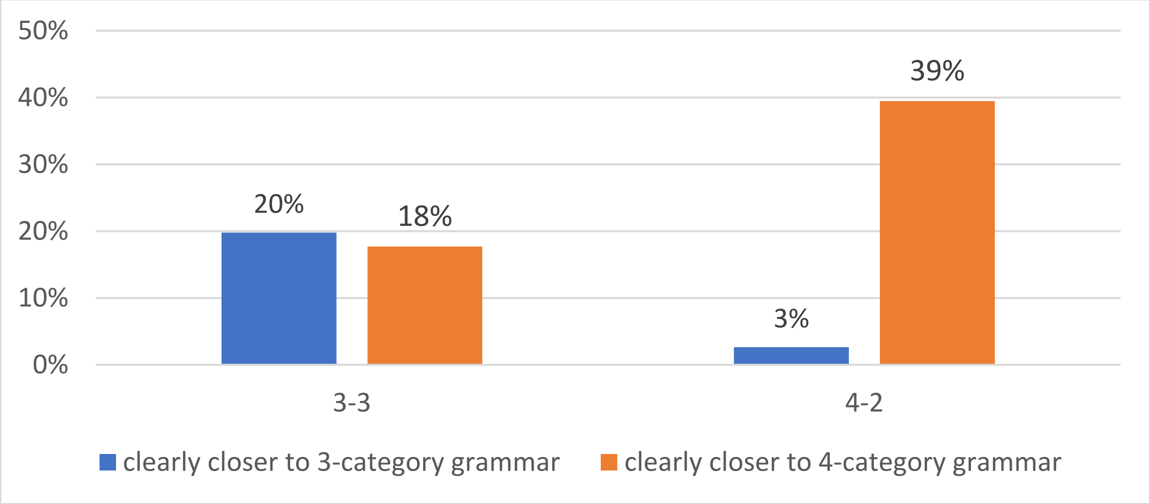


Figure 2. Fraction of common code clearly closer to the exact 3-category or 4-category language.

6.2 Linguistic Relativity

We have established with the above results that the emergent language varies between treatments and, from Result 2, 3 and Appendix A1 and A2, that the red-faced emojis are categorized systematically differently between the two treatments. In order to test linguistic relativity, we must now identify an emoji whose interpretation is likely to vary depending on its categorization. As discussed above, we conducted a baseline categorization task to identify such candidates. Of the three red emojis, only the red-faced smiling emoji qualified as a candidate, with 69% of the subjects indicating that it meant “happy” (insignificantly different from 75%, 2-sided t-test, p-value=0.24). Regarding the other emojis, 32% of the subjects indicated the frowning red-faced emoji meant “sad”, and 58% indicated the open-mouth red-faced emoji was “surprised”. Both of these rates are significantly different from 75% (p-value≤0.001 for both cases, 2-sided t-test).

In view of this, the Linguistic Relativity Hypothesis is that the red-faced smiling emoji is less likely to be categorized as happy in 4-2 than 3-3. The relevant frequencies are shown in Figure 3. We find that 75% of participants categorized the red-faced emoji as “happy” in 3-3, while only 57% did so in 4-2. The difference is in the hypothesized direction and is statistically significant (p=0.02, 1-sided t-test).[16] This evidence supports the Linguistic Relativity Hypothesis.

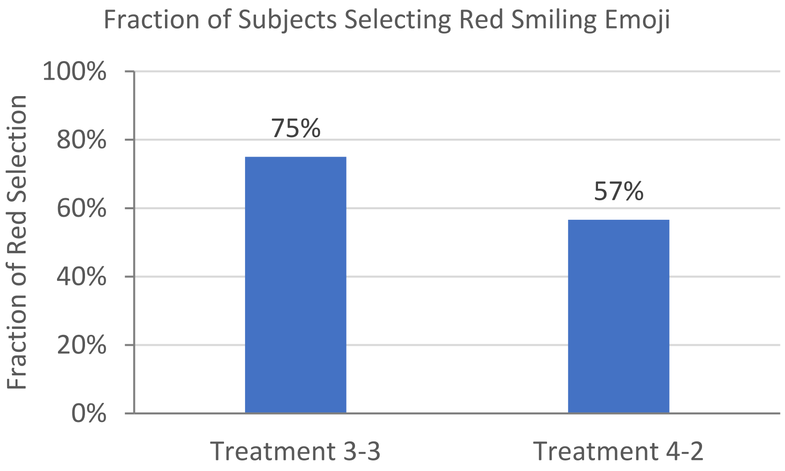


Figure 3. Fraction red smiling emoji 😊 selected as meaning "happy"

7. Conclusion

We developed a method to assign language randomly to participants in a laboratory experiment and used this approach to analyze linguistic relativity. Our analysis was based on data from a communication game where participants coordinated on fruits expressions used to describe emoji objects. Our paper offers both methodological and substantive contributions. Methodologically, we demonstrated that by varying the incentives of the coordination game one can reliably create languages that include different semantic categories. This insight is important, as it opens the door to studies investigating links among, for example, language, culture, expectations and beliefs.

