SimLang can simulate the evolutionary trend of language in color naming task

Yi Ren (renyi.joshua@gmail.com)

Abstract

World color survey (WCS) dataset is a large color-naming dataset gathered from participants in those preindustrialized cultures all around the world. Such a dataset is widely applided in various research fields, e.g., cognitive science, linguistics, society science, etc. The community of evolutionary linguistics also derives many inspiring hypothesis from analyzing this dataset. However, as the dataset is obtained from real people, it is hard to enhance the dataset to a larger scale. Furthermore, we cannot go back to the past and gather more detailed data to analyze how language evolves, which is the main focus of the evolutionary linguistics. Hence people proposed a simulation tool named SimLang, which can simulate the language evolution procedure on any pre-defined object-naming tasks, e.g., color-naming task. The goal of this paper is to demonstrate that the simulated language generated by SimLang can represent some crucial trend of the real WCS dataset using some visualization tools, and meanwhile, can provide additional information of how the language evolves. Some statistics of the SimLang's output are also demonstrated to show that this simulation tool is robust on various environment settings.

1 Introduction

I might put the related work part in three parts (i.e., not use a separate section): introduction part to provide an overview; WCS data part to provide works on real data collection; SimLang part to provide some evolutionary linguistics stuffs.[Background about linguistics]

Among all the capabilities we human beings have, the language ability is the most special one to tell us apart from other primate animals: we can use language to communicate, cooperate, record things, or even express our motions and feelings. From the perspective of information theory, human's language is a coding scheme for knowledge delivery among agents (i.e., people). Among all the features that human's languages have, the compositionality, which enables us to express more complex concepts using a systematical combination of several simpler concepts, is the most important one [ref]. With the help of compositionality, people can express infinitely many concepts using limited amount of words or phrases. That is why the modern languages have similar information capacity compared with the acient ones, even thorough the world is becoming more and more complex [ref].

[Introduce WCS and its limitations]

To find out how this striking feature of language occurs, many evolutionary linguists not only propose various explanations for this phenomenon, but also keep searching for evidences to support their hypothesis. One of the widely used dataset for this topic is the world color survey (WCS) dataset [ref]. As the entire language describing the real world is too complex to find any pattern in it, the authors of [ref] narrow down their focus on a simple but representative task, i.e., the color-naming task. In this task, volunteers from different regions are asked to describe different colors on a Munsell color card (a standard color card [ref]) using their first languages. To get knowledge from the ancient languages, some of the participants are carefully selected from those preindustrialized cultures that had limited contact with modern, industrialized society. Although the WCS dataset can provide many insights about the language evolution process, it still has some limitations. For example, as the data is collected by requiring the participants to fill questionnaires, it is hard to get larger scale dataset in this way. Furthermore, as the timescale of language evolution is hundreds or thousands of years, it is hard for us gather enough chronological data on different languages. In other words, the data we have in WCS are only some snap shots of specific languages.

[Introduce SimLang as a solution to aforementioned problems]

To overcome this problem, some researchers propose to use artificial intelligence technology (AI) to simulate the language evolution process. They believe that the emergent language generated by the neural agent system can also shed light on how language evolves on some extent [ref]. Using the neural agents rather than real people, they can run simulations under different system settings on various tasks. Among all these tools, SimLang, proposed in [ref], is considered to be an effective framework to do such simulations. With the help fo SimLang, many exciting works emerges recently [ref].

[Propose our goal: use vis to introduce SimLang and show that it can indeed represent important trends in WCS]

However, there is also some opposite voice on using such AI simulation tool to study how human language evolves [ref]. They assert that the emergent language of neural agents is fundamentally different from human language, because people and computer are biological different. Although the defenders of SimLang claim that the computer simulations can provide insights from information theoretical perspective [ref], some more straightforward ways to illustrate the validity of SimLang are still needed. Hence in this paper, we will use some visualization tools to show that the results obtained from SimLang is robust and effective enough to represent some crucial trends and features of the WCS on the color-naming task. The main goal of this paper is to persuade people that SimLang can indeed assist the research on evolutionary linguistics.

The remainder of this paper will be organized as follows. In section 2, we will visually explain what is color-naming task and what are the quantitative metrics we care about in analyzing the language evolution process. In section 3, we will provide a brief introduction of the WCS data, together with some pre-processing process of the data. In section 4, we will show what the data generated by SimLang looks like. The statistics and dynamics of the generated language are also illustrated to prove the effectiveness of SimLang. In section 5, we may make thorough comparisons between WCS data and simulated data. And in section 6, we draw a conclusion and provide some discussions about future directions.

