Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 5 (6): 952-964, 2005
ISSN 1812-5654

Frequency Value Grammar and Information Theory

Asa M. Stepak
P.O. Box 540325, Lake Worth, Fl. 33454, USA

Abstract: Previous efforts to calculate the entropy of written English were based upon inadequate forward (one-way) n-gram models that will result in an overestimation of the entropy. These models failed to recognize that in English, generally, function words preceding an open-class word within a phrasal unit is not homomorphic with the forward n-gram analysis since these words represent decompressions of cognitive objects in the information source probability space that tend to be homomorphic with a backward n-gram analysis. A backward n-gram analysis in these instances will result in word frequency values more representative of cognitive object co-occurrences in the 'information source' and a more accurate calculation of the upper bound entropy. The present study demonstrates how Frequency Value Grammar (FVG) can be used as a model for improving the upper bound calculation of entropy of written English implying that FVG is a more valid grammar than generative grammar or Probability Syntax (PS). Generative grammar is passive and descriptive and has no bearing on calculating the entropy of language whereas the theoretical framework of PS can support only a non-directional trigram analysis of language which disregards mutual information and results in a higher upper bound for the entropy of English. In demonstrating that FVG serves to validity an improved upper bound for the entropy of English, entropy has, in fact, also, been used in this study as a formal method for evaluating the validity of a formal grammar. This study demonstrates, therefore, that in addition to entropy being used as a metric for describing Natural language, it also can be used as a formal method for evaluating grammatical approaches to Natural language which further suggests its plausible use as a formal method for evaluating other formal systems such as artificial language and logic systems in general and for complementing denotational semantics and other formal methods. In this study, however, entropy was used only with respect to evaluating Natural language and grammatical descriptions of Natural language.

Key words: Frequency Value Grammar (FVG), information theory, entropy, mutual information, Iconic

INTRODUCTION

The groundwork for Frequency Value Grammar (FVG) is described by Stepak[1]. FVG is a formal syntax theoretically based in large part on information theory principles. FVG relies on dynamic physical principles external to the corpus which shape and mould the corpus whereas generative grammar and other formal syntactic theories are based exclusively on patterns (fractals) found occurring within the well-formed portion of the corpus. However, FVG should not be confused with Probability Syntax (PS), as described by Manning[2]. PS is a corpus based approach that will yield the probability distribution of possible syntax constructions over a fixed corpus. PS makes no distinction between well and ill formed sentence constructions and assumes everything found in the corpus is well formed. In contrast, FVG's primary objective is to distinguish between well and ill formed sentence constructions and, in so doing, relies on corpus based parameters which determine sentence competency. In PS, a syntax of high probability will not necessarily yield a well formed sentence. However, in FVG, a syntax or sentence construction of high frequency value should yield a well-formed sentence, at least, 95% of the time satisfying most empirical standards. Moreover, in FVG, a sentence construction of 'high frequency value' could very well be represented by an underlying syntactic construction of low probability as determined by PS. The characteristic frequency values calculated in FVG are not measures of probability but rather are fundamentally determined values derived from exogenous principles which impact and determine corpus based parameters serving as an index of sentence competency.

What distinguishes FVG from other formal syntactic or computational linguistic approaches is that, in FVG, language is regarded as part of a broader dynamic framework with directionally. FVG utilizes exogenous physical principles based upon information theory principles in formally describing language and in NLP.

A better understanding might be gained of the significance of relying on external principles versus internal patterns by considering as a simile of language, rolling pebbles kicked in the sand. The paths of pebbles follow a similar configuration of paths every time they are

kicked in a similar fashion. We could ultimately establish a formal syntax which describes the patterns of kicked pebbles and disregard deviating patterns as flawed or extraordinary as do the generative grammarians disregard extra-linguistic phenomenon. However, what if the pebbles are intentionally kicked in a slightly different fashion with let's say the toes of the foot pointing slightly towards the right or the left? Do we disregard the ensuing patterns of the pebbles merely because they now deviate from our previously determined formal syntax for pebbles? A similar problem arises in linguistics. How do we determine whether deviations from well established syntactic patterns are to be disregarded as merely extra-linguistic or considered a fundamental failing of our formal syntactic approach? If internal patterns or fractals are all that we have to base our determination, it is likely we will merely dismiss the anomaly as extra-linguistic when it is actually the formal syntactic approach that is to blame. Now consider another possibility. We evaluate the rolling pebbles not merely in terms of their patterns but, also, in terms of their physics. We physically consider the point of contact of the foot, the momentum of the foot, the speed and angle of the foot at the point of contact and we physically translate these parameters into the directional mechanical energy transferred to the pebbles. Here, then, we have a better basis for determining if the roll of the pebbles is consistent with the point of contact of the foot, or , perhaps, reflects some extra-physical anomaly such as obstructions on the ground, high wind, or imperfections in the pebbles. Unfortunately, linguists have merely looked at language patterns and ignored the physics behind the patterns, having been influenced by the notion advocated by the Chomskyans that by closely analyzing, solely, the patterns associated with well-formed language we will eventually find the light at the end of the tunnel that explains all that needs to be known about language. I contend, however, that such a narrow pattern analysis approach is inadequate and will lead to our 'spinning of wheels' preventing us from reaching the end of the tunnel. Something more is needed.

In taking an information theory approach to language, the mutual information at the information source end or what is, also, referred to as relative entropy is factored in our descriptions of language. We need to account for, in some mathematical and realistic way, the interactions that take place in the symbolic units which represent language since these interactions are a fundamental part of language. In this regard, we cannot dismiss any linguistic phenomenon as extra-linguistic or anomalous merely because they do not conform with a particular formal representation until we come to recognize, at least approximately, the driving force behind the phenomenon.

Fully describing the language probability space requires our fully describing the mutual information/relative entropy of language symbolic elements in the probability space which, at first glance, would appear to involve an insurmountable level of complexity. However, we can circumvent the complexity by relying on simplified models of interactions which accurately describe the mutual information/relative entropy interactions. FVG is a model that takes into account the mutual information/relative entropy in the language probability space and, thus, unlike other linguistic approaches, attempts to distinguish, so to speak, when the pebbles rolling in the sand represent extra-physical anomalies or a genuine configuration directly determinable by the physical processes which set them in motion. FVG accomplishes this objective by assigning to terminals and non-terminals numeric values based upon mutual information/relative entropy phenomenon occurring in the language probability space.

