THEORY OF A FUNDAMENTAL LANGUAGE

Introduction

The invention of Generative Pre-trained Transformers **(GPT)** is an impressive leap in computer science.  Its linguistic abilities have led many to describe it as having “understanding” or “cognitive ability” since it feels like someone is chatting with you when in fact there are actually statistical methods, machine learning algorithms and search engines working in the background. When it has “hallucinations”, tweaks are added to try and eliminate them. One can view GPTs as statistical engines applied to language.  They may discover, and undoubtedly will, new relations in its domain, outputting a range of linguistic statistical connections previously unstudied. Yet, what is their connection with Natural Languages (**NL**)? What is the connection of NL to knowledge in general? This latest question reduces to the ancient one of “Do we think in words?” As a former simultaneous interpreter, I have a strong opinion on that, and as an inquirer in the fascinating world of AI, I also have a point of view, and the reasoning behind that point of view is the subject of this paper.

The Concept

From the perspective of an interpreter, I am often asked to check the Large Language Model **(LLM)** translation in context of one word in one language into another language. Often, the reply is processed, “Well, it is either this word or that word, which is closer?” This mental process is common not only to professionals, but also to most multilingual speakers. LLMs do an adequate job of this process, but the mechanism in GPTs, being purely statistical, is often inaccurate. Language at its base is not statistical, but much deeper. Language at its base is as deep as the universe itself.  I will call this language the Fundamental Language (**FL**). In my view, FL is a form of interaction. FL is essentially the *rules of communication*, as the rules of physics are dictated by quantum mechanics **(QM)** and General Relativity **(GR)**. These are all laws that were either created or instantiated in the first few seconds of the universe.

FL is a theoretical framework or a set of universal rules that underpins and structures information systems. It is meta-linguistic and meta-communicative in nature. It serves as the scaffold upon which various forms of knowledge and interaction are built. FL is concerned with the fundamental principles that govern the instantiation of interactions, whether they be linguistic, visual, symbolic, or quantum. It is an attempt to formalize the mechanics of these interactions in a mathematically rigorous way, using tools like Voronoi diagrams ([Voronoi diagram - Wikipedia](https://en.wikipedia.org/wiki/Voronoi_diagram#:~:text=In%20mathematics%2C%20a%20Voronoi%20diagram%20is%20a%20partition,in%20the%20plane%20%28called%20seeds%2C%20sites%2C%20or%20generators%29.)) to model these relationships. In essence, FL is a generative system that can produce a variety of "languages" or systems of representation, encompassing natural languages, spatial-temporal representations, and even the symbolic systems of quantum mechanics. FL is the underlying structure or the rules of the game, while communication is the game being played. NL is the human instantiation of the game. FL is more abstract; it defines the principles that apply across all possible forms of communication. Communication is an instantiation of those principles, manifesting in diverse forms, from spoken languages to the dance of subatomic particles.

There is therefore a hierarchy where FL together with the other laws present at the start of our universe, stand at the top. Below the laws of meta-laws of physics are the laws of QM and GR, for example. I use the term meta-law in that classical mechanics is a law under QM. We do not yet have a unified theory of QM and GR. We do not know for certain that QM or GR stand at the top, but something must. QM is a meta-law in that it (fully) describes classical mechanics. Below QM, for example, there are the meta-laws governing biological systems. Below that, more branches, including the laws governing NLs. Finally, below that are LLMs. So, what is the relationship between QM or GR to FL? QM, RQM GR etc. are formalizations of empirical knowledge. This formalization is a mapping from the abstract to the concrete. This interaction is found not only at an analytical level, but also at the level of expressing one’s ideas to communicate them to others. This is what FL studies.