[16] The results for the other two red-faced emojis are in the expected direction but are not statistically significant. The open-mouth red-faced emoji was chosen to represent "surprise" with 63% frequency in 3-3, and 60% in 4-2. The frequencies for the frowning red-faced emoji are 25% and 21% in 3-3 and 4-2, respectively. Note that we constructed our sample (including those who achieved a common code in the first or second rounds) in order to maximize numbers of observations. Importantly, as detailed by Table B.1, our qualitative findings remain unchanged regardless of the sample used, though statistical significance can change due to differing sample sizes.

Substantively, advantaged by the ability to assign languages to participants randomly, we provided rigorous evidence on linguistic relativity. In particular, our data reveal that a red-faced smiling emoji is significantly more likely to be interpreted as expressing “happy” in treatments where that emoji is more likeley to be grouped semantically with other “happy” (but not red-faced) emojis, as compared to treatments where it is more likely to be semantically catagorized with other red-faced emojis expressing non-happy emotions.

Some have suggested that our results are perhaps due to “pure coordination”, as opposed to actual changes in interpretation of meaning. The argument is that people rely on the groupings formed with their counterpart in the Part 2 game when coordinating with that same Part 2 counterpart on the emojis that mean, for example, “happy”. In particular, some point out that this coordination can be accomplished regardless of their actual belief regarding the meaning of the emojis. This explanation conflicts, however, with the poor coordination rates for other emojis. Moreover, our data offer direct evidence against this alternative explanation.

Recall that we include participants who coordinated in the Part 1 or Part 2 games, implying that we observe decisions by people who coordinated only in Part 1, but not Part 2. If the effect we find is due exclusively to “pure coordination”, then one would anticipate that the former group would be far less likely to display a between-treatment difference in the likelihood of using the red-faced smiling emoji for “happy”. In fact, among the 11 participants in treatment 3-3 who coordinated in the first but not second part of the experiment, nine (82%) chose the red-faced smiling emoji to indicate “happy”. Moreover, only three (42%) of the seven participants in treatment 4-2 did so. While we recognize the sample is small, this difference is nevertheless weakly statistically significant (one-sided Fisher exact test, p-value = 0.07).

Linguistic relativity raises the possibility of a complex intertwining of language and culture. The methods and data reported in this paper represent a useful step towards unraveling this complexity, and providing a deeper understanding of, and respect for, differences across cultures and peoples. Further studies taking advantage of the methods developed in this paper hold the promise of promoting shared perspectives and peaceful interactions among disparate cultures and societies.

References

[1] Au, T. K. F. (1983). “Chinese and English counterfactuals: the Sapir-Whorf hypothesis revisited”. *Cognition*, 15(1-3), 155-187.

[2] Athanasopoulos, Panos. (2009). “Cognitive representation of colour in bilinguals: The case of Greek blues.” *Bilingualism: Language and Cognition* 12 (1):83-95.

[3] Berlin, B. and Kay, P. (1991). *Basic color terms: Their universality and evolution.* Univ of California Press.

[4] Bloom, A. H. (1979). “The impact of Chinese linguistic structure on cognitive style”. *Current anthropology*, 20(3), 585-586.

[5] Blume, A., DeJong, D., Kim, Y., & Sprinkle, G. (1998). “Experimental evidence on the evolution of meaning of messages in sender–receiver games”. *American Economic Review*, 88(5), 1323–1340.

[6] Blume, A., (2000). “Coordination and learning with a partial language”. *Journal of Economic Theory*, *95*(1), pp.1-36.

[7] Blume, A. (2004). “A learning-efficiency explanation of structure in language”. *Theory and Decision*, *57*(3), 265-285.

[8] Boroditsky, L. (2001). “Does language shape thought? Mandarin and English speakers conceptions of time”. *Cognitive Psychology*, 43, 1–22.

