

Abstract

What is knowledge and how can it be represented? How does it differ from reality? How does it morph under communication? Definitions for existence, experience, and reality must be confronted. Opinions of philosophers and modern researchers are presented along with a proposal for a mathematical formalism of thought.

“Words, as is well known, are the great foes of reality” (Joseph Conrad) [Conrad, p1]. What is reality? Can language represent reality? What does it mean to *know* something? and why have philosophers from Plato to Wittgenstein never let go of these questions? Even though such queries yield few definitive results, the ensuing discussions can be fascinating.

Knowledge is the only thing of value that one can give away but still retain (communicable diseases are of little value). Some ideas are so simple they can be communicated in a single word. The word “square” probably invokes in your mind an image of a figure with four equal sides. If I see a square, I can say the word “square” to you. You can then *know* what I saw. You can see it in your mind’s eye, as it were. Many complex steps are necessary even for this simple idea to be transported from one mind to another [Jackendoff, p8]. The physiological analysis of these steps will not be the focus of this paper. The focus will be on Knowledge Representation in general (KR): What is knowledge? In what ways can it be organized? and how does it relate to reality?

Plato said that a philosopher is one who’s “passion it is to seek the truth. ... [their] affections are set, in every case, on the reality.” [Plato, p183ff]. He said reality is that which is true. There is also the set of things that are untrue, or unreal. There is yet another set: those things that are neither real nor unreal. By Plato’s terminology (using English words, of course), *knowledge* is the cognition of real things, *ignorance* is the cognition of unreal things, and *belief* is the cognition of those things which are neither real nor unreal. Perhaps Plato inspired Gödel. This third set would contain those unprovable statements which Gödel proved can arise from any formal system. If we are

among those who seek the truth, where do we look? Fox Mulder said, “The truth is out there,” but all each of us has is our internal experience of the world through our senses, each experience creating what Aristotle would call a *quale*, and that quale cannot be truly shared.

It would seem an easy matter to differentiate between knowledge, belief, and ignorance, because the real is provable, the unreal is disprovable, and beliefs are neither. Do I *know* who won WWII? I think I know that the Allies defeated the Axis powers in the European theatre because I read about it, I heard about it, and I’ve seen some war memorabilia. I lived in Europe for 5 years and I don’t have any reason to think differently, but can my knowledge be *proved* in the scientific sense? No. Scientific proof requires repeatability, and I don’t know anybody who wants to repeat WWII. My *knowledge*, therefore, is actually *belief* in Plato’s sense. If WWII was indeed won by the Allies, then the object of my belief is in the set of the real, contrary to Plato’s statement that the object of belief is neither real nor unreal. If I believe that Nazi Germany won instead, can I be scientifically proven ignorant? Even when proof is available, Plato believed that our perceived experience is just a shadow of reality. What we see and hear are mere shadows cast by the real upon our minds through our senses. The writer of *Hebrews* said, “For the word of God is living and active. Sharper than any double-edged sword, it penetrates even to dividing soul and spirit, joints and marrow; it judges the thoughts and attitudes of the heart” [Holy Bible, Heb 4:12]. It is not always easy to divide between knowledge belief and ignorance.

René Descartes believed that the physical world was separate from the mental world. Knowledge for him was not represented in this world. It existed in a non-physical and inscrutable plane. He clearly disagreed with the notion that we are different from clockwork merely in our level of complexity. Putting our mind in another plane protected our status as having dignity, rights, and infinite value. According to him we *are* our minds [Pinker, tBS p8ff]. And these minds exist elsewhere. Hence, we are elsewhere - beyond the status of a highly complex machine. Our mind is inhabited with noble convictions and responsible choices, not the paltry quantities being added and subtracted the brain that Hobbes wrote about. Descartes said understanding in the mental plane corresponds to facts in the physical plane. By this line of thinking there is no

scientific way to study knowledge within the mind. Descartes considered the mind to be wholly indivisible. We now know that the mind comprises many processes and stored facts that can be selectively damaged. Such dissociability of mind abilities shows division and connection to physical brain substrate [Fodor, p47,98]. Even such fundamental things as the ability to treat other people with proper manners can be removed by an injury without affecting perception, memory, language, or motor functions. Just such an injury happened to Phineas Gage in the nineteenth century when a spike went through his ventromedial prefrontal cortex [Pinker, tBS, p42]. The mind is dissociable, and there are definite physical regions of the human brain that are necessary for certain mind functions. Mind, and therefore knowledge, depend on the physical world, and must therefore submit to physical limitations.

