# Galactica: A Universal Map of Understanding

Dorian Spiegel
dspiegel27@sjs.org

December 31, 2023

**Abstract**

Modern languages, with their arbitrarily constructed systems of characters and symbols, present a significant challenge in learning and understanding. This challenge arises from the lack of structure or similarity between alphabetically (or symbolically) similar words. For example, words with similar meanings such as "truck" and "car" share no phonetical resemblance. This paper introduces Galactica, a language purely based on the intuitive similarity between words. Galactica leverages modern natural language processing to create a visual map of understanding, based on semantic similarities between words. Although still non-conclusive, our research presents the idea that this language may be much more efficient to learn and understand and allows for more nuanced expression with translations that don't exist in languages. This paper details an overview of Galactica, and the initial research and findings that may support a more comprehensive study in the future.

## 1 Introduction

Almost all languages are based off vocabulary and symbols created thousands of years ago. Latin, the basis for most modern languages, was created in 7th century BCE. These languages are extremely primitive, there is almost no correlation to intuitive meaning and phonetical or morphological similarities, making learning and understanding these languages more difficult than neccessary. "Buy" and "purchase" have the same intuitive meaning but no phonetical similarity; "Flour" and "flower" have both phonetical and morphological similarity but mean completely different things. Another problem is that the world doesn't have a universal language, leaving billions of people unable to communicate without translations.

However, with the advent of natural langauge processing and modern word embeddings, there's an opportunity to reexamine the primitive languages and create something more universal and comprehensive.

This paper introduces 'Galactica', a framework of understanding based off the semantic similarities between words as quantified by natural language processing. A randomized map of symbols and shapes is generated, and the points

on that map that correspond to different understandings (or words) is created by reduced dimension word embeddings created by Word2Vec. Simply, words with similar meanings will have closer points on the map, and words with farther meanings will be farther away on the map. A symbol (or word) is created taking a subsection of the map centered around a point, with a visual transform to make it easier to comprehend.

Our extremely preliminary research explores the feasability of Galactica, whether its possible to remember the symbols themselves, and correlate the position on the map to different understandings. While our findings are initial and need more statistical rigor, they present the idea of a universal map of understanding, unbounded by a list of words or languages. This paper details the conceptual and technical foundation of Galactica, and the initial promising findings.

## 2 Word Embeddings

Natural language processing (NLP) word embeddings are a technique used to represent words as a dense series of number or vectors. Common embedding models such as Word2Vec, Glove, and FastText learn these embeddings from text data, and allow similar words to have similar embeddings representations. These vectors are typically multidimensional (often in the hundreds of dimensions), so it's hard for us to comprehend these vectors on a graph in their raw form.

However, we can use statistical methods to reduce these high-dimensional vectors into things we can graph and understanding. t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique for converting high dimensional data (like our word embeddings), into two and three dimensional points.

Using Word2Vec and reduced dimensionality with t-SNE, here is a visual map of 20 common fruits, vegetables, foods, and drinks.

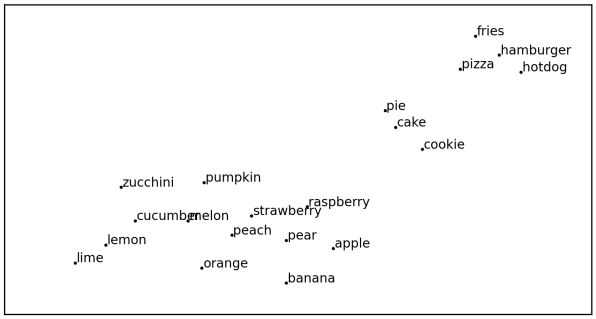


Figure 1: t-SNE map of 20 common fruits, vegetables, foods, and drinks

It's also important to note that t-SNE becomes more accurate the more data it has because it's a machine learning algorithm, so expect a graph with thousands of points to be extremely accurate compared to Figure 1.

# 3 Map

In order to turn these points into symbols, we need to create a map of symbols and shapes that are easily human recognizable. We can use a thresholded and simplified map of Perlin noise. Perlin noise is a gradient noise generation, meaning it interpolates between gradients to create a smooth pattern.