[P.S. My experience about this topic]

As a Ph.D. candidate in computer science, I think I can write code and analyze the raw data from WCS. For the linguistics and SimLang part, I do have some previous work on that topic. I choose this topic because I think the knowledge I gained from CPSC 547 can help me provide better visual illustrations on SimLang. During the rebuttal phase of my previous work, the reviewers asked a plenty of questions on the information visualization part. I think with the help of the tools and principles I learned from this course, I can explain my ideas better using good figures.

2 Color-naming task and Two metrics

In this section, we will first introduce the color-naming task and explain why it is suitable for the research of language evolution. Then we will provide two quantitative metrics that are highly correlated with the emergence of compositionality during language evolution.

2.1 Color-naming task

People use language to describe different concepts in their living environments, hence the more complex the world people lives in, the more complex their language is. Even the most ancient language (that have records) contains thousands of distinct concepts [ref], which means that observing the whole language is impossible. As a result, many researchers narrow down their focus to some simple tasks on small groups of concepts, e.g., numerals, positions, colors [ref]. Among all of these tasks, color-naming task is an efficient way to analyze how compositionality emerge during language evolution, because the different color cells in the Munsell color system are usually defined by two (hue, brightness) or three (plus saturation) attributes.

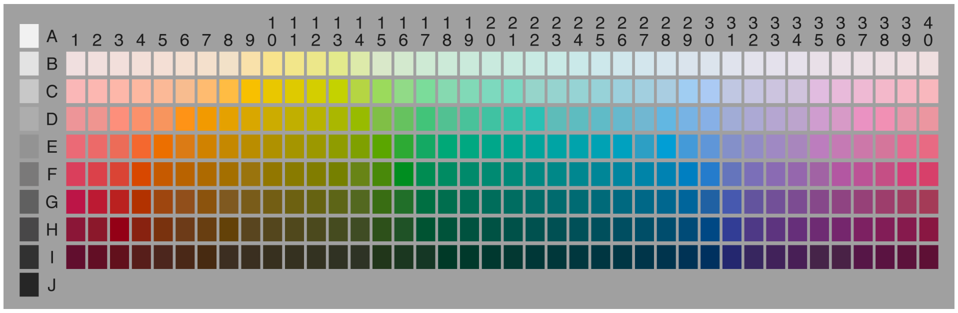


Figure 1: Standard Munsell color card.

In color-naming task, we use a standard Munsell color system, as illustrated in Figure 1. In this color card, 330 different colors are arranged in two groups: the left column contains blocks with 10 different gray scales (brightness, denoted by $\mathbb{B} = \{A, ..., J\}$) and the right region contains 320 different colors. In the right region, there are 40 columns labeled as $\mathbb{H} = \{1, ..., 40\}$ representing different hues, and 8 rows with $\mathbb{B}_{color} = \{B, ..., I\}$ representing different brightness. Note that in the right region, the brightness only ranges from B to I[1]. All of these colors are **randomly** indexed with a chip number ranges from 1 to 330. Hence we can refer to a specific color whether using its index (denoted by $idx \in \{1, 330\}$) or by a tuple $(h, b) \in \mathbb{H} \times \mathbb{B}$, e.g., $(A, 01)$[2].

In this paper, we discuss two different ways of getting data (phrases in languages) by doing this task:

- MCS data: letting volunteers fill a questionnaire to describe each cell in the color card;
- SimLang data: letting neural agents play a referential game to generate emergent language [ref];

These two ways can both generate data samples like $(A, 0) =$ **light red**. The details of the generating process, so as the structure of the raw data obtained, will be elaborated later in the paper.

2.2 Word types and topological similarity

To study how compositionality emerges during language evolution, the following two quantitative metrics play important roles:

- Number of word types: denoted by $N_V = \{1, 2, ...\}$, the number of different word types used in describing all 330 blocks in Munsell;
- Topological similarity: denoted by $\rho \in [0, 1]$, a metric calculated from original data to evaluate how compositional the mappings are [ref];

In this paper, rather than providing the formal definition of ρ, which can be found in [ref], we use Figure 2 to give an intuition of what is the difference between high-ρ and low-ρ languages. In this figure, the mapping of high-ρ language disentangled well: the same word always refers to the same concept in all phrases. For example, B always means ”light” in both $B1$ and $B30$. But in low-ρ language, B means ”dark” in $B1$ but ”light” in $B30$, which violate the aforementioned alignment. From the results obtained in [ref], we know that higher ρ means the language is more compositional, which is the main trend of language evolution.

