Information in Language

Final Project Proposal
Undergraduate Laboratory at Berkeley
Data Science ULAB
Linguistic Leopards
Due December 6, 2020

Mentors

Amrut Nadgir
Jeremy Yeung

Group Members

Varun Agrawal
Harshika Jalan
Emilyne Kim
Jiaxun Li
Ethan Qiu
Srinath Rangan
Siddhant Satapathy
Cameron Sylber

Certification

I certify that this project proposal is an original collaborative effort that I have carried out in conjunction with only those listed here and with the assistance of the subgroup mentors. I affirm that I have properly cited all references, including journals, textbooks, and other resources.

Amrut Nadgir [AN]
Jeremy Yeung [JY]
Varun Agrawal [VA]
Harshika Jalan [HJ]
Emilyne Kim [EK]
Jiaxun Li [JL]
Ethan Qiu [EQ]
Srinath Ragan [SR]
Siddhant Satapathy [SS]
Cameron Sylber [CS]

1. Background

Information theory was first brought to worldwide attention by Claude E. Shannon in 1948. In “A Mathematical Theory of Communication,”[1] Shannon proposed a mathematical model of the statistical processes underlying information theory. Since then, main ways of measuring information and complexity have involved analyzing underlying categorical distributions of different aspects of the language. This method can only be applied in limited cases due to a lack of experimental data; thus, researchers have tried to find new ways to represent language.

In 1997, Siegalmann[2] proposed a more flexible approach -- using the neural complexity of a recurrent neural network to learn different simple languages. Neural complexity is defined as the number of nodes a neural network has. This approach was validated by Kon and Plastoka (2000)[3], in which they state that neural complexity measures lower bounds for information needed about the desired input output function. Therefore, in order to achieve the most efficient and effective results, we will employ neural networks in our research.

The study of linguistics utilizes scientific methods by which languages are structured. Although different languages follow a pattern of wide but constrained variation, all languages are shaped by the need for communicative efficiency -- which is achieved at the optimal tradeoff between simple and informative systems. In other words, languages should support effective communication by being simple (minimizing cognitive load) and informative (maximizing communicative effectiveness). This is most amply seen in naming colors across different languages. Since color is a continuous domain[4], how colors are named has led to tenuous debate on language and thought. The difficulty in differentiation of color has much to do with hue and its connection with saturation and value, which makes the range of specific dominant colors extend across the color spectrum further than others.

2. Research Question

We are trying to devise a method to easily understand and measure the information carried in language. We will attempt to measure the information content of various aspects of language by modeling them with a neural network optimized for minimal complexity.

In the paper “Efficient compression in color naming and its evolution,”[5] Regier and other researchers demonstrate that color naming conventions in multiple languages maximize efficiency by optimizing between complexity and accuracy. Regier compares the complexity of these languages using mutual information. This serves as the basis for our project.

This type of complexity measurement can be difficult to compute with real-world data because it requires knowledge of an underlying probability distribution. Using neural networks offers an advantage because they can be modified to closely resemble the brain and help understand information bounds from a neuroscience perspective. We believe that neural networks may also be an applicable tool to understand the complexity of language due to its similarity with the brain.

To get meaningful results, we will model color naming conventions in different languages/cultures and determine their complexities. Using our technique, we will attempt to recreate some of the information theoretic aspects of Reiger’s work that are seen in the Figure[5] below.