Other approaches that rely on numbers or weights such as probability syntax, PS, are purely based upon frequency of occurrence, fixed corpus based parameters which approximate probability but disregard mutual information dynamics. As such, PS is not based upon nor is a measure of well formedness though there may exist some positive correlation between high probability syntax and well formedness. But the positive correlation which PS provides is not nearly high enough to satisfy the empirical requirement of being correct, at least, 95% of the time in determining a well formed sentence.

FVG is based upon statistical information theory principles and, thus, can be described as a Mathematical treatment of linguistic phenomenon relying on the mutual information aspects as well as the formal pattern forming aspects of language. The fundamental theory was introduced in Stepak[1] which, will not, in its entirety, be repeated here. Any attempt to model the mutual information end of the equation can only be approximate at best but a good first approximation can resolve many of insurmountable difficulties and problems we are often confronted with in describing and processing language. However, some of the details of FVG still need to be finalized such as the handling of 'questions' and 'parenthetical phrases' which follow a different dynamics, as distinct units, when compared to core declarative sentences. Questions, parenthetical phrases and subordinating clauses can be viewed as a sort of 'backtracking' which utilize a modified or somewhat of a reversal dynamics when compared to the dynamics in the declarative sentence core. Clauses in the declarative core resulting in sentence ambiguity or ill formedness can be converted to subordinating clauses that disambiguate and restore well-formedness to the sentence. The

subordinating clause will carry a characteristic phonetic stress depending upon its position in the sentence. Details that need to be further developed concern the interplay between iconic markers and phonetic stress.

To be sure, there is an interplay between the strength of an iconic marker and phonetic markedness and the strength of one diminishes the need for the strength of the other. This is vividly seen in transposing declarative sentences into 'questions' or vice of versa. Questions require an elevated frequency value towards the end of the sentence relative to the beginning whereas the reverse is true for declarative sentences. Thus, comparing the questions, 1. 'That is who?' versus 2. 'Who is that?' we find that 'who' in 1 requires greater phonological stress than 'that' in 2. since 'that', as a general determiner, is a much stronger iconic marker than 'who'. Transposing 1 and 2 into declarative sentences is accomplished by changing the phonological stresses on 'who' and that'. Since frequency value must diminish as we progress through a declarative sentence, we find that 'who' no longer requires phonological stress in transposed sentence 1. Furthermore, in transposed sentence 2. we have to be very careful we don't place any phonological stress on 'that' since, otherwise we are again asking a question. The phonological stress on that in transposed sentence 2. has to be lowered to below baseline stress for 2. to be fully understood as a declarative sentence whereas in transposed sentence 1., the phonological stress on 'who' need not be lowered below baseline stress. These phenomenon involving the interplay of iconicity and phonetic stress and non-iconic factors, in their totality, can be described as part of the physical dynamic principles related to the conditional or mutual information/relative entropies existing in the probability space in the language information source. (The treatment in the following sections primarily refers to the declaratory core sentence which comprises the major portion of the corpus).

FVG as a model of mutual information: The mutual information equation is as follows:

$$MI(X;Y) = H(X) - H(X|Y) \quad (1)$$

As reported by Cover-Thomas[3], it is the reduction of uncertainty in one random variable due to the knowledge of the other.

In the language probability space there are numerous such interdependencies between variables, too numerous to reasonably calculate and determine. However, as was described by Stepak[1], we can separate and distinguish two broad classifications of variables--iconic and non-iconic. The iconic variables interact with one another and determine the H(X|Y) probability space whereas the non-iconic variables would comprise of variables that can be calculated independently based upon their frequency of occurrence trigrams and determine the H(X) probability space. Confirming that language utilizes information theory principles is the fact that prominent in all languages are most common word lists, which bear resemblance to efficient artificial coding strategies utilizing a Huffman algorithm Stepak[1]. It can further be confirmed that MCW's in language require a second condition of iconicity since short coded words randomly chosen and completely devoid of iconicity are not found on most common word lists. Whether this theory is correct or not may be subject for further debate but, regardless, a mutual information model based upon the theory can allow us to reach certain conclusions about sentence structure that, otherwise, would not be possible and provide us with the tools for making correct language determinations.

FVG's basic premise is that well formedness of sentence structures requires the correct distribution of iconic markers in sentence structure in accordance with information theory principles Stepak[1]. Iconic markers have conditional intra-dependencies with each other and conditional interdependencies with non-iconic words. Iconic markers of higher frequency value take precedence in word order over those of lower frequency value based upon information theory principles, i.e., Det>adv>adj>prep. Furthermore, iconic markers are strongly associated with specific word categories, i.e., determiner-noun; preposition-object noun, adj.-noun etc. which comprise the iconic interdependencies with non-iconic words. These iconic interactions reduce the overall entropy of the probability space and can be viewed as a simplified model of mutual information as defined in Eq. 1. above. This model can, also, be used to get a better estimate of the overall entropy of natural language.

The inadequacy of previous estimates of language entropy: In 1950, Shannon[4] calculated the entropy of written English to be 1.3 bits per letter. The Shannon[4] calculation is flawed in several respects. (Obviously, Shannon was not a Linguist.) Firstly, it was not based upon a broad representative text of the English language. Secondly, it relied on introspective judgments (guesses) by subjects that would likely vary widely depending upon the education and demographic profile of the subjects. Shannon provides little or no information regarding experimental controls for the selection of his subjects and seems to have simply required that English be the natural language of the subjects. Finally, the data compiled by Shannon[4] is based upon introspective judgments at the

post decoding juncture of the information transmission paradigm whereas it was Shannon's objective to measure entropy of written language in the transmission channel. According to information theory principles which Shannon[4] himself developed, the entropy of the transmission at the post decoding juncture would not be the same as the entropy of the transmitted code. Similarly, the entropy of the transmitted code would not be the same as the entropy in the information source probability space.