3 WCS data visualization

In this section, we will introduce the raw WCS data in detail. Some visualization methods are applied to provide a good overview of the data. We will also discuss how to calculate N_V and ρ from raw data.

[1] That is because A is pure white and J is pure black.

[2] The indexing of the color card can be found in the *chip.txt* file in WCS dataset (`www1.icsi.berkeley.edu/wcs/data.html`).

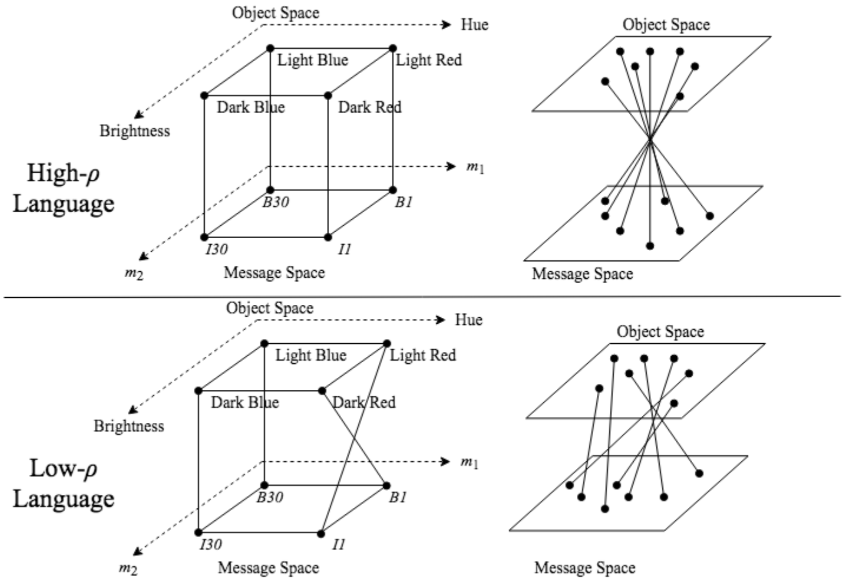


Figure 2: Toy example on languages with different toplogical similarity.

3.1 Raw data from WCS project

Not all ready yet, I am still reading some related papers to find some interesting features of the data to demonstrate in this section. In WCS project, an average of 24 native speakers of each of 110 languages were asked to name each of 330 Munsell chips. Then the participants speaking the same language will discuss and pick out the best example (from the names they used before) of each chips after seeing a palette of these chips. The results are stored following the structure of Table 1. In WCS project, we have in total XXXX pieces of recordings[3].

Language ID	Speaker ID	Focus Response	Term Used	Described Chip
1	1	1	LF	A0
1	1	2	WK	D9
1	2	6	WK	D10
...	...	...	...	...

Table 1: Examples of raw WCS data, stored in *foci-exp.txt*.

There are five columns in this table:

- Language ID: denote which language this piece of recording is, ranges from 1 to 110;
- Speaker ID: denote which speaker this piece of recording comes from;
- Focus response: Not clear yet, need checking
- Term used: the abbreviation[4] of the term used to describe the chip;
- Described chip: the (b, h) representation of the chip described in this recording.

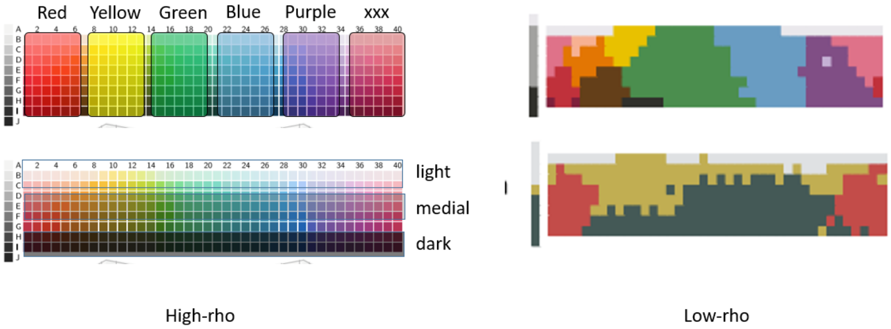
From the raw data, we can directly count the distinct word types in each language and get $N_V(l), l = \{1, ..., 110\}$ for different language or feed these data to the algorithm proposed in [ref] to calculate topological similarity (i.e., $\rho(l)$). But before that, we will provide some statistics and visualizations to give a high-level overview of what the data is (and should be) look like.

