Title: A Proposed Mathematical Theory Explaining Word Order Typology
Author: Asa M. Stepak

e-mail:amstepak@lycos.com

Abstract
   In this paper I attempt to lay the groundwork for an algorithm that measures sentence competency.
   Heretofore, competency of sentences was determined by interviewing speakers of the language. The data compiled forms the basis for grammatical rules that establish the generative grammar of a language. However, the generative grammar, once established, does not filter out all incompetent sentences. Chomsky has noted that there are many sentences that are grammatical but do not satisfy the notion of competency and, similarly, many non-grammatical constructions that do.
   I propose that generative grammar constructions as well as formal theory frameworks such as Transformational Grammar, Minimalist Theory, and Government and Binding do not represent the most irreducible component of a language that determines sentence competency. I propose a Mathematical Theory governing word order typology that explains not only the established generative grammar rules of a language but, also, lays the groundwork for understanding sentence competency in terms of irreducible components that has not been accounted for in previous formal theories. I have done so by relying on a mathematical analysis of word frequency relationships based upon large, representative corpuses that represents a more basic component of sentence construction overlooked by current text processing and artificial intelligence parsing systems and unaccounted for by the generative grammar rules of a language.

1. Introduction
        In this paper I propose a Mathematical Theory and mathematical treatment of linguistic phenomenon. I refer to this theory as mathematical since it is derived from Information Theory which is a Mathematical Theory. But since the theory I propose concerns linguistic phenomenon and can only be validated or refuted based upon observable linguistic phenomenon, the theory might be better referred to as a Physical Linguistic Theory. Also, many of my suppositions and premises which I propose as a basis for the Theory are linguistic or cognitive in nature rather than mathematical. However, before I proceed to detail the actual Theory and how I have chosen to implement the Theory, I think it would be apropos if I first provide some of the broader theoretical background and differing theoretical perspectives. And, of course, since I will be proposing a new theoretical perspective, much of my effort will be directed to advancing my own theoretical perspective.
   In this section I will describe three perspectives of Natural Language Acquisition, (NLA), the first being the Chomskyan perspective based upon the notion of 'Poverty of the Stimulus', the second a Learning Theory perspective, and a third perspective which is the one I propose based upon Information Theory principles. The Chomskyan and Learning Theory perspectives can be regarded as non-contradictory and consistent with one another. However, the Information Theory perspective entails an entirely different framework and does contradict the Chomskyan and Learning Theory perspectives. Based upon Information Theory principles it is my contention that the Chomskyan and Learning Theory proposals are inadequate in many respects.
   Chomsky's explanation of Natural Language Acquisition is based on the notion of 'Poverty of the Stimulus'. From this perspective, since sentences can acquire infinite complex characteristics, the child needs to learn a set of rules governing the language rather than a set of patterns matching the language. The child's learning of these rules is guided internally since there are not sufficient data provided the child for language acquisition to occur through an inductive reasoning. Chomsky proposes we are born with an innate set of rules described as a 'Language Acquisition Device', LAD, that limits the choices we have to make in determining the rules of a particular language. The set of internalized rules we are born with comprise the Universal Grammar, UG,--common limiting constraints found in all languages. The innate Universal Grammar is, also, referred to as the 'Principles and Parameters Theory' which states we are born with innate principles governing our use of language that for each language can be varied within finite parameter ranges. However, the shortcoming of the Chomsky perspective is that there is no attempt to provide an internal independent validity for the innate principles or universal grammar other than what is

observed externally in all languages. Furthermore, Chomsky makes no effort to interconnect the many different features and aspects of the Universal Grammar by a common overriding principle. Thus, many of the principles comprising the Universal Grammar often appear disconnected and incidental, lacking a common cognitive or neuro-physiological justification. In fact, Chomsky points to many peculiarities and anomalies in language that seem to defy even a grammatical explanation and, therefore, presupposes language acquisition requires an innate LAD that cannot be fully described[1]. Chomsky, therefore, portrays the LAD as a 'Black Box' comprising of unknown underlying principles and serving as a common repository for all the known universal features of language. Therefore, if, in fact, there are universal features in language that have not yet been identified, these features would not be included in our description of the principle constraints of the LAD model. The LAD mirrors what is known of the universality of language through linguistic observation, what is not known or yet to be discovered is unaccounted for by the LAD model. Furthermore, the proposal of the LAD model excludes any factors that might be known about underlying cognitive neuro-physiological implementation mechanisms. The LAD's construction is purely based upon observations of language. The Chomsky LAD is intended to serve as a model for directing neuro-physiologists and cognitive scientists to make discoveries that would be consistent with an underlying implementation of the LAD but, the LAD, in its own right, is regarded inviolable by any counter opposing neuro-physiological findings.

As stated above, Chomsky makes no effort to find an underlying principle that might interconnect and explain why the universal features of language are what they are. And, in fact, in so doing many of the hand picked Universal Grammar constraints that have been collected over the years based upon observation of all languages can be disconnected and unrelated to one another. In fact, Chomsky makes no requirement that they be related or connected in some way. Chomsky makes no effort to take a bottom-up approach that would draw upon principles that would provide independent justification and be able to independently predict the collection of grammar constraints attributed to a LAD and, in this regard, Chomsky treats the LAD as a 'Black Box'. His justification for so doing is based upon his drawing upon a wide array of peculiarities in language that appear to defy any known linguistic explanation but, nevertheless, comprise of an integral part of language that is learned in the natural language acquisition phase. The Chomsky perspective, in this sense, is highly subjective and could be regarded as unproven or within the gamut of conceptual formulations that cannot be proven.