Another estimate of language entropy by Brown *et al.*[5] yielded a result of 1.75 bits per letter. Here, the methodology utilized represents an improvement over Shannon's effort in that no introspective data was relied upon and the entropy of the actual coded transmitted message is what was measured. However, it would appear that the Brown *et al.*[5] study, likewise, is flawed since they relied on a trigram language model based upon a Shannon-McMillan-Breiman equation that fails to give due consideration to mutual information factors given by Eq. 1. above. (Shannon's calculation of entropy, likewise, failed to give due consideration to mutual information.) Both calculations, that of Shannon of 1.3 bits per letter and that of Brown *et al.*[5] of 1.75 bits per letter would appear as overestimating the actual entropy when considering humans have been experimentally shown as having the capacity to comprehend up to 25 phonetic segments (letters) per second, Liberman[6], (albeit typical reading speed would be somewhat slower). Assuming a 25 phonetic segment capacity per second and that greater than one bit of information per letter would require 2 neural inputs per letter, the total number of required neural inputs per minute would total 3000. This figure would have to be doubled to account for a lower bound 2 neural outputs per letter and then multiplied 5 times to account for a reasonable 20% efficiency, (Hopfield models are actually only 10% efficient), resulting in a rough estimate of 30,000 neural synapse activations per minute. Greater than one bit per letter may represent too great a neuro-physiological load and one would expect, therefore, the actual entropy of English to be substantially less than one bit per letter.

The frequency value equation used in FVG shares similarities with a Bayesian equation modeled on a noisy channel paradigm that is typically used in spell checking. What is sought is the maximum or above threshold values of iconic marker probabilities. If the sentence frequency value is not above a minimum threshold value, the sentence is deemed ill-formed. However, the metric employed is not true probability since a constant K, representing the size of the largest syntactic category, is factored into the equation to normalize non-iconic components of the equation. The iconic components, thus, are represented as some multiple fraction of K whereas the non-iconic components are represented as some rational number substantially less than K and insignificant in terms of contributing to the overall frequency value of the sentence. The frequency value equation can be represented by a Lambda equation that is easily adapted to a functional programming language such as Lisp, Scheme, or Haskell which simplifies programming and assures rapid computation. The basic FVG equation used in natural language processing for a sentence with 3 iconic components is as follows:

$$\text{Lambda } [x^6y^4z^2]/K^{11} \quad (2)$$

The lambda arguments are the frequency values of iconic markers in function argument complexes representing some fraction of K. (2) can be generalized to:

$$\text{Lambda } [\,[x_1^{2n}][x_2^{(2n-2)}]\ldots\ldots \ldots\ldots[x_n^{(2n-(2(n-1)))}]\,]/(K^{(n\wedge 2+n-1)}) \quad (3)$$

n represents the number of iconic markers in the sentence and the number of bracketed terms in Eq. 3. The frequency value equation is designed so that frequency value can be plotted versus progression through the sentence. For a sentence comprising of 3 iconic complexes the data points for the y axis are, x^2/K; $[x^4y^2]/K^5$; $[x^6y^4z^2]/K^{11}$, with the variables being bound by their corresponding iconic marker values serving as arguments. With the y axis representing frequency value, a well formed sentence will give high negative slope and greatest area under the curve whereas ill formed sentences will result in flatter negative slopes and less area under the curve and will, also, tend to yield lower overall sentence frequency values. Sentences that were devoid of requisite iconic markers in any of the argument function complexes (these comprise of the obvious ungrammatical sentences) would result in the frequency value of Eq. 3 to drop to zero resulting in an uncharacteristic vertical line in the graphical representation. As a result, the sentence would immediately be recognized as ill-formed. Of course, Eq. 3 could be further refined or modified based upon facility of use in programming, software application, or corpus requirements.

The frequency value equation, with a few modifications, could serve as a model for mutual information in calculating the entropy of English. By merely removing the K factor and simplifying Eq. 3 to $[(p(x_1))(p(x_2))\ldots\ldots(p(x_n))]$, we essentially convert Eq. 3 into a probability model for measuring iconic marker sentence configuration. The probability model equation ignores

probability occurrences of non-iconic words so we have a means of estimating the frequency occurrence of iconic markers in the context of function-argument complexes and sentence structure. Typically, declarative core sentences will have 4 to 8 iconic markers embedded in function-argument complexes that assign frequency value to a sentence. And iconic markers per sentence in declaratory-core sentences typically will have an overall frequency of occurrence of anywhere from 1/64 to 1/128,000 (a value of at least ¼ will be above the threshold value for well formedness in short sentences). Questions, parenthetical phrases, and subordinating clauses would require separate treatments based upon the frequency of occurrence parameters in the question parenthetical and subordinate clause portion of the corpus. Let's assume we train on the declaratory core sentence portion of a large representative corpus such as the one relied upon by Brown et al. and find that the average number of iconic markers per sentence is 5. Furthermore, let us assume the average frequency value we compute per sentence for iconic markers is 1/1024. The iconic markers we measure have an overall frequency of occurrence in the corpus of 33% and have an average word length of 3 letters as compared to 6 for non-iconic words Nadas[7]. We assume the average sentence length is 15 words (conjunctions are regarded as new sentences). Thus, we can improve upon the Brown et al estimation of entropy per letter by doing some averaging and calculating the entropy of iconic markers. Treating sentences as the basic unit in a Shannon-Mcmillan-Breiman treatment, the entropy of iconic markers per sentence would be 10. Since there are 5 iconic markers in the sentence the entropy per iconic marker is 10/5 and per iconic marker letter is 10/15 or 67. There are 60 non-iconic letters in the sentence that carry 1.75 bits of information based upon the calculation of Brown *et al.*[5] and 15 iconic marker letters that require. 67 bits of information based upon our calculation. Thus, overall, the average bits per letter drops to 1.53. My purpose here was merely to illustrate the concept since the data inputs do not represent actual data but the kind of data that could be reasonably expected. But, in actuality, what I have just illustrated is not a calculation based upon mutual information but merely an intuitive arithmetic method that on the surface would appear to represent a means for obtaining a better approximation of the upper bound of entropy. Upon closer examination, the above calculation is inadequate for calculating the entropy of natural language based upon a mutual information model.