[3] The examples in Table 1 are samples stored in *foci-exp.txt*, which stores the best term used to describe each chip in Munsell. Actually, we can also use the raw data stored in *term.txt*, which contains for each consultant and each stimulus chip the term with which the consultant named the chip, to do similar things.

[4] The mapping from the abbreviation form to the full form is stored in *dict.txt*.

3.2 WCS data visualization

I am not quite familiar with this part. I should first carefully think about what information I plan to deliver in this subsection. The goal of it is to show readers what is high ρ language looks like in the context of color-naming task, like something in the following figure 3

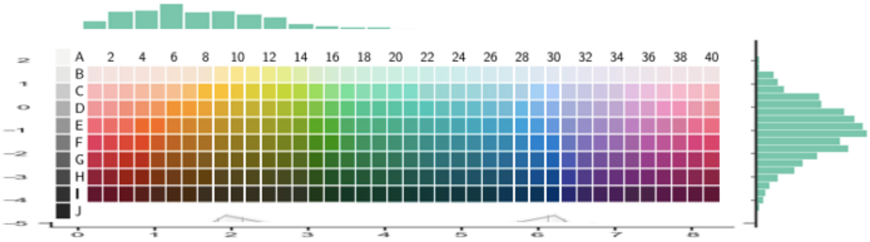


High-rho

Low-rho

Figure 3: Example of WCS data visulization on what is high-ρ or low-ρ language looks like.

Also, I plan to provide some statistics of the WCS data, e.g., when selection the best term, which part on Munsell has most/least disagreements. The figure might like this:



We may merge some chips on the Munsell, and use some other channel to represent the ratio of # winning name to the # total

Figure 4: Example of WCS data statistics on how participants agrees on different region.

Some fundamental regression chart, scatter chart can also used here. Again, the core problem for me is to determine what information I want to deliver. I think I should read more paper about this, and see what the statistics of the data really looks like.

Remember to add some figures to show the calculated N_V and ρ, even though the comparison are mainly discussed in chapter 5.

4 SimLang data and statistics

As mentioned before, the WCS dataset suffers from two limitations when studying the language evolution problem, i.e., the size of the dataset is relatively small and the detailed chronological information of different languages are also hard to obtain. To tackle this problem, the authors of [ref] proposed a computer simulation tool called SimLang, which can simulate the language evolution procedure on a referential task [ref]. In this section, we will first introduce what the data samples generated by SimLang looks like, so as the differences between data obtained from SimLang and WCS project. Then we will provide some statistics of the SimLang data to prove it's robustness under different environmental settings.

4.1 Raw data generated by SimLang

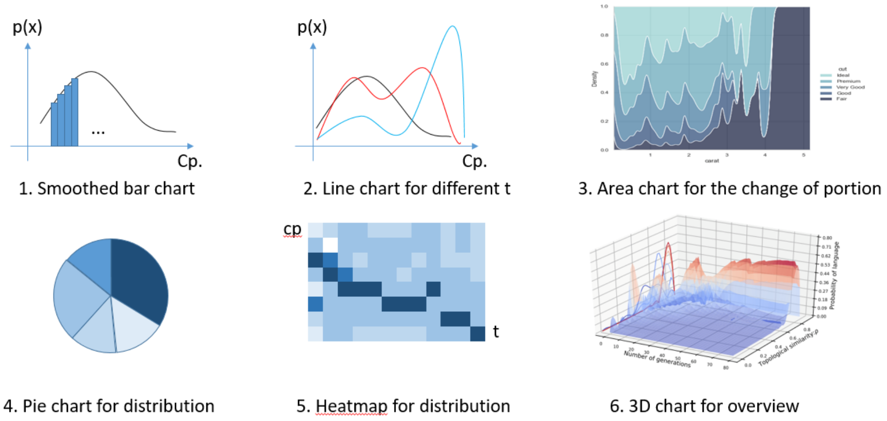
In the SimLang, we can create several neural agents (agents implemented using neural network) and let them play a referential game. In this game, the agents must use phrases (the words are selected in a vocabulary) to encode each chip in Munsell. Then the agents should use these phrases to communicate with each other to accomplish some tasks, e.g., selecting the referred chip given the phrase. The agents will play the game for several rounds, and the language (i.e., the mappings from chips to phrases) will gradually converge. After the game playing phase, we may create some new agents (we call them a new generation of agents), and let them learn from the old agents. This simulates the evolution process of language through time. Thus from SimLang, we can not only analyze languages at the end of one specific generation, but can also analyze how languages changes generation by generation, which is quite helpful in evolutionary linguistics.