A second approach to explaining Natural Language Acquisition is the Learning Theory or Computation Theory approach. Here, the basic argument is that since regular languages comprise of a subset of all possible languages to be learned and regular languages, unlike the complement of regular languages, can be either finite or infinite, there is no theoretical or practical way to decide through an inductive reasoning that is required for Natural Language Acquisition whether a language being learned was an infinite or finite language. Thus, theoretically, there would be a need for some internalized constraints represented as a Universal Grammar, UG, that would limit the search space. A review of this perspective is provided by Nowak (2002). However, it appears that Nowak's review may have oversimplified the required task in Natural Language Acquisition since Learning Theory makes no requirement that language be constrained in order to contain viable information that can be effectively communicated. From the Learning Theory perspective, a language comprising of as little as one or two words or even ten words would be amongst the multitude of possible languages a subject would have to choose from in NLA. Clearly, such abbreviated languages would not fulfill the information containing constraints required of a language. Similarly, languages comprising of one or two element characters would result in lengthy word strings that would not satisfy efficiency constraints of a language. Such languages also would be excluded from the effectual search space due to efficiency constraints. In NLA, the choice of choosing either a finite or infinite language is not at issue—the issue is choosing only a viable language that efficiently communicates information. Many languages, both finite and infinite, are automatically excluded because they do not fulfill this criteria. The fact that natural language can be described as belonging to the infinite class of languages based upon its infinite recursive properties is really irrelevant to the choices that would have to be made during the natural language learning phase. Thus, the search space would be constrained in its own

---

[1] A classic example of idiosyncratic sentence structure suggested by James Garson would be found in the comparison of the following two sentences, "The dog that I saw has rabies." compare with "The dog that bit me has rabies." In the former, 'that' can be omitted without affecting the competency of the sentence whereas, in the latter, 'that' cannot be omitted. Chomsky claims these idiosyncratic phenomenon defy a linguistic explanation and can only be explained by a LAD serving as a 'Black Box'.

right and the effectual search space would be substantially reduced and finite since many of the possible choices of a language simply do not satisfy the Information Theory criteria of serving as a viable platform for communicating information.

Thus, the Learning Theory approach proves nothing and certainly does not prove the necessity for the existence of an internalized UG or LAD. The Learning and Computation Theory approach overlook an important feature of the search space—that the search space is finite in its own right due to Information Theory constraints that requires a language medium to convey viable information efficiently for a specific information source.

The third approach to Natural Language Learning is the one I have proposed based upon Information Theory principles. In taking an Information Theory approach, it becomes rather meaningless to think in terms of whether the search space is limited in its own right or by an external processing such as a LAD. Based upon Information Theory principles, the search space and the processing of the search space are both constrained and it becomes futile to attempt to make a distinction between the two. Information Theory principles require that an optimal homeostasis or equilibrium be achieved between the search space and the processing of the search space and places constraints on both simultaneously. From this vantage point, the mind as the embodiment of the brain takes on a whole new dimension since the information search space includes both and serves as a viable alternative to the more traditional notion of embodiment in a biological structure.

From an Information Theory perspective there is no need to model a theory based upon an internal versus external dichotomy. Statistical Information Theory accounts for simultaneous events occurring within the organism and without and provides the schema that enables the information search space to be effectively used. The adjustments and modifications that are needed do not parallel the artificial boundary line established by the biological boundary of internal versus external. The objective is to use the information space as effectively as possible. Here the programmer is evolution as those organisms that do utilize the information search space most effectively are more certain to survive. But Information Theory is based purely on a statistical notion and is not impeded or detoured by the biological notion of embodiment. Thus, the contours for an effective information processing are not solely determined by an internalized processor. There is ,also, an externalized refinement of the information to be processed and it becomes difficult to draw the line where externalization ends and internalization begins. They both overlap to large extents and may, in fact, be indistinguishable. In Information Theory terms, we don't think in terms of a LAD or an internalized Universal Grammar. We think in terms of encoders and decoders, channels, entropy of information source, information transmission and efficient coding. And the terminology that is used in Information Theory is ubiquitous, it can be applied to any communication of information or physical commodity and no less can it be applied to language. In language, the language is the code, the pharyngeal, oral cavity, tongue and lips produce the transmission signal, somewhere in the interface between neuronal stimulators and the pharyngeal and oral cavity lies the encoder, the air molecules are the channel, somewhere in the interface between the eardrum and neuronal receptors lies the decoder. The information source is our external and internal environments which embodies us. The nice thing about Information Theory, it is based upon mathematical statistical relationships that are linear. Thus, when applied to linguistics it can provide a format for obtaining simple, and quick solutions to otherwise thought to be complex issues in linguistics that, heretofore, one could only make an attempt to understand heuristically. And the concept of the LAD is precisely one such issue that might be better understood or put to rest based upon an Information Theory approach.