One question was satisfactorily settled by Descartes: Do we exist? This question of course bears upon the topic of this paper. If the *real* doesn't even exist, then why are we talking about it knowledge of the real? His logic that an evil deceiver who made us believe we were real would in fact require a victim - a real one. If we think we exist, then we must. He did not successfully convince us, however, that a human mind can't be just a brain floating in a vat with all our perceptions input directly into it a la *The Matrix*. The distance between thought and reality remains significant. Solipsism purports that self is the only thing that can be known and verified. I know I exist. I know I know things. I believe that others exist as well, but do I really have proof? Could others exist merely in my senses? Is there any point discussing a reality outside of our senses?

Aristotle said that there is a direct connection between an object in reality and that object's representation in our minds [Haugeland, p16]. Pictures in our mind represent things that exist. If you imagine the house in which you grew up, you can walk through the different rooms, or count the windows, or even pretend to paint the walls in your mind. There are two important concepts in Aristotle's resemblance theory of representation: 1. The different parts of the representation correspond to parts of the real object, 2. The structure of relations among the corresponding parts is the same in both. The layout of the house in your mind has the same shape and dimensions as that of the real house [Haugeland, p221ff]. The house in your mind does not, of course, take up just

as much physical space. Concepts such as *shape* and *dimensions* are merely numbers which can be applied to thoughts as well as to physical objects.

John Locke said that objects have two sets of properties: primary and secondary. The primary properties are those the object has apart from us and our experience. The secondary properties are those which objects appear to have because of the way our senses function [Haugeland, p23]. Locke believed in the blank slate, but he saw the difficulty in explaining our ability to learn and act upon new knowledge. How does a blank slate *know* how to process sensory data and use it to make plans? Modern studies in Artificial Intelligence have taught us that some of our most effortless skills require immense computations and various assumptions. Locke tries to wiggle his way out by describing what he called “*The Understanding*”, which *looks* at the marks on the blank slate and carries out the recognizing and associating [Pinker, tBS, p34]. This circular argument reeks of homunculus, and begs the question, “How does *The Understanding* understand?”

Thomas Hobbes is considered by some to be the grandfather of AI, though his intention was to better understand human intelligence. He said that thinking involves mathematical symbolic operations, or *ratiocination*. These operations happen in the mind and are not represented by words or symbols. They exist as *phantasms* or *thought-parcels* (the original *mentalese*). Inspired by Galileo, Hobbes believed that all of reality must be governed by mathematical laws, nothing more than tiny moving particles. He reasoned that the brain was made of those same tiny particles. The senses assign qualities to objects such as taste, color, or pitch, but Hobbes knew that these were not qualities of the objects themselves. If an object appears blue, it is because the senses describe it as such by creating a tiny phantasm with a mental symbol for blue [Haugeland, p23ff]. Nothing in this universe is inherently blue, or ugly, or sour. These sensory categories are meaningless apart from the system they inhabit. We successfully communicate these ideas only because of the relative homogeneity of human sensory systems.

In a way that is similar to these sensory symbols, Ludwig Wittgenstein believed that words have no meaning apart from those who understand them. He takes the point even further to say that unless two or more people are using a word to communicate, it is meaningless. If spoken symbols weren't already hard enough to pin down, he discussed