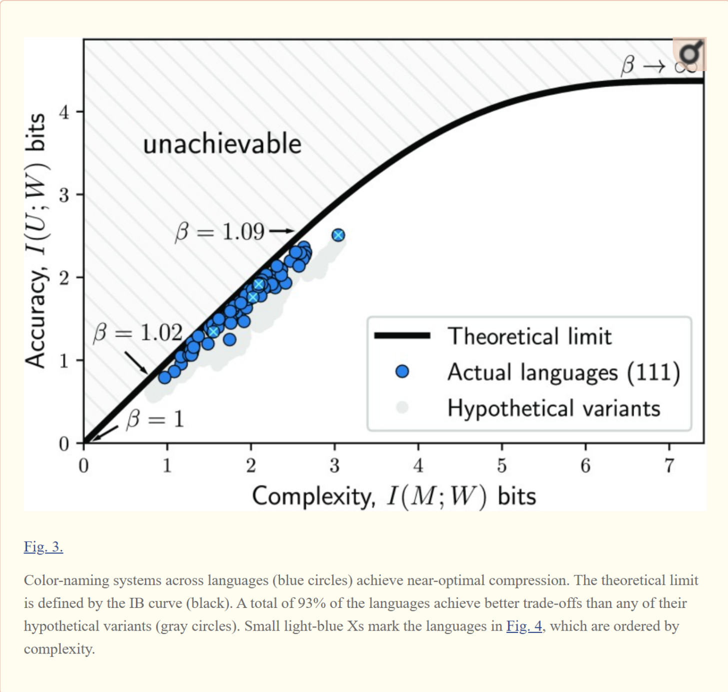

Fig. 3.

Color-naming systems across languages (blue circles) achieve near-optimal compression. The theoretical limit is defined by the IB curve (black). A total of 93% of the languages achieve better trade-offs than any of their hypothetical variants (gray circles). Small light-blue Xs mark the languages in Fig. 4, which are ordered by complexity.

Figure: We will recreate the plot above but replace the mutual information complexity on the x-axis with our notion of complexity.

We will ensure that our method yields the same results as the exact mutual information by analyzing our minimal complexity networks using the Information Bottleneck principle[6] and comparing their complexity in bits to the complexity calculated using mutual information.

We will also compare the similarities of different color naming conventions by training a minimal complexity network to recognize color naming conventions in multiple languages simultaneously. In addition, we will include noise to gauge the robustness of different color categorizations.

Should our method prove successful, we will also build on Regier's experiment regarding the complexity of the portrayal of kinship[7].

In particular, our experiment will serve as a proof of concept for using neural networks optimized for minimal complexity as the basis for measuring similarity and complexity across languages.

3. Methods

(1) Modelling Color Naming Systems:
Instead of constructing the source distributions, here we attempt to create a minimal complexity neural network that models the color naming conventions in each language. To do this, we will one-hot encode the color words and pass it into the network. We will also pass in the x and y coordinates of a color chip into the network, and we expect the network to output 1 if the color name describes the associated color name category and 0 otherwise. For example, if a language has two colors, red and blue, the input layer of the network would consist of 4 nodes. The first two nodes would have inputs of either 1 and 0 or 0 and 1 depending on whether the color was red and blue. The last two nodes will have inputs that define the x and y coordinates of the color chip. These inputs will take on evenly spaced values between -1 and 1. We choose to use both coordinates of the color grid instead of parametrizing it with a single coordinate because, with both coordinates, nearby colors will have similar coordinates, so it will be easier for the network to learn the structure of the naming conventions.

Since the size of our training datasets for each language is relatively small, we can use the entire dataset to train our networks.

The outputs of the network will be fed through a sigmoid function to represent probabilities. Then, they will be rounded to either 1 or 0. A successful network will output 1 if the coordinates of the color chip matches the input color name and 0 otherwise with no error.

(2) Measuring Complexity:
To find such a network, we adjust the number of hidden layers and the number of neurons per layer to find out the simplest network structure that is able to represent the “encoder” of the language. When defining the simplest network structure, we are using the “neural complexity”[3] which is equivalent to the number of nodes in a network. Our goal is to find the network structure with the lowest neural complexity that achieves perfect accuracy on the color naming data for a given language.

(3) Dataset:
We will use the same WCS data as the one used in Regier’s paper “Efficient Compression in Color Naming and Its Evolution”, obtained from www1.icsi.berkeley.edu/wcs. As seen in the figure below from the paper, “Word meanings

across languages support efficient communication”[7], for each language, each color chip will have a single word associated with it.