[9] Bresnan, J. (1982). *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

[10] Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., & Clark, S. (2018). “Emergent Communication through Negotiation.” *arXiv preprint arXiv*:1804.03980.

[11] Chan, C., Tardif, T., Chen, J., Pulverman, R., Zhu, L., & Meng, X. (2011). “English- and Chinese-learning infants map novel labels to objects and actions differently”. *Developmental Psychology*, 47(5), 1459–1471.

[12] Chen, M. Keith. (2013). “The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets”. *The American Economic Review* 103 (2):690-731.

[13] Chomsky, N. (1956). “Three models for the description of language”. *IRE Transactions on information theory*, 2(3), 113-124.

[14] Claidière, Nicolas, Yasmina Jraissati, and Coralie Chevallier. (2008). “A colour sorting task reveals the limits of the universalist/relativist dichotomy: colour categories can be both language specific and perceptual”. Journal *of Cognition and Culture* 8 (3):211-233.

[15] Crawford, V., & Haller, H. (1990). “Learning how to cooperate: Optimal play in repeated coordination games.” *Econometrica*, 58(3), 571—595.

[16] Crawford, V., Sobel, J., (1982). “Strategic information transmission”. *Econometrica* 50, 1431–1451.

[17] Cremer, J., Garicano, L., & Prat, A. (2007). “Language and the theory of the firm”. *Quarterly Journal of Economics*, 122, 373-407.

[18] Davidoff, Jules, Ian Davies, and Debi Roberson. 1999. "Colour categories in a stone-age tribe." *Nature* 398 (6724):203-204.

[19] Devetag, G. (2005). “Precedent transfer in coordination games: An experiment.” *Economics Letters*, 89(2), 227—232.

[20] Foerster, J., Assael, I. A., de Freitas, N., & Whiteson, S. (2016). “Learning to communicate with deep multi-agent reinforcement learning.” *Advances in Neural Information Processing Systems* (pp. 2137-2145).

[21] Franke, M., (2014). “Creative compositionality from reinforcement learning in signaling games.” In: Cartmill, Erica A., et al. (Eds.), *The Evolution of Language: Proceedings of the 10th International Conference (Evolang 10)*. World Scientific, Singapore, pp. 82–89.

[22] Franke, M., (2016). “The evolution of compositionality in signaling games.” *J. Logic, Lang. Inf*. 25, 355–377.

[23] Galantucci, B., Garrod, S., (2010). “Experimental semiotics: a new approach for studying the emergence and the evolution of human communication.” *Interaction Studies* 11, 1–13.

[24] Galantucci, B., Garrod, S., (2011). “Experimental semiotics: a review. Front.” *Human Neurosci*. 17, 1–15.

[25] Gentner, D. (1982). “Why nouns are learned before verbs: Linguistic relativity vs. natural partitioning.” In S. A. Kuczaj (Ed.), *Language development. Language, thought and culture* (Vol. 2). Hillsdale, NJ: Lawrence Erlbaum.

[26] Hong, Fuhai, Wooyoung Lim, and Xiaojian Zhao. 2017. "The emergence of compositional grammars in artificial codes." *Games and Economic Behavior* 102:255-268.

[27] Hong, F., & Zhao, X. (2017). “The emergence of language differences in artificial codes.” *Experimental Economics*, *20*(4), 924-945.

[28] Ji, L. J., Zhang, Z., & Nisbett, R. E. (2004). “Is it culture, or is it language? Examination of language effects in cross-cultural research on categorization.” *Journal of Personality and Social Psychology*, 87(1), 57–65.

[29] Kay, Paul, and Terry Regier. 2003. "Resolving the question of color naming universals." *Proceedings of the National Academy of Sciences* 100 (15):9085-9089.

[30] Lazaridou, A., Peysakhovich, A., & Baroni, M. (2016). “Multi-agent cooperation and the emergence of (natural) language.” *arXiv preprint arXiv:1612.07182.*

[31] Lipman, B. (2003). “Language and economics.” In N. Dimitri, M. Basili, & I. Gilboa (Eds.), *Cognitive processes and rationality in economics*. London: Routledge.

[32] Lucy, J. A. (1992). *Language diversity and thought: A reformulation of the linguistic relativity hypothesis.* Cambridge: Cambridge University Press.

[33] Mordatch, I., & Abbeel, P. (2017). “Emergence of grounded compositional language in multi-agent populations.” *arXiv preprint arXiv:1703.04908.*

[34] Marschak, J. (1965). “Economics of language.” *Behavioral Science*, 10(2), 135–140.