how they can morph under ambiguity. He gives an example in which a subject is asked to paint one patch of a canvas red and another patch green. The subject uses ideal standard green and red paint for the task. Next the subject is shown a pile of leaves. Some of them are a reddish-brown and others are a greenish yellow. He is asked to separate the pile into red leaves and green leaves. The subject divides the pile by considering the reddish-brown leaves as red, and the greenish-yellow leaves as yellow. A normal subject would never start the first task by painting a reddish-brown patch on the canvas [Wittgenstein, p137]. Why is reddish-brown now considered *red*? If the pile was of reddish-brown and bright red leaves would the result have been different? The meaning of the word *red* did not change between task 1 and 2. It was merely blurred. When presented with objects that did not fit into the known buckets, the subject made them fit by making his own understanding more fuzzy. The same thing happened when Europeans first arrived in South America and discovered the *ananas*. The Castellano word for it is *piña* (Castellano for pinecone). The prevailing belief at the time was that their knowledge of science was complete - therefore all of Botany was already known. When the pineapple (as it became known in England) didn't fit nicely with the content of their books they merely blurred their definition of *pinecone* so that it applied to this new fruit [westegg.com]. Argentinians and many others still call them *ananas*.

Another word, like *bank*, for example, has multiple definitions that do not require blurring. It can refer to a place to store your hard-earned cash, or the edge of a river. These two uses refer to concepts that are entirely unrelated. The meaning of a word can therefore be modified in two ways: its definition can be blurred so that it refers to a larger set of objects, or a mapping can be added to a new concept without blurring the original definition [Wittgenstein, p137].

In the 1940's C. E. Shannon contributed to a field now known as Information Theory. *IT* was primarily concerned with communication of information. This information has to be represented, which of course is at the heart of this discussion. Shannon avoided this problem by saying that although messages are correlated with certain physical or conceptual entities, these semantic aspects are irrelevant to the engineering problem [Shannon]. Perhaps so, but they are relevant to KR. Understanding

how much computation is required for a mental process under a given theory can assist researchers by invalidating inaccurate theories. For example, to compare the relative size of two numbers, the brain must first recognize them, translate them to their internal representations in order to compare them. The answer must then be communicated through an output channel. Our ability to comprehend the mind's inputs, outputs, and internal calculations are at stake. For a theory to be matched to data, the theory must be analyzed under *IT*. Such analysis would provide a guide. In one study, human subjects were asked to compare two numbers. The closer those two numbers were, more time was needed to respond, with a peak on either side of the number itself. Relatively distant numbers were rapidly and accurately compared [Dehaene, p73ff]. Information Theory does not negate the theory that humans compare large numbers on an inexact number line. *IT* supports this idea. *IT* does negate the theory that we compare numbers as computers typically do. The time required for a computer to compare two numbers does not vary inversely to the distance between the two. Unless something unexpected and inaccessible is occurring in our brains, the present theory on numerical comparison bears close examination under *IT*. This theory supplies a viable model for explaining the *Distance Effect*. The symbols our mind uses to compare two numbers is not a set of exact registers which can be lined up. Approximations that are themselves approximately the same require a closer look, or a higher resolution analysis. It all comes down to a greater processing time requirement. The mind is a limited computational engine, and its limits give clues about its internal organization, and what it *knows* ahead of time [Hoffman, p13ff].

Two important classifications to make are those of implemented knowledge and accessible knowledge. They are not mutually exclusive. To speak any language a person must learn a highly complex set of rules, exceptions, and words. One may speak perfectly without being able to describe a single rule that they follow. These rules are implemented knowledge. If the person does not know what they are, then they are inaccessible. Once a linguist comes along and discovers these rules, they do not cease to be considered implemented. They merely pass into the realm of accessible knowledge. Most of what common people consider *knowledge*, is accessible knowledge - things we can describe or see or feel. What is the capital of Vermont? When is my sister's

birthday? etc. Implemented knowledge, on the other hand, is often overlooked because so much of it is inaccessible. How do I tell where one spoken word ends and the next one begins? Why do I think that there's sadness behind that smile? or even, How do I make my heart beat? Somehow the brain *knows* how to maintain proper activity throughout our autonomic systems. Some implemented knowledge will always be inaccessible, and a great deal of it is congenital. Hume, Lord Herbert, Leibniz, and Chomsky all subscribed to a belief in innate knowledge [Chomsky, p263].