Now, having embarked on an Information Theory approach and having described briefly language in Information Theory terms, perhaps, we can now do some of things that Chomsky has chosen not to do, that is, provide an independent, underlying justification for the observable phenomenon we refer to as Universal Grammar constraints that appear disconnected otherwise. And, in so doing, what has otherwise been referred to as a LAD as a 'Black Box' might otherwise be better referred to as a fundamental underlying principle we have evolved to implement language communication effectively. And we might suppose that such a fundamental principle based upon Information Theory, also, represents a common pathway to our other faculties such as walking, vision, etc. We can still think in terms of a highly specialized area of the brain devoted to language use, perhaps, based upon a Hopfield connectionist model, but, do not have to think of this specialized area relying on mystifying 'Black Box' principles separate and distinct from other cognitive functions. The LAD, in fact, is demystified and the 'Black Box' notion unraveled.

2. An Information Theory Model for Language

In artificial coding there are two coding strategies to consider. 1. To minimize the entropy of the information source as much as possible. 2. To rely on coding that is the most efficient in transmitting a given information source. These two overriding principles can be seen as, also, operating on a natural language coding level.

What is to be encoded in a natural language channel is our cognitive awareness that we choose for any given time duration to communicate rather than contemplate introspectively. Our cognitive awareness is the information source with a characteristic entropy. However, in achieving 1., above, the effective entropy can be reduced if portions of this information source do not have to be searched or their probability of occurrence are reduced to zero. Grammatical coding in language or what is more commonly referred to as syntax accomplishes this objective. The onset of one particular grammatical category in a sentence affects the probability of occurrence of the grammatical category that is to follow. For instance, in the case of the determiner, a noun has an extremely high probability of occurring immediately afterwards and, thus, other probabilities of occurrence of other grammatical categories are reduced to zero or a minimum. Grammatical coding in sentence structure, thus, serves to reduce the overall entropy of the given information source at any given point in time allowing for a more simplistic and efficient encoding. In achieving 2. ,above, an Information Theory Huffman algorithm is utilized. This is evident by the characteristic frequent word lists found in all languages. There are a predetermined small subset of words comprising of most common words in all languages that comprise of the majority of words utilized in the language. For English, roughly the 300 most frequently used words comprise of 65% of the words utilized in the language. Assuming there are roughly 300,000 available lexemes in a language, in effect, .1% of available words comprise of 65% of the words used. Furthermore, the most frequently used words tend to be the simplest encoded words of a language as is evident by their shorter lengths, fewer number of phonetic segments they comprise of and their shorter temporal duration. This observation is consistent with the implementation of a Huffman algorithm that determines the most efficient coding scheme for a specific n-ary information source in artificial coding. The Huffman algorithm assigns to information source items of high probability of occurrence the simplest codes resulting in a lowest overall average word length for a code.

Zelig Harris (1988) was one of the first proponents of information constraint implementation in natural language. His view was that phonemic choices for words were constrained by the underlying requirement that words carry information. Harris illustrated this by the characteristic spikes in permissible phoneme choices evident at word boundaries, thus, verifying that the selection of a phoneme within a word or morpheme was constrained when compared to word boundaries. Similar constraints are revealed when examining a language's most frequently used words or MCW's. Here it is evident that the most frequently used words are, also, the simplest coded words consistent with Huffman but a substantial portion of the available phonetic combinations or permutations are never used for simplistic encoding. Thus, unlike artificial coding that relies on Huffman where all available combinations of simplistic codes are relied upon, there appears to be a constraint imposed on natural language that precludes the use of all phonetic combinations. As was independently suggested by Harris based upon word boundary spiking data, this additional constraint is an information carrying constraint or semantic constraint. Such constraints, also, would have to include an encoding for word boundaries so that in oral communication individual words could be distinguished without pauses that would otherwise markedly slow down oral communication and likely render it ineffectual.

The two strategies described above in making language more code efficient and easier to utilize relative to an information source are, also, the strategies that determine the linear sequencing of grammatical categories in sentence structure. Sentences can be infinitely long and complex or extremely short or fragmental. But in terms of what grammatical categories occur most frequently in a language, here too, the strategy would be to front end the most frequent categories and back end the less frequent categories in a phrase or sentence structure. The same principle applies to constituents of categories, the most frequent being in the head or front and the less frequent or non-frequent being placed towards the tail or end of the phrase. This single unifying principle governs the sequencing of grammatical units in a particular language which, in turn, can be ascertained by the frequency of the grammatical units and constituents thereof appearing in large, representative corpuses of the language. To the extent the frequencies would be different in corpuses of different languages, the sequencing would be different to reflect the differences in frequencies.

Now, what I just described for you above would seem to follow an Information Theory approach as applied to natural language but there is a subtle difference when compared to other artificial, man made forms of communication. In artificial communication, discrete events take place based upon a pre-coding of the information source. In others words, a telephone conversation that has to be conveyed over transmission lines is already pre-structured and coded at the source which is the telephone. Here the objective is to convey the message that has already been coded at the source. In natural language, however, there appears to be an ongoing processing or coding of the source message. In other words, the source is not completely coded or structured but exists in a partial gestation state. This results in codes not being transmitted merely on a predetermined discrete time event schedules, but allows for the most frequent coded words to take precedence in transmission over less frequent coded words to the extent other coding constraints are not violated. The advantage to this approach is obvious, it allows for the sequencing of grammatical units and their constituents based upon frequency precedence that would not be achievable otherwise, thus, promoting more efficient encoding. Thus, there is an added means by which efficiency in coding is achieved in natural language not available in artificial coding. Not only are information source data of high probability of occurrence attributed simple codes, but, also, the high frequency simple codes are front ended in the coding. This exemplifies what I stated in section 1, that the natural language information source and the processing of the same almost become indistinguishable and much of the refinement and processing of the information source occur simultaneously or in tandem. Thus it would be incorrect to describe the natural language phenomenon as determined by a static repository of rules and principles such as a LAD as Chomsky does. Language is a dynamic process that is governed by both internal and external processes that are in equilibrium with one another based upon the statistical properties of Information Theory applied to a representative corpus.