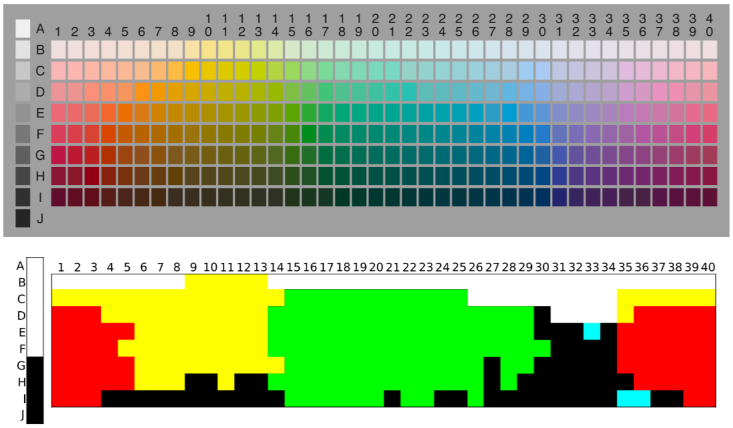

Figure 3: (Upper panel) Color naming stimulus grid. (Lower panel) Mode map for the Iduna language (Austronesian, Papua New Guinea), mapped against the stimulus grid. Each colored blob indicates the extension of a color term in this language. There are 5 major color terms, corresponding roughly to black, white, red, yellow/orange, and green/blue. There are also a small number of chips for which the modal response was another category (shown in light blue).

(4) Further research - 1) similarities of different color naming conventions & 2) complexity of the portrayal of kinship (using recurrent networks)

For 1), We measure the difference between the summed network complexity required to model each language individually and the network complexity required to model both simultaneously can act as a measure of how similar two languages are. In addition, we will include noise to gauge the robustness of different color categorizations.

For 2), the complexity of kinship could also be portrayed using a similar method with neural networks. RNN would have to be utilized in this instance because of how kinships in various languages utilize multiple words to describe a particular relation. These varied input sizes suggest the use of RNN.

4. Proposed Collaborations

- Terry Regier - Linguistics
  Language and Cognition Lab (https://lclab.berkeley.edu/) -
  - Our lab investigates the relation of language and thought, in computational terms.
  - How is the structure of the human mind reflected in the world's languages? Are there concepts that are universally available to speakers of all languages? Or does language create and shape our thoughts, such that speakers of different languages think about the world in fundamentally different ways? We probe questions such as these, and do so in an integrative fashion that seeks to move beyond simple nature vs. nurture conceptions of the mind. We seek to understand which aspects of the mind shape language, and which are shaped by it - and what general principles govern the interaction of our language and our thoughts.

5. Budget and Equipment

Due to having to train many neural networks, we may need supercomputing time as well as access to machines or software running parallel computation. Thus we may need credit for running our classifier using Google servers or on AWS.

6. References

1. A Mathematical Theory of Communication—Shannon—1948—Bell System Technical Journal—Wiley Online Library. (n.d.). Retrieved December 2, 2020, from https://onlinelibrary.wiley.com/doi/10.1002/j.1538-7305.1948.tb01338.x
2. Siegelmann, H. T., Sontag, E. D., & Giles, C. L. (1992, September). The Complexity of Language Recognition by Neural Networks. In IFIP Congress (1) (pp. 329-335).https://doi.org/10.1016/S0925-2312(97)00015-5
3. Kon, M., & Plaskota, L. (2000). Information complexity of neural networks. Neural Networks, 13(3), 365-375. doi: 10.1016/s0893-6080(00)00015-0
4. Masculine or Feminine? (And Why It Matters). (n.d.). Psychology Today. Retrieved December 2, 2020, from http://www.psychologytoday.com/blog/culture-conscious/201209/masculine-or-feminine-and-why-it-matters

5. Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. Proceedings Of The National Academy Of Sciences, 115(31), 7937-7942. doi: 10.1073/pnas.1800521115
6. Tishby, N., & Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle. ArXiv:1503.02406 [Cs]. http://arxiv.org/abs/1503.02406
7. Regier, T., Kemp, C., & Kay, P. (n.d.). Word meanings across languages support efficient communication. 22.