As with other aspects of human intelligence, we have learned a great deal about knowledge by attempting to synthesize it. Taking the Hobbesian point of view, an atom of knowledge can be represented in a computer simply by marking some internal registers. In what respect does that internal state represent the knowledge atom? Can the computer be said to *know* the atom in question? If so, does it still know the atom of knowledge when that program is not running? What if the program has been uninstalled? Clearly a prerequisite for knowledge is retrievability. The retrieval process could also be described as translation or conversion. The internal symbol is converted by the program that *wrote* it, to another format which can be understood by another agent. The converted form is then output through some communication channel to the other agent. The program in this case plays the role of Locke's *Understanding*. For example, a program may store a set of numerical quantities corresponding to each field of a spreadsheet. When that sheet is displayed on a monitor you can see what those quantities are. The channel of communication is comprised of a monitor, some photons, and your visual input systems. You are the other agent.

Computers and digital systems are more than an analogy for knowledge and thought. There's no way to "wiggle out" of explaining the mechanics of thought. Developing such systems has been an exploratory journey for modern cognitive psychology and a surprise for AI researchers. The tasks we find easiest as humans appear to be the hardest ones of all in AI. Many programmers believed that problems such as computer vision and speech recognition would be relatively solved by the early 1990's. Technology keeps improving, but it still seems just as far from competing with what we each of us does every day with ease.

All agents of knowledge acquisition, retrieval, and computation share similarities, be they organic or synthetic. Some of these become salient in a slightly more detailed description of a typical computer process. Most computer processes are managed by an Operating System. The OS allows a program to be started and grants it access to memory (RAM). This memory is addressable. When a program runs it can request memory access from the OS, which can grant blocks of it to the program while it is running. If that program stores a datum for future reference it puts the datum somewhere within the memory block granted by the OS. It goes without saying that before the datum is stored, it is temporarily stored within the *state* of the program. Once done, the program needs only three things to retrieve the datum, the memory address where the datum starts, some way to know where it ends, and parsing instructions (*Locke's Understanding*). If the program ends, a modern OS will do "garbage cleanup" - the OS takes back the allocated memory block. This does not mean that these locations in memory will be made 0 (Memory is binary - either 0 or 1), but merely that this particular block will now be considered free to be allocated dynamically to other programs. Once the program stops running, the datum which was stored in that block is now irretrievable. Since most programs will use memory in proprietary ways, there is little chance of another program understanding the datum that was left behind (although expert hackers sometimes exploit a situation just like this one). Without doing any erasing, the datum is now functionally gone - forgotten. The computer can no longer be said to *know* the datum. A more stable storage medium for data is the hard drive. Most programs in this case use a non-proprietary method of data storage: the *file*. Of course there are a relatively infinite number of file types out there, but all programs on a specific computer typically adhere to a non-proprietary *file-system*. This way, a program can store information in a non-volatile manner. Also, the program does not need to record the memory address of each file - this is handled by the disk itself. The program only needs to indicate file name and hierarchical location.

If, for example, an internet request is received by a web server, the web server program (which must be running), understands the request because it was listening for a specific communication protocol (language) on a specific port (channel). In simple impulse-response, the program takes the name of the requested object (which was parsed

from the request) and attempts to retrieve it from the hard disk. Most web pages are stored on the hard disk of a server as flat text files. If the page is in the repository, the program then hands the object to the output channel as a series of binary digits (much detail has been omitted for the reader's sake). Upon receipt, the requestor must still make sense of the data. The program used to make the request will parse the received data and print it to the screen. This is what a browser does.

The explosion of connected devices since 2000 has spurred innovative ways to deal with knowledge representation. The concept of middleware became very important for connecting disparate endpoints. The first connected handheld computers had small black and white screens and painfully slow, not to mention undependable, data connections. How could one use such a device to read a large email attachment? or a large web page with animations or high-resolution images? A middle ground was necessary: *middleware*. The handheld could send an internet request to a middleware server, which would act at first like a browser by retrieving the object of the request from a standard web server, but instead of displaying it on a local screen, it would translate the page into a format and protocol appropriate for the handheld in question and send it there. The handheld would then be able to display the smaller digested version of the web page on its limited display.