The above calculation does not measure the joint probabilities representing the relationship between iconic marker sentence dependent components and the sentence. Rather, it treats sentence dependent components and the remaining portion of the corpus as comprising of independent probability densities. Typically, averaging independent probability densities will result in an accurate calculation of the overall entropy of the system. However, in natural language, joint Entropies must be considered. In natural language, the sentence dependent probabilities or grammaticality of the sentence is part of a joint probability, mutually dependent upon the sentence dependent and non-sentence dependent components of the corpus. Thus to accurately calculate the overall entropy we must subtract the entropy representing the joint probability (grammaticality) from the entropy representing the overall word based corpus frequencies calculated by one-way trigrams. A calculation based upon Eq. 1 will be illustrated in this study.

But before continuing, it should be noted that an assumption that the n-gram models provide us with a pure measure of the non-conditional, dis-joint entropy would not be correct. There is a substantial amount of reduction of the H(X) value derived from a n-gram based calculation due to grammaticality effects. Thus, H(X), also, represents grammaticality to the extent corpus based n-gram frequencies will be effected by grammaticality. The correcting factor, H(X|Y), represents grammaticality effects stemming from word positional constraints and sentence structure not taken into account in a n-gram analysis. For the sake of simplicity and to avoid confusion with H(X|Y), in this study, H(X) is operationally defined and referred to as a non-grammatical, disjoint, n-gram based entropy value when, in actuality, it does include a grammatical component.

Mutual information-A closer look: Other approaches to calculating English entropy, Shannon[4], Brown[5] rely on linear n-grams that assume one probability density, (or disjoint multiple probability densities) and assume that the calculation will fully account for all frequency letter or word occurrences in the language. A somewhat modified approach by Teahan and Cleary[8] rely on a PPM algorithm which merely calculates a probability distribution of letters based upon previous contexts of letters of varying lengths. Here again, language is treated as having disjoint probability densities and it is assumed that the PPM algorithm will adequately account for all probability interactions between letters and words. N-grams and PPM's are essentially smoothing techniques that calculate an overall average of interactions, smoothing out the bumps along the way which otherwise could be attributable to mutual information interactions and multiple joint probability densities. A FVG calculation, on the other hand, attempts to accentuate and highlight the bumps along the way by relying on two joint probability densities, the probability density associated with the

sentence core interacting with the overall sequential word probability space. A jointed two-probability density approach visibly is more valid since the probability of any given word in a sentence merely does not only depend on words preceding it but, also, on its position in the sentence. For instance, the probability of a determiner occurring at the beginning of the sentence is greater than near the end of the sentence regardless of what appears before it. If one is to consider punctuation as words, then, the determiner would become more predictable but, then, we are faced with the dilemma of predicting the occurrence of the end of sentence 'period' punctuation mark which would increase in probability as one progresses through the sentence irrespective of what appears before it. And there are many positions in the sentence where the determiner would have a zero probability such as between an adverb and adjective, after an attributive adjective, or preceding a preposition. The objective, then, is to fully account for the probability density associated with the sentence structure to obtain a more accurate estimate of the entropy of English. N-grams and PPM's do not accomplish so much in spite of these techniques being able to produce random texts that appear to perform well, but, on closer analysis, compromise to a large degree well-formedness. Substantial improvement would be achieved by embodying the sentence core in a separate probability density space which interacts with the general corpus.

The question, thus, arises, how can this be best accomplished. We must first identify an overall probability of a sentence as a base unit. We can establish a probability of a sentence by calculating the probabilities of sentence dependent constituents in their local environments such as the closed class functions word categories which, in FVG, comprise of a good portion of the iconic markers or MCW's. The remaining words are sentence independent such as the open class nouns, non-modal verbs etc. However, in FVG, adjectives are considered a closed class and iconic marker due to their characteristic phonetic stress. This is because in any sentence or phrase an adjective proceeding and describing a generic open class noun will always have a characteristic stress which, in the context of the sentence, is sentence dependent.

Phonetic stress serves as cues even in written English. The fact that any generic noun can only be preceded but not followed by a generic attributive adjective serves as a cue that the adjective will carry phonetic stress. Though the stress in its own right is not significant in written English, the reader knows its there to accentuate the attributive adjective that will always precede the noun its describing. Such phonetic stress would be absent if the attributive adjective had a natural positional relation to the noun as in French. By natural I mean determined by the relative frequencies of the noun and adjective so that,as in French, attributive adjectives generally follow nouns except when the adjective is frequently used. Thus, in French, the adjective would not be regarded a sentence dependent component. Its frequency of occurrence would be its corpus based frequency. This, however, does not mean that French would have a higher overall entropy when compared to English. In French, the lack of grammaticality associated with the adjective is compensated for by the increased use of the determiner which would lower the entropy. The details of the concepts and axioms relied upon in FVG become rather complex and will not be fully described in this study. Further explanation is provided by Stepak[1]. Suffice it to say, using what are sentence dependent constituents, an overall sentence probability can be calculated that would comprise of a separate probability density and represent the mutual information part of the calculation for determining a more accurate overall entropy of the language.

To illustrate this approach, we rely on some of the concepts introduced by Stepak[1]. K is a constant representing the size of the largest syntactic category of words which in the English language and most languages are nouns. K is a large number but is not infinite even though an infinite number of nouns could be invented. K represents the number of noun lemmas in the lexicon. Similarly, other syntactic categories have attributed sizes relative to the size of the noun category. Thus, based upon Johanssen and Hofland[9] mathematical analyses of the LOB corpus, the size of verbs would be K/10, adj-K/3, adv-K/25. Also, based upon Johanssen and Hofland, we can approximate the relative probability of occurrence of sentence dependent words based upon their associated nonsentence dependent words. Thus, P(Det|Noun) = 1/4, or sentence dependent frequency of occurrence of Det is 1/4. We disregard the frequency of occurrence of the noun since it is not sentence dependent. All that the sentence requires is that one or more nouns exist in the sentence comprising of any noun in the noun lexicon. Similarly, other sentence dependent frequency of occurrence values for other iconic markers of the sentence are: Preposition=1/40; Pronoun=1/22.

Values presented here are approximations based upon the frequency analysis of Johansson and Hofland[9]. Any revision of values would probably be to the upside which would result in a further decrease of the joint entropy when compared to the entropy calculated by the given figures.