3. A Mathematical Analysis of Most Common Words (MCW)

A simple probability analysis shows that the selection of codes for language is not random but subject to several constraints. The first constraint is that the most frequent codes be simplest in form so to minimize the average word length of the entire coding. A second constraint limiting the first is that the coding be such that it carries meaning and includes coding that distinguishes the boundaries between words so that oral communication can proceed at an effectual rate. The meaning constraint or semantic constraint comes in two types, iconic and non-iconic mental lexicon based. The retrieval of words belonging to the first type is autonomic, not requiring an underlying conceptual recognition of what is being communicated. Mental lexicon based words, on the other hand, do require some sort of on-line conceptual cognition corresponding to the words being used. It is words of the first type that serve as iconic markers in sentences that determine the grammatical competency of sentences since the iconic words are the MCW's of a language that, as described in the previous section, are coded to be frontloaded in phrases based on their frequency precedence to render the overall coding more efficient. The iconicism attributed to MCW's is verified by the substantial constraints placed on the coding that goes far beyond what would be needed for an arbitrary semantic coding of words. Mental lexicon words do not serve as iconic markers for grammatical purposes, (albeit, they may retain a remnant of semantic iconicity) and, therefore, are typically backloaded in phrases.

Using Trager Smith's system of 32 phonetic segments that was used in the A. Hood Roberts study (Roberts, 1965) and adding one null element so we have 33 possible choices, the number of possible combinations comprising of two phonetic segments or less is 33^2 or 1089. Based upon the Roberts study, only 50 words in Horn's 10,000 most frequent word list were found to have only two phonetic segments or less. These 50 words, however, accounted for approximately 33% of the words used in American English and, thus, are in the class of MCW's. Based upon a model that attempts to account for semantic constraints, an alternating pattern of vowel and consonant would be a better approximation of the available permutations for a word containing two phonetic segments. Such a model yields $8 * 25 * 2 = 400$ possible choices. The factor of 2 is used since two-phonetic segment words exist in two canonical forms, VC and CV as was determined in the Roberts study. Thus, the conclusion one must reach is that the 350 additional available permutations that are not used are unsuitable since they do not satisfy semantic constraints or word boundary constraints. But if semantic coding is arbitrary and non-iconic then there should be no additional semantic constraint imposed in choosing additional available two or less phonetic segments permutations other than when it results in a one-phonetic segment word. In one –phonetic segment words, identical vowels appearing in the first or second position would represent the same code and one would have to be eliminated due to redundancy. The same would hold true for consonants. But, here, the most that would have to be eliminated would be eight potential choices for vowels and 24 for consonants. That still

leaves 318 available unused choices. The limiting factor would, thus, have to be considered a word boundary constraint. But it is difficult to imagine that word boundary constraints alone would eliminate the use of the unused 318 available permutations. My conclusion, therefore, is that word boundary constraints serve only as a partial constraint and alone do not account for the 318 additional available permutations not being used. The additional constraint that accounts for additional available permutations not being used is an iconic semantic constraint that requires meaning be nestled in some aspect of articulation that metaphorically or otherwise relates to meaning in our internal or external environments (Stepak 2002). The 318 permutations of two phonetic segments that are not used, other than those that are eliminated due to word boundary constraints, cannot satisfy the requirement of iconicism and are not consistent with the Oral Metaphor Construct, OMC, (Stepak 2002,2003), of the language which serves as a characteristic iconic imprint for a language. Only having utilized 50 two-phonetic segment permutations even though more are available , language coding proceeds to utilize three phonetic segments. This fact alone would seem to contradict a Huffman algorithm hypothesis applied to natural language which utilizes the simplest available coding for the words of highest probability of occurrence. However, a Huffman application to natural language would not be refuted if we consider the parameter of simplicity not to be the character length of the code or the number of phonemes but rather the codes iconicity. In so doing, not choosing the additional 318 two-phonetic segment permutations due to iconic constraints would be consistent with a Huffman algorithm application to natural language.

   We can, also, do a statistical calculation of the available three-phonemic segment combinations and permutations in relation to the number that are actually used based upon the Roberts study and we will reach similar conclusions. Stated simply, there are a high number of three-phonetic segment permutations not utilized that cannot be merely explained by word boundary constraints or non-iconic semantic constraints. Eliminating the null character in this treatment since we want the exact number of three-phonetic segment combination or permutations and, not less than three, the number of three-phonetic combinations is 32^3 which equals 32,768 available combinations. Based upon the Roberts study, only 552 words out of a most frequent word list of 10,000 comprise of three phonemic segments. These three phonetic segment words, nevertheless, had an overall frequency of usage of approximately 26.5 %. They are, thus, included in the class of MCW's. Grouping semi-vowels with consonants, the canonical forms of these words based upon the Roberts study are CVC, or VCC. However, the frequency of the VCC canonical form is so low it can be ignored in this mathematical treatment. Therefore, based upon the predominating canonical form of CVC we can approximate the number of permutations to be $24 * 8 * 24 =$ 4608. Here again, the number of available permutations is exceedingly much higher than the 520 actual number of three-phonetic segment words that are used based upon the CVC canonical form. So great a disparity can only be explained by a substantial constraint that can only originate from a semantic iconic constraint where semantic coding is not arbitrary, but, intrinsically derived by some aspect of articulation that conveys semantic significance.