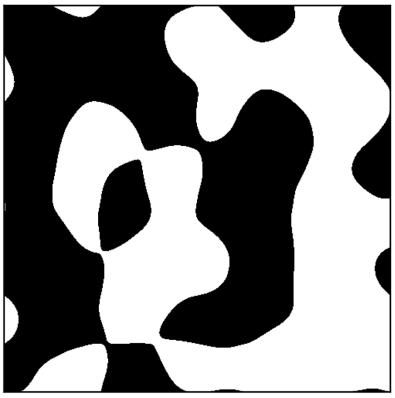

Figure 2: Perlin noise map (500x500), with 3 octaves, 0.1 persistence, and 2.0 lacunarity, thresholded at 0

# 4 Galactica

With a map and points that can be plotted in the same dimensional-space, we can start to create our symbols. First, we plot all our t-SNE points on the map:

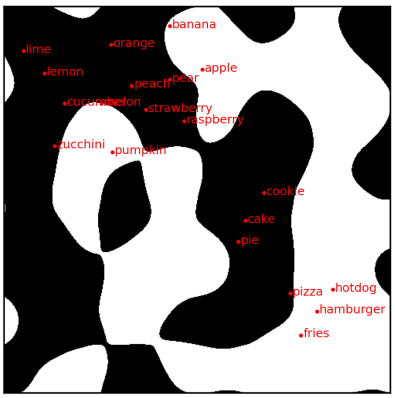


Figure 3: Figure 1 normalized and plotted on Figure 2

If we "zoom" in on two submaps centered around a point, we start to see how these points can be used as symbols.

Figure 4: Symbol on the left is “cookie”, and the symbol on the right is “pie”

After some brief testing (with an unbiased participant), they anecdotaly said that reading symbols in this form was extremely difficult. Instead of limiting ourselves to a small subsection of the map, we can apply a fisheye (or barrel) transformation to the image so the center of the point is zoomed in, but you can still see a wider expanse of the map.

Figure 5: Symbol on the left is “cookie”, and the symbol on the right is “cake”

The general concept is that humans will be able to learn this map, and base their understanding off of their location on the map. The fisheye transformation gives the ability to see minute changes in position (the extremely zoomed in part), but also allows us to see other landmarks in the distance that help us base our position. **Note:** some participants expressed that they couldn't comprehend the fisheye transformation (they thought it was a sphere), but others found it exremely helpful. Additionally, making a colored map is also possible which would make landmark-identification even easier. The most interesting part of using these locations is it is possible to interpolate between these points and create symbols with shared or new meanings and nuance. The amount of words possible to interpolate is virtually infinite, which leads to lots of exciting possibiliites.

# 5 Feasability

Given that this is such a novel way of thinking and understanding language, it is first neccessary to show that it's possible for humans to correlate geographical context to understanding

64 symbols were generated from a list of common fruits, vegetables, drinks, and foods. These were shown to two participants in two randomly selected batches via a learning algorithm. The first batch was 44 symbols, and the second batch was the remaining 20. The learning algorithm randomly shows the participant a symbol, which the participant has to try and guess. If the participant guesses correct, that symbol is removed and considered "learned". If the participant guesses incorrect, the algorithm tells the user the correct answer and the cosine and euclidean distance they were to the correct answer (how close the meanings are)

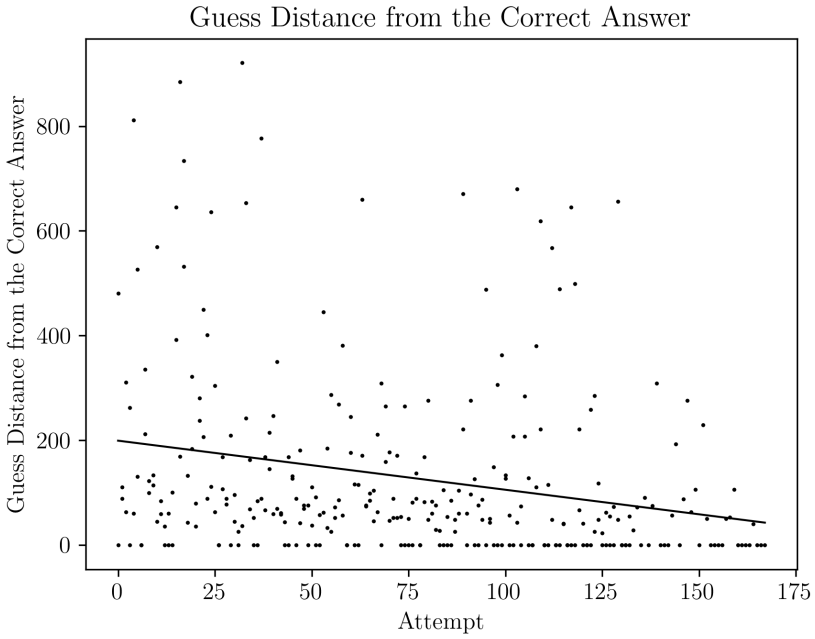


Figure 6: Shows the distance (euclidean) from the guessed answer to the correct answer. Initially the distance to the correct asnwer and correct answers (labelled as 0 distance), are sparse. The distance to the correct answer and number of correct answers increases as time goes on.