From the above introduction, we know that the raw language generated by SimLang from a single generation can have the same form as the examples illustrated in Table 1[5]. The main differences between WCS and SimLang data is the meaning of language ID. In WCS, that is only a categorical value but in SimLang, it can be a chronological index. Furthermore, we can change the environmental settings of the game used in SimLang and generate different groups of languages. Hence we add an additional column (called ”settings ID” to Table 1) to distinguish language generated from different system settings.

4.2 Statistics of the data generated by SimLang

Different from section 3.2, which provides a plenty of visualizations on what the raw data of WCS project looks like, we will provide some figures on the statistics of the results gathered from SimLang. The statistics about terms usage, number of word types, or topological similarity can demonstrate that the simulation result is stable and can indeed represent some trends of language evolution.

Here I plan to re-consider what statistics to present and how to present them. The first goal is to show the robustness across different settings, the second goal is to show some trends that cannot obtained from WCS data. Figures might like this (still need to think about how to effectively show robustness:



1. Smoothed bar chart

2. Line chart for different t

3. Area chart for the change of portion

4. Pie chart for distribution

5. Heatmap for distribution

6. 3D chart for overview

Figure 5: Example of statistics of SimLang data.

5 Compare data from WCS and SimLang

To goal of SimLang is to provide larger amount of simulated language data in when system settings varies (including time) which cannot offered by WCS project. So whether the simulated data can correctly represent some important trends of the real data is the premise of using it in research.

[5] The language ID can be used to denote the number of generation, the speaker ID is the index of agents.

Hence in this section, we use information visualization tools to compare the data from this two approaches from different perspective.

To make a good comparison, I plan to provide two groups of figures. The first group is like the upper line in the following figure, in which we parallel draw two figures from different dataset and show that they share a lot of similarities. The figures can be about some statistics, distributions, etc. The second group is like the bottom line of the following figure. In this one, each generated language will be represented by a 2-tuple (e.g., $(N_V(l), \rho(l))$, i.e., a point on x-y plane. We might use regression to see their correlations.

What's more, we can use the techniques learned from the discussion of Ballotmaps to show the bias in both WCS and SimLang, we hope they share similar bias, e.g., more modern language tend to be more compositional. Or some bias in WCS dataset, e.g., whether the age/gender of the participants influence their results (depends on whether it is correlated with the main topic of this paper).

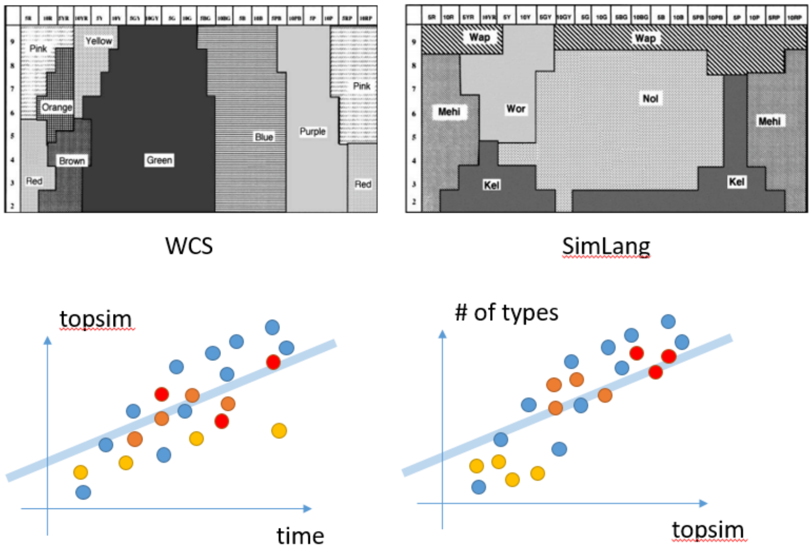


Figure 6: Example of comparison between two systems.