    Before ending this section and moving on to the next, I should state that in the above treatment I did not make reference to bound MCW's which the Robert's study did not consider. It is my view that common inflections and affixes of words, based upon there frequency of occurrence, should be considered with the class of MCW's. A basic tenet of my work on grammatical sequencing is that bound MCW's share the same features and serve the same functions as non-bound MCW's. Both serve as iconic markers in sentence structure. However, in American English, the bound MCW's comprise of a relatively small group, perhaps, no more than 30 such bound MCW's are present and the majority fall into the three-phonetic segment class. Thus, the mathematical conclusions reached in this section still would remain valid even having not considered the bound MCW's in the mathematical treatment.

4. From a Cognitive Perspective

   A term often used in describing phenomenon from a cognitive perspective is the term 'object'. However, in my review of the literature on this subject I have failed to find anyone attempting to provide an exact definition of the term 'object'. To eliminate the plausibility of this term being misconstrued by those not having exposure to the cognitive sciences, I propose the following definition. **'***An object represents a foci of our attention that can be described as a unit distinguishable from its immediate environment***'**. In so far as discrete objects are the means by which we gain recognition of our environment and interact with the environment, our perception of objects in a non-language sense is, also, likely to be governed by Information Theory principles. In other words, we perceive and interact with objects in a discrete time event sense even though the true physical reality of our environments might be better portrayed as

continuous, and non-discrete. Thus, herein lies the connection between what we perceive cognitively and that which we communicate linguistically since language use, likewise, is best described as a discrete time event phenomenon.

Objects can be in the form of words in which case we refer to them as word objects or they can be entities we interact with in a non-language sense in which case they would be referred to as cognitive objects. The number of objects we can possibly conceive of and routinely perceive is a very high number but it is not infinite. There are objects that we have not previously perceived that could be brought to our attention through instructive enlightenment. It could be said that enlightenment is a process by which the number of objects we routinely perceive is modified in some way. Objects can be reduced or subdivided and considered as new objects or objects of different classes might be grouped together as distinctly new objects. Objects can be tangible or conceptual, they can be in our sensory fields or beyond our sensory fields.

The point to be made is that our conception of cognitive objects to a large extent dictates our information processing and representation of those objects in our language medium. Perception of cognitive objects categories determines the relative importance of corresponding word object categories and their frequencies in a language and their ultimate grammatical sequencing. To the extent that cognitive perceptions would tend to differ amongst different populations, so too would their corresponding languages. It is a basic tenet of this paper that a mathematical analysis of large representative corpuses of a language in terms of the relative sizes of word categories is a good measure of the relative size of classes of cognitive objects in the cognitive field of the users of the language. The point to be made is there are many perspectives and vantages points from which to view a cognitive field and these differences ultimately are reflected in different frequency precedent values for corresponding word categories in different languages that results in a unique sequencing of grammatical units in different languages.

However, I do not advocate a pure cognitive perspective since it is my view that what I have described above is ultimately constrained by Information Theory considerations. In other words, it may be useful to attempt to construct cognitive schema to gain a semantic understanding of intra-sentence word dependencies in terms of subject, object etc. but, once having been established, grammatical sequencing becomes primarily an Information Theory phenomenon. The cognitive side of grammatical sequencing resides only in the cognitive discrimination of iconic markers and their proper precedence in a sentence structure based upon frequency values, which, in turn, is determined by the application of Information Theory principles to the cognitive field of the users of the language.

5. The Proposed Theory

In the previous sections I have attempted to provide the reader some of the broader theoretical background so I would not have to devote as much time and effort explaining the more detailed axioms and premises my proposed theory is based upon. Every theory requires axioms and premises. If the theory is able to predict phenomenon, then it can be assumed the axioms and premises upon which it is based are generally valid. However, that does not preclude the design of experimentation that could potentially test the validity of individual premises upon which a theory is constructed or, for that matter, that premises of a theory might be replaced with better premises. Based upon the limited data I have explored which involve only comparisons with English with some of the other Indo-European languages, my proposed theory can only be considered a working hypothesis at this stage subject to further refinement and supplementation. In this section, I will attempt to describe in a summary fashion all the tenets, premises, and perspectives which I have proposed as relevant to the formulation of my proposed theory and then state the Theory in its current form. I should note that the premises that I list may not represent a complete list of premises that should have been listed. Also, I wish to make the reader mindful of the fact that it is my belief that the actual Theory I propose may require further refinement, supplementation, and, perhaps, recapitulation to some extent as more data is explored from the diverse array of languages and dialects but, still, I firmly believe that my basic working hypothesis is valid. That working hypothesis is based upon a Information Theory perspective which claims that the different sequencing of grammatical units across different languages is determined by a fundamental unifying dynamic principle. In the next section, I will, then, apply my Theory to both simple and complex grammatical constructions and will provide justification for different grammatical constructions prevailing in English and French. Some of the premises detailed below have already been discussed in previous sections of this paper. However, in a brief paper such as this, it is not possible to devote as much detailed discussion and background information as I would have liked to all the premises which follow. Some of this void can be filled by reference to my other works described under

'References', page 12, items 4, 5, and 6. Other portions of this void will require my further indulgence in addressing the topics more specifically through additional writings.