QM and FL as symbolic systems share fundamental questions, importantly among them, “what reality do they describe?” Is that question even answerable? In QM, the approach advocated by Carlo Rovelli ([Carlo Rovelli - Wikipedia](https://en.wikipedia.org/wiki/Carlo_Rovelli)), known as Relational Quantum Mechanics **(RQM)**, is the best metaphor for considering the question in FL. In RQM, quantum states are relational information between the “observer” and the “system being observed”. All that is knowable is solely because of interaction. It is the knowable sense that mirrors that of FL knowing. FL is about the instantiation of interactions. Far down the hierarchy of FL, for example, the “system being observed” is the internal state, the “mind”, and the “observer” is the natural language. For multilingual people, going between languages always passes through the “state of mind”. FL is what allows the interactions described in RQM, or any QM system to be expressed, to be symbolized, to be written and to be instantiated. RQM, or any QM system, does not exist by itself, it only exists in relation to the ability of something to interact with it. In this way, Ideas are “theories”, in that they are abstract realities for communication with others.

FL studies the nature of theories’ interactions. These interactions have structures. They interact through the rules of FL to yield structures that build transmittable knowledge. These rules are algorithms, thanks to the Curry-Howard correspondence ([Curry–Howard correspondence - Wikipedia](https://en.wikipedia.org/wiki/Curry%E2%80%93Howard_correspondence)). The Curry-Howard correspondence, also known as the Curry-Howard isomorphism, establishes a direct relationship between computer programs and mathematical proofs. Accordingly, every proof in a constructive logic corresponds to a program in a functional programming language. Conversely, every program corresponds to a proof. This is often summarized by the phrases "proofs as programs" and "programs as proofs."

Thanks to Landauer’s principle ([Landauer's principle - Wikipedia](https://en.wikipedia.org/wiki/Landauer%27s_principle#:~:text=Landauer%27s%20principle%20states%20that%20the%20minimum%20energy%20needed,given%20by%20where%20is%20the%20Boltzmann%20constant.%20)) regarding the minimum energy limit for any computation, we also know that computation is real (assuming non-reversibility, which would not change our discussion here). QM, All types of relativity, RQM, and FL are physical entities, are thus physical realities. If they are physical realities, then they can be studied with the tools used to study physical reality. (Train of thought: The Curry-Howard isomorphism provides a bridge between mathematical logic and computation. It proves a one-to-one correspondence between mathematical proofs and program types. Landauer then links computation to energy expenditure.) QM’s various interpretation are all based on equations, hence on computer programs, hence they are physical. Program types are real. In the context of FL, this isomorphism implies that every valid 'statement' or 'interaction' within the FL framework can be considered a type of mathematical proof. Therefore, FL is not just a symbolic or linguistic construct, but physical entity grounded in a physical reality and mathematical rigor. These building blocks allow FL to interact with quantum systems like RQM, making FL a part of the physical world, open to study and exploration using the tools of QM. The interaction of any formal system (QM, etc.) and FL will yield a physical system.

Later in this presentation, I will be using Voronoi graphs to model FL. It must be kept in mind that because of the preceding discussion, these will also have thermodynamic consequences. In David Deutsch’s Constructor theory ([Constructor theory - Wikipedia](https://en.wikipedia.org/wiki/Constructor_theory)), this would be a possible transformation. One might say that the analogy from QM’s measurement to observation would be calculation of interaction to position. FL is the mapping from some finite space of an actual universe to its symbolic concretization. FL studies the interaction of the objects, the things that are knowable, as its arguments, and produces its output, what is described, what is “known” in the sense that it is somehow proven.

FL’s models model of our present knowledge, be that of physics or some non-empirical realms such as mathematics. The realm of physics, possibly the foundation of physical reality, is built from the most basic constituents, from the most fundamental knowledge, (from Newtonian mechanics to QM). Yet the explanatory value of QM and GR, the most basic explanations of existence, must be at its core. This hierarchy can be decomposed (as we shall see using Voronoi diagrams). Using LLMs (LLMs can now be viewed as functions), let’s generate embedded vectors of these laws. This process also shows how FL has the characteristics of a Von Neumann generator ([Pseudorandom number generator - Wikipedia](https://en.wikipedia.org/wiki/Pseudorandom_number_generator)). Dimensionalities will be discussed separately.