Information Theory analysis of this method of communication highlights two weaknesses: 1. Each web page downloaded by this method must be parsed twice. 2. For n different endpoint formats/protocols, $n!$ translation faculties must be developed. Even if the middleware server had a requested page already stored in its *cache* (a local repository of previously retrieved webpages), it would still need to parse the page again (commonly written in HTML and JavaScript) and translate it into the proper format for the requestor (which presumably cannot render HTML). Between 1998 and 2002 there were many significantly different formats and protocols developed for the new connected devices, and their use was not restricted to the devices themselves. Many were setting up servers that were able to serve pages in some of the new formats. The problem was that these formats were being updated constantly, and new ones were introduced so often that

it was futile to try to make a translation faculty for each possible combination of server format and requestor format.

The brain does not suffer from the same burden of computation required in the method above. When sensory input data is parsed, it does not need to be reparsed when passed to a given mind function. For instance, when the phrase "*He threw the ball up in the air,*" is cognized by a hearer, the spatial coordination center of the hearer's brain does not have to reparse the phrase in order to know where the ball went. Nor does the hearer have to reparse it to know that the ball will eventually come back down. There seems to be some universally understood method of knowledge representation. This *mentalese*, as Pinker puts it, is not always precise, but it leaves nothing to the imagination because it *is* the imagination [Pinker, HtMW, p70].

The weaknesses of the above-described middleware would be ameliorated by a universal way to represent web content. Knowledge representation is one thing, but making knowledge accessible to disparate systems is quite another. One proposed solution is *Empirical Markup Structure* (EMS) [Morehead]. The first time a web page is read by the middleware server it is parsed. This parsed form is then stored as a parsed structure on the hard disk, instead of as a flat file. This would be analogous to diagramming a sentence and storing the diagram instead of the sentence. Each subsequent time that the object is requested it can be loaded in its parsed form and immediately translated into the desired output format. One universal method of internal representation means that only $n + 1$ translation faculties will be required instead of $n!$.

Another strength of EMS is its facile ability to manage advanced transformations. The handheld computers in question had very limited capabilities - so limited that most web pages could not be sent in their entirety, even after translation. Splitting web pages without losing their meaning is not an easy task. Not splitting in the middle of sentences, keeping links with their context, and a multitude of other considerations can complicate the problem of dividing web content into logical bite-sized pieces. Inherent in EMS is a tree structure that maintains all parsed elements (tags, text, images, code,) in a particular branch based on their respective function and relationship to all the other elements of that page. Less computation is required to separate an EMS tree into logical chunks that would be necessary to accomplish the same for a flat HTML file. The goal of

this technology was to create a flexible universal way to represent all the knowledge (textual and visual) contained within any given web object, and make that object accessible to any connected endpoint [Morehead].

Quantum Mechanics offers a distinct perspective on knowledge representation. To describe a particle one needs only to know the quantum state. Once this state is known, one knows absolutely everything there is to know about that particle [Cohen-Tannoudji, p108], and that particle is identical in every way to another particle with the same state - a rose is a rose is a rose.

Young's 2-slit experiment shows just how important 'state knowledge' is in the physical universe. In the famous experiment, Young was emitting electrons at a detector through a shield with two slits. He observed strange interference that he could not explain. It did not make sense because he couldn't imagine what paths the electrons could take to create the interference. He had to find out through which slot each electron was passing. Then he would be able to create a mathematical model. His attempt involved testing for each electron as it passed through a slit. The interference disappeared. His and every experiment since then has shown that just the *knowledge* of the state or location of a particle will change the outcome of the experiment [Cohen-Tannoudji, p11ff]. Similar experiments have proven that if two identical particles have the same quantum state, they are fundamentally indistinguishable. Such things are not a matter of having the right equipment or better scientific methods, they are fundamental to the universe. This area of research is uncovering many challenges for a future in which we use elementary particles for computer memory storage. We are already developing a platform for cryptography that may prove unbreakable because of the very laws of physics. Quantum effects will be used to ensure the disappearance of any information encoded within the states of transmitted photons and electrons unless the data stream is read by the approved recipient. No amount of computation could ever decode such a message because the eavesdropped data would be perfectly random.