My Proposal:

1. In natural language learning and usage, the subject relies on iconic markers comprising of the set of MCW's of a language that enable the user to distinguish a competent sentence. The iconic markers need to be sequenced in a sentence in a prescribed format based upon Information Theory principles for the sentence to be competent.

2. In natural language acquisition, the subject not only learns the set of words comprising of the iconic markers in a language, but, also, learns the degree of iconicism of the markers. Competence of sentence structure is achieved by the subject's intuitive ability to front end strong iconic markers and backend weak iconic markers in a sentence structure or phrase. In learning the degree of iconicism of iconic markers, the subject learns the relative precedence of iconic markers in a sentence structure or phrase.

3. MCW's are iconic and serve as iconic markers.

4. MCW's are cognated differently than uncommon words. Thus, the subject can distinguish between MCW's and uncommon words on a cognitive level. Iconic or MCW's are autonomic words, and their use does not require a concurrent semantic cognition. Uncommon words or mental lexicon based words, when used, due require to some extent a concurrent semantic cognition.

5. All MCW's are iconic (with some few exceptions such as proper nouns and depending on the corpus). However, the iconicity of MCW's might be masked by a surface structure hardwiring that has evolved over long evolutionary epochs. A word will not become a MCW unless it is iconic.

6. In written communication, MCW's can be represented as being bound to uncommon words. In oral communication, the distinction is less clear. Still, bound MCW's serve the same purpose and function as unbound MCW's in serving as iconic markers. Bound MCW's usually come in the form of common inflections and affixes.

7. The Information Theory based strategy of language is to front end grammatical categories and constituents of high frequency and to back end grammatical categories and constituents of low frequency in a phrase or sentence. The subject accomplishes this objective by relying on iconic markers. Each phrase or grammatical unit has an associated iconic marker from which the frequency value of the constituents of the phrase are determined.

8. Strong stresses on individual syllables and words as in the English language can attribute to a grammatical category iconicism that would not otherwise reside in MCW iconic markers alone. Examples are the adjective, adverb, and abstract noun. Corresponding strong stresses are not apparent in French or Spanish explaining the differing sequencing of grammatical units involving these categories.

Theorem:

   Relying on large representative corpuses of a language, one can establish the relative size of word categories. Furthermore, by relying on the frequency of most common words and bound most common words, one can establish a category frequency of occurrence of an iconic marker. Multiplying the category frequency of occurrence times the size of the category of which the word is a constituent will yield positive frequency values for iconic markers and normalize non-iconic words to a value of 1. The size of the largest word category is arbitrarily assigned the value of K which is a high positive number but not infinite. The sizes of the other smaller categories are designated values proportionate to their relative size to the largest category in terms of K, i.e. 1/2K, 1/3K etc. To determine the overall frequency precedence of grammatical units within a sentence, rather, than merely within a phrase within the sentence, it is necessary to determine the intra-sentence function-argument word dependencies[2] based upon a semantic analysis. Once established, in all instances where words are arguments of other words serving as functions, the upper limit of the class size of the argument is taken to be the class size of the word serving as the function. Determiners , prepositions, general determiners are not considered as having a fixed class size and are assigned the class size of the function to which they serve as argument, however, are considered as having a calculable frequency of occurrence based upon their frequency ranking among MCW's. Being that pronouns can serve as primary functions as do nouns, they are attributed as having the same class size of nouns. The above treatment does not exclude other factors such as phonetic stress that can attribute to a

---

[2] Zelig Harris used word dependencies in his study of language except that he referred to word dependencies as operator-argument dependencies. Harris Z 1988 *Language and information*, New York, Columbia University Press.

non-iconic category constituent a high frequency value. However, the frequency value attributed to stress cannot be independently approximated using a representative corpus and can only be fitted to the grammatical data.

Based upon the methodology described above the following four principles govern the sequencing of grammatical units in a phrase or sentence.

1. The frequency value of linguistic units relative to the frequency values of other linguistic unit determines the linguistic unit's relative precedence in a phrase or sentence. The higher the frequency value, the higher the precedence.

2. It is possible for frequency values to be too high immediately prior to a non-iconic constituent so that the non-iconic constituent is masked due to iconic persistence. This may occur when a non-iconic constituent is preceded by another constituent having two origins of iconicism such as in the adverb, one being stress and the other being the iconic suffix 'ly' .

3. Principle 2 above can serve to constrain principle 1 above. When such constraints occur, a linguistic unit of lower frequency value can proceed one of higher value.

4. Sentence structure can be viewed as comprising of connecting function-argument units or complexes which transduce iconic valence through the sentence—each function-argument complex must carry an iconic valence which one can measure in terms of frequency values.

6. Implementation and Examples

Using the above suggested methodology of the theorem, anyone familiar with corpuses could attempt to do their own mathematical computation of frequency values to justify grammatical sequence. There is still some guess work in approximation involved as to how one best goes about calculating frequency values. And I do not expect anyone's own calculation to match exactly my own. The important thing is that the calculations in some fundamental way explains the sequencing of grammatical units within language. And at this early stage there is still some fitting of the data that might have to be accomplished through refinement of the technique. For instance, there is no independent method for assigning frequency value attributed from phonetic stress. Here the assignment would solely depend on using a value that would explain the grammatical sequencing and hope that the same value works under other scenarios. The methodology still requires further testing across differing languages and dialects utilizing different corpuses. The results that follow were based on a mathematical analysis of the LOB corpuses as computed by Johansson and Hofland (1989).