Adjectives acquire part of their sentence dependent probability from phonetic stress since characteristic sentence dependent stresses are considered common

feature attributes which serves to increase the frequency of occurrence parameter. Adverbs acquire their sentence dependent probability from phonetic stress and commonality of the suffix which serves as an iconic marker. Thus, we have:

Adj. = (1/10)(1/3) = 1/30; Adv. = (1)(1/25) = 1/25

where the first factor represents the frequency of occurrence and the second factor the relative size of the category yielding the product representing the sentence dependent probability. There are other phonetic stress related markers such as carried by mass, abstract and plural nouns which substitute for the absence of the determiner.

Based upon the values, above, immediately one notices the sentence dependent iconic markers carry a substantially higher probability then their corresponding corpus based frequency of occurrence. For instance, in the sentence dependent treatment, the determiner carries a probability of 25% whereas in the word based corpus calculation it typically carries a lesser frequency value for determiner tokens, i.e. the-6%, a-3%, an-1.5%. Similarly, the calculation of frequency of occurrence of other sentence dependent iconic markers is substantially greater than the corresponding frequency values determined by a word based corpus frequency. Thus, if we were to consider the corpus comprising of two independent probability densities, averaging the two probabilities would be justified in calculating an overall language entropy and the higher the frequency of the sentence dependent components, the lower the overall entropy. But our mutual information model proposes that the two probabilities densities are mutually dependent so that each sentence dependent component has two attributed frequency values, one determined by its local environment which is sentence dependent and the other determined by an overall corpus based frequency. In the mutual information model, the higher the frequency values of sentence dependent components the higher the overall entropy of the language as can be seen by Eq. 1 as page 3. Nonetheless, relying on sentence dependent frequency determinations for sentence dependent constituents (grammaticality), rather, than corpus word based frequencies alone, results in a reduced overall entropy based upon Eq. 1 The sentence dependent components comprise of the probabilities that result in H(X|Y) of Eq. 1. The non-sentence dependent components such as the open class nouns and non-modal verbs do not contribute to the value of H(X|Y) since any word of the noun or verb open-classes fulfill the basic requirement of the sentence comprising of a noun and verb. Since there are K such nouns and K/10 such verbs their sentence dependent frequency values are K/K and (K/10)/(K/10) respectively or equal to 1.

However, it should be noted that the assumption that open-class nouns and verbs carry a sentence dependent frequency of 1, of course, represents an oversimplification since the frequency of open-class nouns and verbs are not evenly distributed in the corpus. Thus, some sentences may require a restricted subset of the open-class nouns or verbs. But for purposes of calculating an upper bound language entropy this assumption will suit our purposes for now since it assures we will arrive at the maximum possible overall entropy. The lower the joint (grammatical) entropy, the higher the overall entropy.

We can now proceed to calculate the overall entropy based upon mutual information as expressed by Eq. 1. We once again rely on the reasonably expected hypothetical data used earlier, page 5 and the entropy calculation by Brown *et al.*[5] of 1.75 bits per character. The average number of iconic markers per sentence is 5 and the average sentence dependent frequency value we compute is 1/1028. (Typically, there are more than 5 iconic markers per average sentence and the sentence dependent frequency value is less that which is indicated when one also includes the non-declaratory core portion of the corpus. But since we are merely interested in calculating a theoretical upper bound for the overall entropy, an overstated sentence frequency value would suffice since it would not artificially reduce the final entropy value). There are on average 15 words per sentence and 75 characters per sentence so the corpus based entropy per sentence is 1.75x75 = 131.25 which represents H(X) per sentence. Based upon the sentence dependent frequency value of 1/1028, the sentence dependent joint entropy (grammaticality) is 10 which represents H(X|Y) per sentence. Thus, based upon Eq. 1, the overall entropy drops by 10 bits per sentence and 133 bits per character resulting in an improved upper bound of 1.62 bits per character. My purpose here, again, was merely to illustrate the concept of extracting sentence dependent features representing a second mutually dependent probability distribution that serves to reduce the overall entropy calculation when compared to a calculation based upon corpus word frequencies alone. Here, the amount of the reduction is not really crucial from a theoretical standpoint. The theoretical point to be made is that the upper bound calculations relying on one way forward n-grams first utilized by Shannon[4] and later adopted by others does not represent the upper-bound. This is an important point since using only the Shannon one way forward n-gram approach in comparing the Entropies of different languages and genres is likely to lead to misleading comparisons.

Based upon the above approach, it can now be concluded that:

$$\text{English entropy} < H(X) \quad (4)$$

where, H(X) represents the entropy calculated by the forward n-gram analysis. The significance of equation (4) is that no matter how sophisticated or refined a forward n-gram analysis might be, the actual theoretical upper bound will be less than that which is calculated by the forward n-gram. The inequality of (4) can be removed by a simple modification that accounts for conditional joint entropy as follows:

$$\text{English entropy} = H(X) - H(X|Y), \quad (5)$$

which is equivalent to Eq. 1.

The mutually dependent probabilities within a language can be seen as a measure of the language's grammaticality. The greater the grammaticality, the greater the entropy of the mutual dependent probabilities and the lower the overall entropy of the language. Grammaticality in its own right can be viewed as being constrained by an information theory requirement to keep the entropy directly associated with it below a threshold level. Grammaticalization can not be too extensive since, then, grammatical rules become too burdensome to learn and may even begin to overlap or become self contradictory which defeats the purpose of grammaticalization. There is an inherent upper limit information theory threshold, therefore, that limits the attainable entropy associated with grammaticality, thus, assuring that grammaticality can only reduce the overall entropy of a language up to its grammatical entropy threshold upper limit. Above the grammatical entropy threshold level, grammatically becomes too cognitively expensive in its own right rendering impossible its efficient usage. The grammaticalization entropy threshold is less than the corpus word based frequency entropy of the language, thus, grammatical entropy cannot reduce to zero the overall entropy. Natural language can be seen as comprising of two Entropies, the entropy associated with the corpus word based frequencies calculated by forward n-grams and the joint entropy associated with grammaticality. In comparatively measuring the overall Entropies of different languages and genres, therefore, merely calculating the corpus based entropy associated with forward n-grams will yield erroneous results.