True randomness is hard to create. Our actions have patterns. Machines operate according to patterns. The entities in nature that do create true randomness depend on processes that are wholly unpredictable, such as nuclear K-capture (a type of nuclear

decay). Identical binary random lists can be created using K-capture data and then used as “one-time pads” for coded communication. The encoded message is run through a binary operation with the next chunk of the one-time pad. The encoded message is therefore also perfectly random. A perfectly random message by definition knows nothing. All things being equal, an attempted code breaker would have an equal chance of transforming the message into any possible string of characters. The original message is only obtainable by reversing the binary operation with an identical chunk of an identical one-time pad. Any other attempt is futile because the original message is inaccessible - fundamentally. Knowledge is present only when there is data and a parser/translator. They are both inseparable from KR.

The internet is the largest repository of knowledge that the world has ever known (using the word *knowledge* very loosely). Attempts to organize and categorize this behemoth have been thwarted by sheer scale and by another problem that humans find simple: semantic comprehension. Useful categorization requires semantic parsing. How can a computer find the meaning of a block of text? How can a computer be made to know the difference between *bank* (the side of a river), and *bank* (the place where money is stored)? Humans could theoretically do it, but it would take too long - the web changes too quickly. The best technology presently available searches for content based on a set of words. This kind of search is merely textual matching - there is no consideration of semantics.

A major hurdle to automatic semantic comprehension is the innate impreciseness of language. Many words have multiple meanings, and some take otherwise unrelated meanings when combined with others words. Words represent concepts, not reality. These concepts do not themselves represent reality either. They are only an approximation, and words are an approximation to concepts. In a Gödelian sort of way, one has to wonder if the nature of KR really can be discussed in a form as limited as written language.

So much of how we understand words has to do with the world we know and experience. Voltaire said, “What is an idea? It is an image that paints itself in my brain.” Much information that is necessary for semantic comprehension is not contained within

the text. How does one comprehend that “*walked into a store*” means shopping, while “*walked into a wall*” means a painful experience with a possible broken nose. If you needed some milk and bread, or you wanted to offer someone a Band-Aid, a computer could not tell you which person you need to talk to. What generalized rule-system could correctly discern both “*turned into a side-street*” and “*turned into a pumpkin*”? In our minds we simulate the action described or visualize the subject of a phrase in order to comprehend it. Our *mentalese* comprises at least partly our faculties for visualization and simulations. Knowledge of the world as we have perceived it thus far is instrumental in making sense of linguistic phrases that do not carry enough information in themselves for complete comprehension [Barsalou]. We use “mirror neurons” to run simulations and fill in the necessary context [Tettamanti, et al. ; Rizzolatti and Craighero]. How could a computer be made to simulate or visualize everything that a human can? For now, we’ll have to cut corners.

One attempt at a better categorization of the web is being called “*The Semantic Web*.” The SW would comprise the standard internet with the inclusion of many semantic notations that corresponding to preexisting web objects. Its construction requires automatic semantic annotation: the computerized attribution of semantic information to every link. Spider programs will scan the web and attempt to glean and digest context words associated with each web object. These words will be stored separately from the web objects themselves but will be retrievable with a new browser made just for the Semantic Web. Such data would also be available for the development of a new search engine which would consider this extra data corresponding to each web object as that object’s semantic context [Kiryakov, et al.].

A smaller scale project in Denmark has as its goal to “develop theories and methods for content-based information retrieval by devising a description language whose expressive power covers not only the purely linguistic analysis of texts, but also the description of a concept ontology and the analysis of queries” [www.ontoquery.dk]. Concepts are divided into different nested levels of ontologies, also called “fuzzy nested aggregation-based descriptions” [Andreason and Nilsson]. Much human effort will be

required to construct a prototype system, but it would be able to give meaningful answers to semantic requests.

Another interesting project that is available over the internet allows users to navigate a semantic network. It is called the “Visual Thesaurus.” A user starts by searching for a word. The word appears on the screen in a box with lines of association connecting it to other words that may be similar, may describe it, or may be described by the word. The user can progress down any line of association by clicking on the secondary word or on the line itself. The Java-based interface is easy to learn and educational. [www.visualthesaurus.com].