Based upon the Johansson and Hofland study, my approximation for word class size and category frequency of occurrence (frequency of words relative to their class size) are as follows:

Class Size: Noun Objects = K; Verb Objects = K/10; Adj. Objects = K/3; Adv Objects = K/25

Avg. Frequency of Occurrence: Nouns = 1/K; Verbs = 10/K; Adj = 1/10 from phonetic stress, fitting of the data, Adv.= 1 from 'ly' suffix and phonetic stress, fitting of the data; Prep = 1/40; Det(a,the) = 1/4; Gen. Det. = 1/6; Pronouns = 1/22

I should reiterate, that the above frequency of occurrence values are approximations that serve to fit the data but are, also, reasonably deduced calculations based upon Johansson and Hofland's mathematical analysis of the LOB corpus. By reasonable, I mean that the values fit relative to one another but not necessarily reflect true absolute values. Also, I should indicate that in the following examples where I analyze French constructions, I am assuming the word class sizes determined in the LOB would also apply to the French language. This may not be an entirely correct assumption to make and any analysis of French based upon an English corpus would have to be considered rather limited in scope.

Examples:
1. Det noun
  arg1 f1
  Det arg1: freq. ¼, size K(size of function) = fv K/4
  Noun f1: freq. 1/K, size K = fv 1 (noun is non-iconic)
  Therefore Det precedes Noun. Principle 1
   In French, all nouns require to be fronted by a determiner, unlike English. This difference is due to the difference in amplification of stress in the two languages. In English, stresses carry sufficient amplitude for plural nouns and abstract nouns to stand alone without Determiners. In English, stress attributes to stand

alone nouns iconicism. In French, stress is not of a sufficient amplitude to attribute to nouns iconicism and thus, all French nouns require the determiner. The same holds true for Spanish.

2. Det noun verb adv adj noun. English, attributive adjective.

                    arg1 f1
                arg2 f2
   Noun f1: freq 1/K, size K = fv 1
   Adj arg1, f2 : freq. 1/10, size K/3 = fv K/30 (fv is attributed from stress)
   Adv.arg2: freq. 1, size K/25 = fv K/25 (fv is attributed from stress and suffix "ly".)
   Principle 1 applies.

3. Det noun verb noun adv. English
            f1   arg1 arg1
   Noun arg1: freq 1/K, size K = fv 1 (non-iconic)
   Adv.arg.1: freq 1, size K/25 =fv K/25 (doubly iconic due to stress, results in iconic persistence.)
   Principles 2 and 3 apply.

4. Det noun verb adv noun. French
            f1   arg1 arg1
   Adv.arg1: freq 1/10, size K/25 = fv K/250(No stress, singly iconic in French, 'ment' replaces 'ly')
   Noun arg1: freq 1/K, size K = fv 1 (non-iconic)
   Principle 1 applies.

5. Det noun verb noun adj. French, attributive adjective.
            f1   arg1
              f2   arg2
   Noun: arg1, f2: freq 1/K, size K = fv 1 (non-iconic)
    Adj arg2: freq. K/3, size 3/K = fv 1 (non-iconic due to no stress)   Explanation: Since there is minimal stress in French, noun and adj have virtually equal frequency values and adj can follow noun or vice of versa depending upon the relative frequency values of the two constituents based upon a representative corpus.

    Following are examples of more complex constructions. Here we do not have to rely on approximate values of frequency or class size but, rather, make an assessment based upon general iconic marker principles in explaining why certain constructions are forbidden or favored over others. The following sentence examples, 6 thru 9, are borrowed from Langacker (2002). Langacker claims that the forbidden constructions, below, have a cognitive grammar explanation and that there is a continuum between the lexicon and grammar. However, Langacker's cognitive schema often appear to be excessively subjective and lacking in formal constraints. In Langacker's cognitive grammar approach, the measuring device is the linguist's own mental cognitive judgment which , in essence, is the very thing being measured. When the measuring device is the thing being measured, it becomes difficult to distinguish between the two and to place verifying controls on the measuring device. A good example of this would be the geocentric astronomy theories of the pre-Galileo era. To what extent can it be shown that the cognitive grammar schema proposed by Langacker are not artifacts of the mental cognition utilized in proposing the schema? In the following sentence interpretations, I rely solely upon objective Information Theory principles and frequency value analysis which provide a much more effective and verifiable analysis of idiosyncratic sentence structure than does the cognitive grammar schema advanced by Langacker. Moreover, my approach is, nonetheless, a cognitive approach being fundamentally based upon the cognitive discrimination and response to iconic markers which are to be arranged in sentence structure to maximize information transfer.

 6. a. *I baked her a cake*. b.?*I mowed her the lawn*.
    b., above, is forbidden since lawn as a direct argument of mowed has a high frequency of occurrence. Mowed is almost idiomatic in that it takes only two arguments, lawn or grass. Thus, "I mowed the lawn for her" maximizes frequency values of the sentence and is required. In a. above, baked takes several potential arguments, any food item that can be baked. So cake does not have as a high a frequency of occurrence as a

direct argument of bake as does lawn as a direct argument of mowed and the iconic marker her can be placed in between cake and baked. Now consider.