Decomposition using Voronoi Diagrams

Human understanding of the universe has evolved from basic observations and reflections to advanced theories like QM and GR. The goal of FL is to model this progression using **vector embedding**, for example in the case of NL. Here, each step in our collective knowledge can be considered a theorem, as discussed above. Each theorem is comprised of one or more vectors. These vectors encapsulate the reduction in entropy or uncertainty about our questions concerning the universe. In QM, the concept of a smallest change is analogous to Planck’s constant. A similar foundational 'quantum' in FL could emerge from its design and representation, which will be discussed later. For these mappings, Voronoi diagrams serve as the mathematical framework.

For the sake of simplicity, let's first focus on a 2-dimensional representation. In this case, the vector embeddings of rules (or algorithms) will generate vertices in a Voronoi diagram. The Voronoi cells can be thought of as 'units of knowledge' within the FL 'balloon.' Each cell naturally segments the space based on proximity principles, much like natural laws or theorems in QM or mathematics. For example, a mathematical theorem ***T*** derived from the set of theorems ***T***(i) to ***T***(j) has cells ***T***(i→j)) as its neighbors. Although we start with a 2D example, these principles can be extended to higher dimensions to capture the complexity of FL's representation more accurately. Each Voronoi cell can be related to the 'amount of uncertainty' or entropy associated with a particular law or theorem compared to its neighbor. A Delaunay edge between adjoining cells Cell ***T***(i)) and Cell(***T***(i→j)) is bisected by the boundaries of the cells. This segment's value is a probability vector. The same Delaunay edges also trace 'walks,' which are vector additions in the sense that a path of n adjoining cells can be represented as a vector segmented into n components, with vectors being added head-to-toe. This introduces an additional metric for distances between points, the Voronoi distance (like the Manhattan distance). We will in the future compare Euclidian and Voronoi distances.

A Voronoi cell and its neighbors exhibit distinct relationships based on their connecting paths, represented by their Delaunay edges connecting them. Consider a Voronoi cell at the 'level 0 ', an arbitrary cell as the smallest region formed by the original vector embedding. Let this cell be represented by a prime number p. The neighboring cells can then be assigned successive prime numbers p\_1,p\_2,… using Fortune's algorithm. This aggregation of a Voronoi cell and its immediate neighbors generated by the sweep of Fortune’s algorithm can be termed as a 'level 1' Voronoi region, denoted as V\_1. We can recursively extend this process of sweeping V\_1 to map to an arbitrary level n, generating V\_n recursively generated regions.



Step 1: Create level 0 vector embeddings.

Step 2: Using algorithms such as Fortune sweep, create a Voronoi graph and label it with (Loop: given points (objects), assign ascending prime numbers to them as Fortune sweep creates the graph. Form the composite number of the cell neighborhoods). Each cell will have 2 numbers, its own prime, and the composite of its neighbors.

Step 3: Create a dictionary data structure of the created primes and the associated composite number mentioned.- This would create V\_1 . (These regions represent ways knowledge is aggregated to produce new knowledge.) Recursively generate V\_2 to V\_k (Jupyter notebooks of basic code can be provided). This “zooming out” is analogous to the theorem proving, hence the “discourse (interaction)” mentioned earlier.

These algorithms may also be a better way to construct “intelligent bots”. Prime of primes idea: at each step, start the prime creation anew. Assign to that level the level and the prime. 5, 3 would be the 5th prime of the 3rd prime.

As noted at the beginning of the discussion here, FL is a physical reality. All algorithms about creations of new cells and paths involve energy. We know from Landauer.