Modern methods of KR in Computer Science utilize known linguistic ideologies even where natural language is not concerned. They originate in the conceptual analysis of language and they have equivalent ability to express concepts [Hartley and Barnden]. The Pierce product, for instance, can be used to describe the relationship between two objects:

$$x [\mathit{role} : y]$$

where x and y are conceptual entities, and role is one of the following:

TMP	(temporal)	
LOC	(locative)	
WRT	(with respect to)	
CHR	(characteristic)	
FOR	(purpose)	
CBY	(caused by)	
CMP	(comprising)	[Andreason and Nilsson]

For instance, the idea that the Ebola virus causes death could be captured as following:

$$\text{death} [\mathit{CBY} : \text{Ebola} [\mathit{CHR} : \text{virus}]]$$

The idea of wearing layers of clothing to stay warm could be:

layering [*WRT* : clothing [*FOR* : retention [*WRT* : heat]]]

To create simple phrases like these will at times be more comfortable if inverses for the roles are defined, but there is no need for this inherent to the logic itself.

Nouns must be categorized into *general* and *named entities*. Named entities are individual instantiations of an entity. The noun, “building,” is general, but “The Empire State Building” is a named entity. There is only one Empire State Building, and it is associated with other named entities such as “New York” (the city), and “New York” (the state). Named entities could be organized in a similar way to general nouns, for example:

Empire State Building [*LOC* : NYC [*LOC* : NY State]]]

but they would still inhabit their own semantic network, apart from the general nouns. Named entities are extremely important to search engines.

By necessity, this paper is written with words. We are able to connect each word to an appropriate concept and can therefore use them as building blocks to construct a meaning, but computers cannot yet fill in the context. I propose an infinite-dimensional space where each imaginable empirical concept could be represented by its coordinates therein. Let this be called *concept space* (CS). More than one word may refer to the same point in CS, and multiple points in CS may be invoked by the same word, but each point has by definition only one meaning. Instead of words, CS coordinates would be the semantic “atoms” in an artificially intelligent system.

CS itself would be language-independent, but to write down the coordinates of a given point, one would have to choose a frame of reference, a *base*. Mathematically, a base is a coordinate system. Coordinates from two separate but parallel coordinate systems (let’s say A and B) can mark out the same space (e.g. Cartesian and Polar). To convert a location written in coordinate system A to its coordinates with respect to B, one

simple matrix operator is needed. To ensure conversion, both A and B must have a 1 to 1 mapping - every point in space addressable by A must be addressable by B. In reality, however, there is no language in the world which can by literal means give coordinates to every point imaginable within CS. And no two languages share an exact 1 to 1 mapping. Human translators hear the first language which they do their best to map onto their own internal CS. From that state of cognition the translator merely engages his or her speech production module in the target language. Each translated concept has three forms: 1. source language, 2. translator's cognitive state, 3. target language. Machine based translation could be accomplished in the same manner once a CS were properly populated and charted with at least two coordinate systems.

Another way to think of CS coordinates is that each coordinate can be expressed as a parameter and a value. If each value is termed "level of excitement" and each parameter corresponds to a specific brain substrate, perhaps such a framework might be a model for human thought. If one began with basic but specific concepts, a unique pattern of brain substrate and respective excitement might be found when each concept is beheld by a subject. Two concepts might possibly share identical substrate, but their excitement pattern could not be identical for this theory to work. Such a framework would be consistent with the mathematics of Lotfi Zadeh and presumably human thought as well [McNeill and Freiburger].

Poetry is an under appreciated area of natural language research in the world of cognitive linguistics. Poetic word usage speaks to us in a way that literal words cannot. We all feel it but most find it ineffable. Poets are exactly those who can speak the ineffable. Aristotle set out to study poetry almost as a scientific subject. He said, "For by deviating in exceptional cases from the normal idiom, the language will gain distinction; while, at the same time, the partial conformity with usage will give perspicuity" [Aristotle, part XXII]. Take, for example, this stanza written by Sylvia Plath after receiving some flowers during her stay in a mental hospital:

The tulips are too red in the first place, they hurt me.
Even through the gift paper I could hear them breathe

Lightly, through their white swaddlings, like an awful baby.