7.a. *I cleared the floor for Bill.* b.? *I cleared for Bill the floor*.
  a. is an improvement over b. because in a. the definite article is front loaded. Remember that the definite article is a constitutent with a high frequency value. This serves to maximize the overall frequency value of the sentence. Now consider.

8.a. *I sent a walrus to Antarctica.* b.?*I sent Antarctica a walrus*.
Walrus as an argument of sent has a much higher frequency of occurrence then does Antarctica as a argument of sent. In fact, the latter probably represents a corpus frequency value of zero. Thus, a. is the favored construction. Now consider.

9. a. *I gave the fence a new coat of paint*. b. ?*I gave a new coat of paint to the fence*.
Here again we are dealing simply with an idiomatic phrase "a new coat of paint" that has very low frequency of occurrence as an argument of gave. Thus, it is backended in the sentence to allow "the fence" to take precedence that has a much higher frequency value due to the determiner "the".

10. Following is the idiosyncratic sentence structure described in footnote 1, page 2.
  a. *The dog that I saw has rabies*. or *The dog I saw has rabies*.
  b. *The dog that bit me has rabies*. or ?*The dog bit me has rabies*.
 Explanation: "That" is an iconic marker that signals a consecutive action event in a sentence. When the consecutive action has embedded its own iconic marker,"that" can be dropped as in a. above where the following pronoun serves as the second iconic marker signaling the second action. In b. above, there is no second iconic marker to signal the second event and thus "that" cannot be dropped. Now consider this.
 "? *The dog bit me that has rabies*." Here we have two iconic markers for two action events but the sentence is still forbidden. The reason is "that" has a high frequency value as a determiner and, thus, needs to be front loaded in the sentence to maximize the overall frequency value of the sentence.

   In conclusion, frequency value analysis can be used to explain theoretical grammatical syntax constructions or word order typology across languages, examples 1 to 5 above, as well as departures from theoretical grammatical syntax constructions, examples 6-10 above. Moreover, frequency value analysis can be used as a method for measuring the linguistic notion of sentence competency.

7. Summary
    From a frequency value perspective, sentence structure can be viewed as comprising of connecting function-argument units or complexes which transduce iconic valence through the sentence—each function-argument complex carries an iconic valence attributed to an iconic source. The iconic source can either be a MCW or stress. Iconic valence is measured by calculating frequency values. The strategy for achieving sentence competency is to front end constituents of high frequency value in a sentence or phrase and, thus, maximize the transduction of iconicism through the sentence or phrase. You might say that iconic transductance on a cognitive level is similar to electrical conductance through a copper wire. The voltage and current at the input end of the copper wire is always greater than that at the output end due to friction and thermodynamics. There is a linguistic friction that results in progressively lower frequency values as one progresses through the sentence. The linguistic friction can be minimized by front loading the sentence with function-argument complexes of high iconic valence (high frequency value).
    Unlike Chomsky's Lad model which fails to suggest or imply any particular neuro-physiological implementation pathway, the frequency value model is extremely compatible with implementation models from several perspectives. It is consistent with a non-supervised learning connectionist model where most frequent sensation-response associations result in the strongest neuronal associations. In this sense, the frequency value model can be considered as the backside of a non-supervised learning connectionist model which, in turn, could be considered as the backside of Statistical Information Theory that serves as the overriding evolutionary driving force that establishes the neurological framework which mediates Statistical Information Theory constraints. Furthermore, the frequency value model would be, also, consistent with a more traditional cognitive model of short term memory spans. Here, the strategy would be to front load strong iconic function-arguments complexes in a sentence or phrase due to the longer attention

spans associated with strong iconic complexes. On the other hand, weaker or weak iconic complexes would succumb to shorter memory attention spans and, thus, would better serve memory retention by being backloaded in the sentence or phrase.

Thus, the frequency value model appears to be compatible with three implementation models, Statistical Information Theory serving as the basic model, and the connectionist and cognitive models serving as two equivalent models that are evolutionary determined by the Statistical Information Theory model. More importantly, the frequency value model is not a parsing model but a function model compatible with functional programming. Thus, if the parsing model fails or is ambiguous, the functional model can resolve the ambiguity by calculating the maximum frequency value for a sentence or phrase which would represent the correct sentence or phrase to be used.

References

1. Nowak M 2002 Computational and Evolutionary Aspects of Language. Nature Vol 417:611-617

2. Harris Z, 1988 Language and Information. New York, Columbia University Press

3. Roberts A. Hood 1965 *A Statistical Linguistic Analysis of American English.* Netherlands, Mouton and Co.

4. Stepak, A. 2002 *Rationalism versus Empiricism: A New Perspective*. A work in progress. (filed in the copyright office, Washington D.C.)

5. Stepak, A 2002 *Oral Metaphor Construct. In Proceedings SSGRR Summer 2002,* http://cogprints.ecs.soton.ac.uk/archive/00002294/

6. Stepak, A 2003 *Oral Metaphor Construct: An Information Theory and Cognitive Perspective*. A work in progress. (filed in the copyright office, Washington D.C.)

7. Johansson S, Hofland K 1989 *Frequency Analysis of English Vocabulary and Grammar Based Upon the LOB corpus, Vol.1&2*, Oxford, Clarendon Press

8. Langacker R, 2002 *The Cognitive Basis of Grammar*, pp 14-15. Berlin, New York, Mouton de Gruyter