Furthermore, I plan to put a figure about how language evolves with time using the data generated by SimLang, their data can be drawn in the line form. The data generated from WCS migth also on this figure, but they are only some scatter points. Even though the language in WCS do not have a clear time stamp, I do remember some linguists' research have a cursory chronological order for these 110 languages. I might take some time to find this.

6 Conclusion

To be done.

References by topic:

WCS and color-naming:

(Regier et al., 2007); (Zaslavsky et al., 2018); (Lindsey & Brown, 2009); (Gibson et al., 2017); (Richard Cook & Regier); (Tan et al., 2008); (Franklin et al., 2008); (Berlin & Kay, 1991)

SimLang and Evolutionary Linguistics:

(Havrylov & Titov, 2017); (Lazaridou et al., 2018); (Brighton & Kirby, 2006); (Andreas, 2019); (Cogswell et al., 2019); (Kirby et al., 2015); (Smith, 2002); (Wagner et al., 2003); (Kirby & Hurford, 2002); (Kirby, 2007); (Ren et al., 2020)

Info vis tools:

(Munzner, 2014); (Wood & Slingsby, 2011)

7 Milestone

As we have a strict timeline for submission, the milestone of this project should follow it.

Updates due (22 days, Oct. 24 to Nov. 17):

- Write code for WCS dataset. (to Oct.30)
- Use SimLang to generate data, analyze data. (to Nov.07)
- Based on some analysis of data, design what the figure in section 3,4,5 should look like. (to Nov. 14)
- Add these new results and figures to these chapters. At least complete section 3. (to Nov. 17)
- Make slides for peer review (to Nov. 19)

Post-update due (9 days, Nov. 17 to Nov. 26):

- Add these new results and figures to these chapters. At least complete section 4. (to Nov. 23)
- Design what should be in section 5 (to Nov.26)

Final presentations due (14 days, Nov. 26 to Dec. 10):

- Accomplish section 5. (to Dec. 3)
- Make PPT (to Dec. 10)
- Polish report. (to Dec. 14)

References

Jacob Andreas. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019.

Brent Berlin and Paul Kay. *Basic color terms: Their universality and evolution*. Univ of California Press, 1991.

Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2):229–242, 2006.

Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. Emergence of compositional language with deep generational transmission. *arXiv preprint arXiv:1904.09067*, 2019.

Anna Franklin, Gilda V Drivonikou, Ally Clifford, Paul Kay, Terry Regier, and Ian RL Davies. Lateralization of categorical perception of color changes with color term acquisition. *Proceedings of the National Academy of Sciences*, 105(47):18221–18225, 2008.

Edward Gibson, Richard Futrell, Julian Jara-Ettinger, Kyle Mahowald, Leon Bergen, Sivalogeswaran Ratnasingam, Mitchell Gibson, Steven T Piantadosi, and Bevil R Conway. Color naming across languages reflects color use. *Proceedings of the National Academy of Sciences*, 114 (40):10785–10790, 2017.

Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems*, 2017.

Simon Kirby. The evolution of meaning-space structure through iterated learning. In *Emergence of communication and language*, pp. 253–267. Springer, 2007.

Simon Kirby and James R Hurford. The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language*, pp. 121–147. Springer, 2002.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations*, 2018.

Delwin T Lindsey and Angela M Brown. World color survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences*, 106(47): 19785–19790, 2009.

Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

Terry Regier, Paul Kay, and Naveen Khetarpal. Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104(4):1436–1441, 2007.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. *arXiv preprint arXiv:2002.01365*, 2020.

Paul Kay Richard Cook and Terry Regier. Wcs data archives. URL `http://www1.icsi.berkeley.edu/wcs/data.html`.

Kenny Smith. The cultural evolution of communication in a population of neural networks. *Connection Science*, 14(1):65–84, 2002.

Li Hai Tan, Alice HD Chan, Paul Kay, Pek-Lan Khong, Lawrance KC Yip, and Kang-Kwong Luke. Language affects patterns of brain activation associated with perceptual decision. *Proceedings of the National Academy of Sciences*, 105(10):4004–4009, 2008.

Kyle Wagner, James A Reggia, Juan Uriagereka, and Gerald S Wilkinson. Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1):37–69, 2003.

Badawood D. Dykes J. Wood, J. and Slingsby. Ballotmaps: Detecting name bias in alphabetically ordered ballot papers. *IEEE Transactions on Visualization and Computer Graphics*, 2011.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942, 2018.