Their redness talks to my wound, it corresponds. [Plath, p19]

The reader knows better than to visualize a bouquet in the act of respiration. By incongruous context, the phrases are immediately considered *poetic*, or, non-literal. Since the reader knows not to map these words onto their literal coordinates in CS, a new uncharted mapping is made, found, discovered by the reader alone. This is why poetry is a personal experience. It is different for everyone, and different every time a person goes back to the same poem, because time changes each one of us. Perhaps poetry is the only linguistic expression that can directly touch Aristotelian *quaila*.

Why do we carve up reality into cognizable chunks and classifications? There is no other way for us to interact with this world. Even if our words do not refer to reality, they are still the most useful tool we have. Many great philosophers have tried to figure out if there is a reality outside our minds, or if there is, how closely does it represent our language, our mathematics, and even our logic. Do the words written on this page exist outside our minds? or would Magritte say, “Ceci n’est pas une page”? Is the universe “written” in mathematics as believed by Galileo [Dehaene, p252], or is that mathematics is the only language through which we can understand it?

Our language may restrict us with Whorfian blinders at times, but it also gives us a leap beyond any other species. Knowledge need not be rediscovered by each successive generation. It can be passed on and shared for mutual benefit. Mankind has learned a great deal so far (or so it seems to us) and our rate of learning is speeding up with no end in sight.

REFERENCES

Andreason and Nilsson (2003) "Grammatical Specification of Domain Ontologies", *Data and Knowledge Engineering* 48, 221-230

Aristotle, "Poetics"

Barsalou, Lawrence W. (1999) "Perceptual Symbol Systems", *Behavioral and Brain Sciences* 22, 577-660

Chomsky, Noam (1986) "Knowledge of Language", Praeger Publishers

Conrad, Joseph, (1996) "Under Western Eyes", Penguin Books Ltd.

Cohen-Tannoudji, et al. (1977) "Quantum Mechanics" vol1, Hermann and John Wiley & Sons

Dehaene, Stanislas (1997) "The Number Sense", Oxford University Press

Fodor, Jerry (1983) "The Modularity of Mind", MIT Press

Hartley and Barnden "Semantic Networks, Visualizations of Knowledge,
(<http://www.cs.nmsu.edu/~rth/publications/TICS.pdf>)

Haugeland, John (1990) "Artificial Intelligence, The Very Idea", The MIT Press

Hoffman, Donald D. (1998) "Visual Intelligence", W W Norton and Company

Jackendoff, Ray (1994) "Patterns in the mind", BasicBooks

Kiryakov, Atanas, et al. "Semantic Annotation, Indexing, and Retrieval",
(http://www.ontotext.com/publications/SemAIR_ISWC169.pdf)

Mcneill and Freiberger, (1993) "Fuzzy Logic", Simon and Schuster

Morehead, Graham (2001) "Empirical Markup Structure" (provisional patent),
(<http://www.pangeon.com/cv/EMS.pdf>)

www.ontoquery.dk

Pinker, Steven (1997) "How the Mind Works", W W Norton & Company

Pinker, Steven (2002) "The Blank Slate", Viking Press

Plath, Sylvia (2004), "Ariel", Harper Collins

Plato, "The Republic" (1980), Oxford University Press

Shannon, C. E. (1948) "A Mathematical Theory of Communication", The Bell System Technical Journal, Vol 27, 1948

Tettamanti, et al. (2001) "Listening to action-related sentences activates fronto-parietal motor circuits":

(http://www.brain.hr/Mind&Brain4/NEW_COURSE_MATERIALS/BUCCINO/Listening%20to%20action-related%20sentences%20activates%20fronto-parietal%20motor%20circu.pdf)

Rizzolatti and Craighero (2004) "The Mirror Neuron System" Annu. Rev. Neurosci. 27:169-92

Westegg, (<http://www.westegg.com/etymology/index.html>)

Wittgenstein, Ludwig (1960) "The Blue and Brown Books", Harper & Row