[35] Montague, R. (1970). “English as a formal language”, in B. Visentini, et al. (eds.), *Linguaggi nella Società e nella Tecnica, Milan: Edizioni di Communita*, 189–224; reprinted in Thomason (ed.) 1974, pp. 188–221.

[36] Montague, R. (1973), “The proper treatment of quantification in ordinary English”, in K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes (eds.), *Approaches to Natural Language* (Synthese Library 49), Dordrecht: Reidel, 221–242. Reprinted in Portner and Partee (eds.) 2002, pp. 17–35.

[37] Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language. William Morrowand Company*.

[38] Pinker, S., (1999). *Words and Rules: The Ingredients of Language*. Harper Perennial, New

York, NY.

[39] Resnick, C., Kulikov, I., Cho, K., & Weston, J. (2018). “Vehicle Community Strategies.” *arXiv preprint arXiv*:1804.07178.

[40] Roberts, S.G., Winters, J. and Chen, K., 2015. "Future tense and economic decisions: controlling for cultural evolution." *PloS one*, *10*(7):0132145.

[41] Rubinstein, A., (1996). “Why are certain properties of binary relations relatively more common in natural language?” *Econometrica* 64, 343–356.

[42] Rubinstein, A., (2000). *Economics and Language*. Cambridge University Press, Cambridge.

[43] De Saussure, F. (1916). “Nature of the linguistic sign.” Course in general linguistics, 65-70.

[44] 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.

[45] Steinert-Threlkeld, S., (2016). “Compositional signaling in a complex world.” *J. Logic, Lang. Inf.* 25 (3), 379–397. December.

[46] Szabó, Z.G., (2013). “Compositionality.” In: Zalta, Edward N. (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2013 edition. http://plato.stanford.edu/ archives/fall2013/entries/compositionality/.

[47] Van Huyck, J. B., Battalio, R. C., & Beil, R. O. (1991). “Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games.” *Quarterly Journal of Economics, 106, 885–911*.

[48] Vygotsky, L.S., (1934). (1987). “Thinking and speech.” In RW Rieber & AS Carton (Eds.), *The collected works of LS Vygtosky: Vol. 1. Problems of general psychology* (pp. 37-285).

[49] Weber, R., & Camerer, C. (2003). “Cultural conflict and merger failure: An experimental approach.” *Management Science*, 49, 400–415.

[50] Whorf, B. (1956). “Language, thought, and reality: Selected writings of Benjamin Lee Whorf.” In J. B. Carroll (Ed.), *Language, thought and reality*. Cambridge, MA: MIT Press.

[51] Whorf, B.L., (1940). *Science and linguistics* (pp. 207-219). Indianapolis, IN: Bobbs-Merrill.

Appendix A1. All Emergent Exact 3-Cateogry Languages

treatment	session	userid	pairid	p1ccode	p2ccode	exact 3-category grammar	exact 4-category grammar	part	🙂	☺️	😀	😊	😯	😮	😲	😳	😕	☹️	😫	😖
Tr 3-3	0328	1	38	0	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0328	9	38	1	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0329	1	44	0	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0329	11	13	1	1	1	0	1	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0329	11	44	1	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0329	13	14	1	0	1	0	1	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0329	14	14	1	1	1	0	1	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0531	1	25	1	1	1	0	1	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0531	3	60	1	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0531	8	25	1	1	1	0	1	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0531	8	60	1	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0601	2	33	1	1	1	0	1	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0601	4	68	0	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0601	5	33	1	1	1	0	1	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓
Tr 3-3	0601	5	68	1	1	1	0	2	🍅	🍅🍅	🍅🍅🍅	🍅🍅🍅🍅	🍌	🍌🍌	🍌🍌🍌	🍌🍌🍌🍌	🍓	🍓🍓	🍓🍓🍓	🍓🍓🍓🍓