Different natural languages have distinct word sets, resulting in unique representations when tessellated in a Voronoi diagram. For example, the English word "shallow" lacks a direct counterpart in Italian, translating to the two-word combination "poco profondo." Similarly, the Italian word "mollica" has no single-word equivalent in English since is refers to “bread crumbs”. These linguistic nuances suggest the existence of intermediate word types in a Universal Language (**UL**).

Mathematically, UL can be viewed as the set-theoretic union of the tokens of all natural languages:

UL =∪┬NL " Tokens(NL) "

In this framework, UL operates on vector embeddings of individual natural languages, essentially serving as a union of all languages. For example, "mollica" would be added to the space corresponding to its Italian embedding. This procedure can be generalized to all natural languages through nearest-neighbor comparisons.

The resulting UL will be a rich set of vectors whose dimensionality is determined by the calculated vector space needed to accommodate all the new vectors. A Voronoi tessellation in this new space will yield cells with new centroids created by the union operation. The cells may either be new "Intersection Tokens", words common to all languages, like the word “man” or “woman”, presumably common to all languages. Or they may be common to a group of languages or found only in one language. (It would be interesting to study the cultural evolution of this tree.) Union Tokens. These distinctions are set theoretic distinctions, and will be notated by subscripts of the tokens, NL\_"language name " . For instance, "mollica" might now be a cell in relation to the concept of 'bread' in English, because the word doesn't exist in English, but does in Italian. Let’s represent the vectors of UL as weighted vectors according to the similarity of their representation. Let’s assign to each different NL a new frame. The new English language frame would have a token for “mollica”. Now any language is NL maximized to describe what any language would represent. The frame will determine the dimensionality of the vector space in UL. All languages would have the same normalized, but weighted set of tokens.

Let us now examine the 'shortest paths' between tokens within the Voronoi tessellation of UL. These Voronoi cells in UL can be considered as 'relativistic cells,' existing relative to their base language. For instance, universal tokens like 'person,' 'man,' or 'woman,' are presumably present in all languages contributing to UL. These tokens serve as a common linguistic ground. The 'relativity' of each token can be quantified by the number of Voronoi cells separating it in UL compared to its native language tessellation. For example, the Italian word 'mollica' exists relative to Italian but not to English. To quantify these observations, we introduce a density function for UL and for each NL 'language name'. The ratio of these densities can be regarded as a 'relative density function,' denoted D\_(NL^' " language name' subscript " UL), which serves as a measure of how a specific language contributes to the overall linguistic landscape of UL."

Let us now examine the 'shortest paths' between tokens within the Voronoi tessellation of UL. These Voronoi cells in UL can be considered as 'relativistic cells,' existing relative to their base language. For instance, universal tokens like 'person,' 'man,' or 'woman,' are presumably present in all languages contributing to UL. These tokens serve as a common linguistic ground. The 'relativity' of each token can be quantified by the number of Voronoi cells separating it in UL compared to its native language tessellation. For example, the Italian word 'mollica' exists relative to Italian but not to English. To quantify these observations, we introduce a density function for UL and for each NL 'language name'. The ratio of these densities can be regarded as a 'relative density function,' denoted D\_(NL^' " language name' subscript " UL), which serves as a measure of how a specific language contributes to the overall linguistic landscape of UL." UL’s Voronoi representation of any individual language may be imagined as a cell of a particular color (or a prime number) all languages have corresponding colors or primes. Composite tokens have composite numbers or colors (which would weigh them). There is therefore a relation between a cell’s assigned number or color and its relative position in the tessellation.

Summary

(FL) is a meta-linguistic scaffold. FL is a framework of universal rules, which are applied to specific instances or "arguments," yielding structured information systems. These systems could range from the familiar constructs of human languages to space-time (2D and 3D Images, Video, etc. (non-NL sounds, e.g.., Music) and eventually to the quantized realms of the domain of quantum mechanics. All of these are instances of changes of state. Through FL I hope to show that they are all equivalent.