English, as do all natural languages, exist in the form of consecutive sentences. These integral sentence units cannot be ignored in any calculation of the entropy of the language. The n-gram models and the PPM models represent oversimplifications that are untenable since probabilities of words (or letters) are sentence (word) position dependent in addition to being pre-word dependent. The model relied upon by n-grams and PPM only considers the preceding words or letters which results in a model that is essentially non-ergodic since the probabilities of sentence dependent words (letters) change as a function of sentence (word) position, assuring that calculations based upon these n-gram non-ergodic models will be inaccurate. The FVG model restores the ergodic nature to the probability space by recognizing that the probability space comprises of two probability distributions that interact with each other that serves to reduce the overall entropy. Thus, the probability stemming from the position of the word in the sentence is fully taken into account before trekking across the sentences to calculate the probabilities of the non-sentence dependent constituents of the sentence.

A cognitive perspective: At the language information source level or deepest level, language is not fully encoded. The language information source comprises largely of conscious mental imagery and unconscious hardwiring determined by neurological constraints that are a by-product of information theory principles. At the language information source, we conceive of objects in a three dimensional space corresponding to a mental imagery space. For a specific given singular noun object, the definite determiner in the mental image is not below or above or before or after the noun object. We mentally conceive of specific noun objects with the determiner as a blended in feature of the noun object. Non-specific noun objects would have the indefinite determiner feature blended in the mental imagery. So are other features of the noun object such as color, size etc. blended in our mental imagery of the object. It is only because written communication and (to a lesser extent oral communication) are two dimensional that decisions have to be made where to place the corresponding word objects representing features of the noun object, but these feature word objects are connected and inter- related with the noun word object, representing the mutual information portion of the information source. Therefore, looking at any given sequence of words which intersect the mutual information portion of the information source, we notice a substantial difference in the sequential probabilities if we compare the forward sequence with the backward sequence.

The probability that any given noun will follow a definite or indefinite determiner is proportional to the corpus based frequency of occurrence of the noun. Similarly, the probability that a determiner will appear anywhere in the sentence (other than positions where it would have zero probability) is proportional to the corpus

frequency of the determiners (indefinite and definite, approximately 8%) which, in turn, is proportional to the frequency of occurrence of noun phrases or nouns. But when one considers the frequency of occurrence of the indefinite or definite determiner preceding the noun one find a much higher frequency of approximately 25%.

P(Det|Nword) >>> P(Nword|Det) (6)

Here, as well as with other sentence dependent components that intersect the mutual information probability space, i.e. Adjectives, adverbs, prepositions, a substantial difference in probabilities is obtained when comparing the forward versus backward N-gram. N-grams used for measuring written English entropy only calculate the forward probabilities and, therefore, are not accurate measures since they do not measure the true dependency of sentence components in the three dimensional cognitive information source. The fact that the noun occurs after the determiner is an non-arbitrary assignment determined by information theory principles Stepak[1], to transpose a three dimensional conceptual space into a two dimensional communicative space. From an entropy perspective, however, what needs to be measured is co-occurrence of these mutual information components irrespective of sequential directionally in a two dimensional communicative medium.

At the mental imagery information source level, cognitive objects can be viewed as compressed sentences and written sentences can be viewed as decompressions of cognitive objects governed by information theory principles. A forward n-gram measurement of the two dimensional sequencing is misleading since, due to information theory constraints, much of the co-occurrences of sentence components is backward sequenced and can only be fully measured in the backward probability sequencing of co-dependent elements. N-grams rely on word based corpus frequencies which are substantially lower than corresponding FVG values since N-grams assume the language information source (cognitive source) is homomorphic with the two dimensional forward coding of the communicative medium. However, the language information source is not homomorphic with the two dimensional forward sequencing in the communicative medium and therefore, a one to one n-gram forward analysis will be inaccurate. The interdependent frequencies of features of cognitive objects existing in the information source do not correspond on a one to one basis with the cognitive decompression two dimensional forward sequencing in the communicative medium. In most instances, information theory constraints promotes homomorphism with the backward sequencing in that words and categories of higher frequency tend to be frontloaded in sentences Stepak[1]. For instance, in a two word sequence comprising of a Determiner followed by a co-occurring noun, the backward frequency of occurrence will always be greater than the forward frequency of occurrence since information theory principles require the Det to be frontloaded in the sequence due to its higher frequency. Therefore, in calculating the frequency of occurrence of the Det-nounword two word sequence, the backward analysis will always be higher since the frequency of the nounword serves as the denominator in the backward analysis whereas in the forward analysis the frequency of the determiner serves as the denominator.

Since frequency of occurrence of Det>>>Nword

Det-Nword/Nword>>>Det-Nword/Det (7)

where, Det-Nword/Nword represents the backward analysis.

Using the FVG values introduced on pages 6-7, Det-Nword (backward)= 1/(4K) and Nword= 1/K, which yields the characteristic value of ¼ for the frequency value of the determiner as a sentence dependent component.

It is this higher backward frequency which is representative of entropy of this two word sequence in the context of sentences due to the high co-occurrence dependency existing between determiner and noun in the cognitive information source. Det-Nword is not homomorphic in a two dimensional forward n-gram analysis of the communicative medium but is homomorphic in the backward analysis.

Entropy as a theoretical basis for

Determining ideal sentence length: What has just been described in the previous sections has significant theoretical implications with respect to the length of sentences. One might ask why does the English language or any natural language comprise of sentences that can be averaged to a certain finite length. The answer resides in the fact that the length of sentences are directly proportional to the number of sentence dependent components comprising primarily of closed class function words. The functions words describe the physical dimensions in which we exist. The function words are compressed or blended in with cognitive objects at the language information source. Sentences are as long as are required to decompress the blended in features of cognitive objects in the information source.

Thus, in language, sentences are as long as they have to be in order to communicate accurately the physical dimensions of our existence. If they were shorter

than this bare minimum requirement, the entropy of the language would increase and information transfer would become more burdensome. For instance, let us assume we wanted the English language to have a shorter average sentence length. To accomplish this we would have to entirely delete from the language one of the sentence dependent function words such as 'in' but at the same time substantially increase the size of the lexicon so that for every non-functional word there was a new word having an 'in' connotation. Doubling the size of the lexicon would reduce the frequency of occurrence of individual words which would effectively increase the overall entropy. Also, removing a function word would decrease grammaticalization and the grammatical entropy which would increase the overall entropy based upon. However, let's say on the other hand we wanted to lengthen the average sentence length. Here we would delete a non-physical dimensional category of words and replace it with a new function word. The overall entropy of language would drop and grammaticalization would increase but our ability to convey the nuances of information contained in the category of words we discarded would be lost. Thus, our ability to convey a substantial amount of information would be lost.