Appendix A2. All Emergent Exact 4-Cateogry Languages

treatment	session	userid	pairid	p1ccode	p2ccode	exact 3-category grammar	exact 4-category grammar	part	🙂	☺️	😀	😊	😯	😮	😲	😳	😕	☹️	😬	😖
Tr 3-3	0329	10	12	1	0	0	1	1	🍅	🍅🍅	🍅🍅🍅	🍓🍓🍓	🍌	🍌🍌	🍌🍌🍌	🍓	🍇	🍇🍇	🍇🍇🍇	🍓🍓
Tr 3-3	0530	2	22	1	1	0	1	1	🍌	🍌🍌	🍌🍌🍌	🍒	🍅	🍅🍅	🍅🍅🍅	🍒🍒	🍉	🍉🍉	🍉🍉🍉	🍒🍒🍒
Tr 3-3	0530	3	22	1	1	0	1	1	🍌	🍌🍌	🍌🍌🍌	🍒	🍅	🍅🍅	🍅🍅🍅	🍒🍒	🍉	🍉🍉	🍉🍉🍉	🍒🍒🍒
Tr 3-3	0602	1	36	1	1	0	1	1	🍌	🍌🍌	🍌🍌🍌	🍓🍓	🍅	🍅🍅	🍅🍅🍅	🍓	🍉	🍉🍉	🍉🍉🍉	🍓🍓🍓
Tr 3-3	0602	2	70	1	1	0	1	2	🍌	🍌🍌	🍌🍌🍌	🍓🍓	🍅	🍅🍅	🍅🍅🍅	🍓	🍉	🍉🍉	🍉🍉🍉	🍓🍓🍓
Tr 3-3	0602	4	36	1	1	0	1	1	🍌	🍌🍌	🍌🍌🍌	🍓🍓	🍅	🍅🍅	🍅🍅🍅	🍓	🍉	🍉🍉	🍉🍉🍉	🍓🍓🍓
Tr 3-3	0602	4	70	1	1	0	1	2	🍌	🍌🍌	🍌🍌🍌	🍓🍓	🍅	🍅🍅	🍅🍅🍅	🍓	🍉	🍉🍉	🍉🍉🍉	🍓🍓🍓
Tr 4-2	0401	1	105	1	0	0	1	1	🍅	🍅🍅	🍅🍅🍅	🍅🍓	🍌	🍌🍌	🍌🍌🍌	🍌🍓	🍇	🍇🍇	🍇🍇🍇	🍇🍓
Tr 4-2	0401	2	105	1	1	0	1	1	🍅	🍅🍅	🍅🍅🍅	🍅🍓	🍌	🍌🍌	🍌🍌🍌	🍌🍓	🍇	🍇🍇	🍇🍇🍇	🍇🍓
Tr 4-2	0401	2	137	1	1	0	1	2	🍅	🍅🍅	🍅🍅🍅	🍅🍓	🍌	🍌🍌	🍌🍌🍌	🍌🍓	🍇	🍇🍇	🍇🍇🍇	🍇🍓
Tr 4-2	0401	5	137	1	1	0	1	2	🍅	🍅🍅	🍅🍅🍅	🍅🍓	🍌	🍌🍌	🍌🍌🍌	🍌🍓	🍇	🍇🍇	🍇🍇🍇	🍇🍓
Tr 4-2	0607	3	132	1	1	0	1	1	🍅	🍅🍅	🍅🍅🍅	🍒	🍌	🍌🍌	🍌🍌🍌	🍒🍒🍒	🍉	🍉🍉	🍉🍉🍉	🍒🍒
Tr 4-2	0607	6	132	1	0	0	1	1	🍅	🍅🍅	🍅🍅🍅	🍒	🍌	🍌🍌	🍌🍌🍌	🍒🍒🍒	🍉	🍉🍉	🍉🍉🍉	🍒🍒

Appendix B

Table B1. The Frequency of Each Red Emoji Being Selected as Expressing the Respective Emotion, by Different Sub-Sample of Subjects with Common Code Formed in Different Parts

	Common code in *Part 1*		**Common code in *Part 2***		**Common Code in *Part 1 or Part 2***		**Common code in *Part 1 & Part 2***		**All Subjects**		
	3-3	***4-2***	***3-3***	***4-2***	***3-3***	***4-2***	***3-3***	***4-2***	***3-3***	***4-2***	***Baseline***
Nr. Subjects	42	30	54	46	64	53	32	23	68	64	92
Happy	74%	53%	74%	59%	75%	57%	72%	57%	76%	61%	69%
1-sided Fisher test p-value	0.061*		0.078*		0.028**		0.186		0.041**		
Sad	26%	14%	26%	22%	25%	21%	28%	13%	28%	24%	32%
1-sided Fisher test p-value	0.168		0.401		0.397		0.158		0.368		
Surprised	62%	60%	67%	57%	63%	60%	69%	52%	65%	64%	58%
1-sided Fisher test p-value	0.531		0.202		0.482		0.167		0.541		