Although still preliminary, the results are promising. The first participant was able to learn 44 symbols in 136 attempts, compared to 182 attempts for the same set of words in Japanese. The second participant learned 44 symbols in 166 attempts, and they have not tested Japanese yet. Both participants began to learn the map, and the distance to the correct answer decreased consistently as time went on. The increase of accuracy in the participants shows that they are learning the map, it's landmarks, and gaining an intuitive feel for it.

After the first 44 symbols, particpants were shown another 20 symbols in similar geographical space they had not seen before.

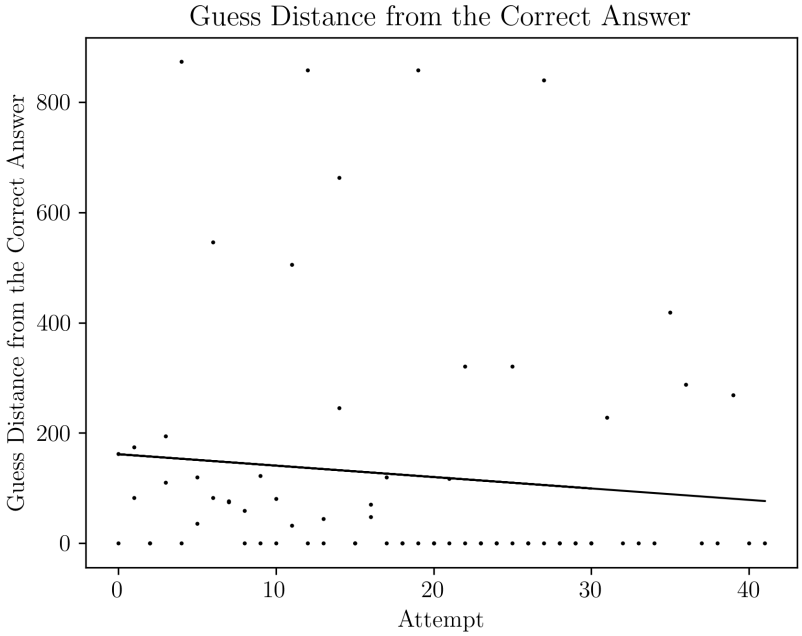


Figure 7: Shows the distance (euclidean) from the guessed answer to the correct answer. Participants learned new words extremely rapidly, with many words being learned the very first time seeing them.

The participants performed extremely well with these new words. Participant one gueussed 10 words correctly on the first try without ever seeing them before, and the learning rate of these new words was significantly faster than the first 44 (across both participants). They took 31 and 42 attempts (respectively) to learn these 20 words.

# 6 Conclusion

This exploratory research presents the idea of Galactica, and shows that it has a promising future.

Particularly, the incredible speed participants could learn new words after being exposed to the map is extremely promising. **Both participants' ability to intuit new words without ever seeing them before is the most promising result of this research.** The three main points that need to be proven are:

- The learning rate of Galactica is faster than traditional languages
- We can intuit new words and understandings purely based on their location on the map
- We can read and understand information faster via Galactica than traditional languages

This study provides preliminary evidence that Galactica can be learned faster, and intuiting new words is possible. However, it's important to stress the results are extremeley preliminary, and need to be proven with more subjects and a larger vocabulary to support the claims with more statistical rigor.

Ideas for future research include:

- Use a larger vocabularly, larger map, and more participants to prove that it can be learned faster than traditional languages
- Switch from using nouns to verbs, and create sentences/strings of words that don't have direct translations to further prove we can intuit new words
- Switch from Word2Vec to BERT or another embedding algorithm that embeds multiple words at once to see if we can quantiy multi-word phrases
- Analyze other factors that could potentially make the map more efficient: colored-map, different visual transformations, etc.
- See if it possible to graph words in different languages on the map, creating a universal language