Thus, the unique finite sentence length found in language could be said to represent an equilibrium between entropy and the need to convey information. A sentence's length is pre-determined by the need to convey the essential physical dimensions in which we exist. Undercutting this bare requirement would result in an increase of entropy whereas a surplusage of the requirement would result in a reduction of the information carrying capacity of the language.

FVG and NLP: Generative grammar approaches fail since they cannot accommodate the context sensitive nature of natural language. Probability syntax fails because the underlying probability distribution of available syntax forms assures that a correct syntax will not be chosen at least 95% of the time. That being said, I do not mean to suggest that I regard Natural language as context sensitive in a mathematical sense. Practically speaking, in applications Natural language mimics some of the aspects of context sensitivity but actually is not a context sensitive language. Natural language represents a separate, distinct, class which should be added to the Chomsky hierarchy. What distinguishes Natural language from other languages is that strings must be shorter than a certain threshold length, practically speaking, from both a cognitive and well formedness perspective. Thus, the pumping lemma can not be used to determine if the language is context free or regular since any elongation of the string could invalidate the string. Furthermore, Natural language displays virtually an infinite number of different types of required finite intra-sentence matchings when one considers word collocations and nonce occurrences that in certain sentences tend to obstruct or override established rules of the language. Thus, Natural language could be depicted as not being describable by any given set of fixed rules unless terminals are given the capacity to produce feedback non-terminals that have the capacity to modify or supplement the previous established productions. FVG accommodates the unique characteristics of Natural language by assigning to terminals and non-terminals numeric values which allow any given high frequency word collocation matching to take precedence over baseline non-terminal productions. From a NLP perspective, therefore, FVG modifies traditional computational approaches by adding to the lexicon numeric values assignable to words and syntactic categories Stepak[1]. The numeric values assigned are based upon mutual information calculations rather than corpus probabilities and, thus, provide a measure of sentence competency and well formedness rather, than, a mere probability of occurrence in the corpus.

Formal mathematical proofs cannot be utilized to prove that Natural language belongs to a particular category of language such as regular, context free, or context sensitive. The application of mathematical formalisms for such purpose by Chomsky and others represents an authoritarian misuse of propositional logic applied to phenomenon that more appropriately requires an additional abductive phase of hypothetical reasoning. (For a detailed description of the term abduction see Magnani[10]). The layer of hypothetical reasoning that is absent pertains to recognizing that Natural language represents a unique irreducible category of language that cannot be assigned to one of the fully described established categories of languages. This notion can be easily formalized by use of the metric entropy.

Natural language is characterized by a finite, fixed entropy. What distinguishes Natural language from other languages is its characteristic fixed entropy. Thus, when applying mathematical formalisms which requires extrapolating a Natural language string to an infinite string in order to categorize it as regular or context-free, the individual words of the string, other than the infinite recursively matched word embeddings, have a frequency of occurrence of 1/infinity, (the inverse of infinity). Alternatively, if we rely on an infinitely long repeating phrase connected by an infinite matched word embedding such as an if-then pairing, at the very least, the end of sentence period, treated as a word, has a frequency of occurrence of 1/infinity. In both cases, therefore, the

entropy of the infinite string is Infinity or undefined which violates the requirement that Natural language has a fixed, finite entropy.

However, practically speaking, from experience and measurements that have been done, we know that entropy of Natural language has a far greater constraint than merely being a fixed, finite value. Entropy of Natural language must be within a restricted range of value. We can conservatively state that the entropy of Natural language is typically less than 2 bits per letter, albeit, the actual upper bound is likely to be substantially less. Thus we can define Natural language 'A' as:

$$NL(A^{Ent<2}) \quad (8)$$

The superscript in Eq. 8 substitutes for the closure star (*) and indicates that combinations and sequences of the characters in strings of the Natural Language 'A' are constrained to the extent that they cannot result in a language having an entropy per character of 2 bits or greater. No such entropy limitation exists in non-natural, artificial languages regardless of their category, i.e. regular, context free, context sensitive, or type 0 recursively enumerable, etc.

Relying on this concept, we now have a blueprint for designing an artificial Natural language that could possibly serve as the first man-made Universal Natural language. There is a vitally important international need for a Universal Natural language that is politically neutral and which provides no advantage to any given Nation or government. Furthermore, such a Universal language would fulfill an important intellectual need in allowing for the creation of scientific journals written in a Universal Natural language where the editors' mother tongue would no longer be a factor in the selection of intellectual works for publication. There is now a visibly clear condition that must be met leading to the creation of a man-made artificial Universal Natural language. The Universal language must be designed so that each character in the language carries no greater uncertainty (Entropy) than 2 bits per character so that the language can be adapted for human use.

DISCUSSION

The Philosopher, Charles Sanders Peirce, introduced the term 'Abduction' more than 100 years ago. 'Abduction' refers to the early stage organizational study of phenomenon that requires something other than an orderly inductive or deductive logic. Historically, it was thought that philosophers should only concentrate on the realm of orderly logic. It was thought that hypothetical inferencing was too deeply embedded in the inner workings of the mind to justify an objective philosophical description. Peirce introduced the notion that the study of mental activity leading to the inference of explanatory hypothesis that precedes the establishment of formalisms was structured in its own right so to justify a disciplined philosophical analysis. This study accepts this latter notion and advances the notion that formalisms and abstractions in science often are detached and even irrelevant to the underlying phenomenon that is alleged to be described by the formalism. What is needed in these cases is additional hypothetical reasoning, what Peirce precisely refers to as 'abductive reasoning', to close the gap between formalisms and what these formalisms allegedly claim to describe. These type of gaps in the field of linguistics remain much too wide.

Thus, in the preceding study, it was my purpose to close this gap somewhat. In so doing, one can not engage in a study that glitters with inviolable formalisms that appeals to everyone. So doing would defeat the purpose of the study. Certainly, abductive reasoning must be utilized but it cannot be said precisely where the abduction ends and formalistic descriptive justification begins. Such is the nature of abductive reasoning.

In this study, I have highlighted two of several instances in science and mathematics where phenomenon have been prematurely and hastily described relying on a propositional logic and mathematical formalism when there should have been a further exploration into the underlying phenomenon's relevance to the formalism through an hypothetical or abductive reasoning. These and other instances confirm that abductive reasoning and descriptive justification do not follow a reversible continuum but that the latter oftentimes can mislead and camouflage the need for the former. The first instance is addressed throughout the study and pertains to the premature and authoritarian formalistic approach utilized by Shannon[4] in calculating the written entropy of English. Here, Shannon[4] relies entirely on mathematical formalisms and theory but never appears concerned with validifying the application of his formalisms to the underlying dynamics of the phenomenon which he claims to be measuring. The scientific community widely accepted without scrutiny Shannon's approach which was indicative of Shannon's authoritarian influence in the scientific community. The second instance, described on pages 1-2, 10-11, pertains to another authoritarian formalistic approach, first introduced by Chomsky and later adopted by others in describing or categorizing natural language which, likewise, was premature and more appropriately required additional abductive reasoning. Here again, the findings were generally accepted by the intellectual community without scrutiny.

Such being the case, it seems to me fairly apparent that these issues are now fundamentally philosophical issues-philosophical issues requiring the examination of the shortcomings of scientific regimen that results in premature descriptive judgments in instances where further hypothetical development and abductive reasoning is required. Some of the philosophical questions are: How could a scientific regimen lead us into delusionary representations of the real world? Is there a test for distinguishing scientific findings that are delusionary from those that are not? How much does authoritarian intimidation, which is a psychological-social phenomenon and not scientific, divert effective scientific inquiry and endeavor? In the study, I have not addressed these questions directly, but they linger in the background and it is my hope that the reader will give consideration to them.

CONCLUSIONS

N-grams were relied upon by Shannon[4] in the calculation of entropy of written English. I have shown in this paper that standard n-gram approaches such as those used by Shannon and others in the calculation of English entropy serve as inadequate models for calculating language entropy. Shannon's approach to language entropy was authoritarian in that he relied on mathematical formalisms in which the relationship between the formalism and the phenomenon being described was not fully verified. Here as well as in many other instances in science, a propositional logic was utilized prematurely since it will yield clear and non-oblique publishable results, easily describable and will appeal to the empirical senses and the mentality of the scientific community. Further, it provides the false sense or false hope that a more painstaking and burdensome abductive type of reasoning can be sidestepped. Unfortunately, the dilemma of our modern age of science is that scientific research is paid for by large and impersonal governmental or corporate research funding bodies and research proposals based upon abductive reasoning fail to be as impressive to funding institutions as those based upon propositional logic and descriptive mathematical formalisms, even when the mathematical formalisms are not justified. For this, the scientist carries part of the blame but so does the funding institution. There is needed a new understanding and realization in the research community that an abductive reasoning can and will more often lead to more significant and long lasting results and discoveries than a formalistic approach that has not been fully justified.

The calculation of language entropy is an important metric that can serve a wide array of useful purposes. Shannon's utilization of mathematical formalisms, albeit impressive in their own right, do not serve as an adequate method for measuring the entropy of language. The crucial point is that there never was any justification to accept Shannon's forward n-gram measurements as true and accurate measurements of language entropy. Nonetheless, the scientific community and linguists have generally accepted Shannon's approach to language entropy without scrutiny due to Shannon's authoritarian influence in the scientific community and his ability to mesmerize the scientific community with novel and hybrid mathematical formalisms that, unfortunately, were not always shown or verified as accurately describing the contextual frameworks to which they were applied such as in the case of language entropy. The entropy of language comprises to a large extent of joint Entropies and Shannon's use of the forward n-gram comprises of compressed independent probabilities which measures disjoint entropy and can only, at best, approximate joint entropy indirectly.

To sum up, entropy is an important metric serving several important functions in language. It is important that a correct and precise model for calculating entropy be employed. It was my purpose in this study to provide some of the hypothetical reasoning needed to justify the implementation of improved methods for measuring entropy and, in so doing, also validate FVG as a more viable grammatical approach when compared to other grammatical approaches such as PS. It is hoped that this study will help to promote a general consensus in these areas.

REFERENCES

1. Stepak, A., 2003. A Proposed mathematical theory explaining word order typology UCREL Technical Papers, Vol.16, Part 2, pp: 744-753. Lancaster Univ.,UK.; Revised Version, TAAL Print Archive, Univ. of Edinburgh. In: Proceedings 4th International Conference on Cognitive Science, 2: 634-640. Sydney, Australia; Website: http://www.e-commercesolutions.net/members/omconstruct/menu.html
2. Manning, C., 2003. Probabilistic Syntax. In Rens Bod, Jennifer Hay and Stefanie Jannedy (Eds), Probabilistic Linguistics. Cambridge, MA: MIT Press, pp: 289-341.
3. Cover-Thomas, 1991. Elements of information theory. New York: John Wiley and Sons, pp: 18.

4. Shannon, C., 1993. Shannon Collected Papers, Sloane, Wyner (Eds.), Prediction and entropy of Printed English, Hoboken, N.J.: John Wiley and Sons, pp: 194-208.
5. Brown *et al.*, 1992. An estimate of upper bound for the entropy of English. Computational Linguistics, 16: 79-85.
6. Liberman, A. *et al.*, 1967. Perception of the speech code. Psychol. Review, 74: 431-461.
7. Nadas, A., 1984. Estimation of probabilities in the language model of the IBM speech recognition system. IEEE Transactions on Acoustics, Speech, Signal Processing, 32: 859-861.
8. Teahan, W. and J. Clearly, 1996. The entropy of English using PPM-based models, Proceedings of the 1996 Data Compression Conference (DCC).
9. Johansson, S. and K. Hofland, 1989. Frequency Analysis of English Vocabulary and Grammar Based Upon the LOB Corpus, Vol. 1 and 2, Oxford, Clarendon Press.
10. Magnani, L., 2001. Abduction, Reason and Science. New York, Kluwer Academic/Plenum.