Frequency Value Grammar and Information Theory

Mar.18th, 2004

Abstract:

This paper convincingly shows that previous efforts to calculate the Entropy of written English are based upon inadequate n-gram models that will result in an overestimation of the Entropy.

Frequency Value Grammar, FVG, is not based upon corpus based word frequencies such as in Probability Syntax, PS. The characteristic 'frequency values' in FVG are based upon 'information source' probabilities which are implied by syntactic category and constituent corpus frequency parameters. The theoretical framework of FVG has broad applications beyond that of formal syntax and NLP. In this paper, I demonstrate how FVG can be used as a model for improving the upper bound Entropy calculation of written English. Generally speaking, when a function word precedes an open class word within a phrasal unit, the backward bi-gram analysis will be homomorphic with the 'information source' probabilities and will result in word frequency values more representative of cognitive object co-occurrences in the 'information source'.

1. Introduction

I previously laid the groundwork for Frequency Value Grammar(FVG) in papers I submitted in the proceedings of the 4th International Conference on Cognitive Science (2003), Sydney Australia, and Corpus Linguistics Conference (2003), Lancaster, UK. 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 (2003). 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' will 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.

The theoretical framework of FVG has broad applications beyond that of formal syntax and NLP. In this paper, I will demonstrate how FVG can be used as a model for improving the upper bound calculation of Entropy of written English.

2. External versus Internal

What distinguishes Frequency Value Grammar, FVG, from other formal syntactic or computational linguistic approaches is that, in FVG, language is regarded as part of a broader dynamic framework with directionality. 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. I introduced the fundamental theory in Stepak (2003) 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).

3. FVG as a Model of Mutual Information

The mutual information equation is as follows:

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

As reported by Cover,Thomas (1991), “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 I described in Stepak(2003), 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(2003) 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(2003). 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 equation (1), above. This model can, also, be used to get a better estimate of the overall Entropy of natural language.

4. The Inadequacy of Previous Estimates of Language Entropy

In 1950, Shannon calculated the Entropy of written English to be 1.3 bits per letter (Shannon 1993). His 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 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 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 (1992) 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 Brown et al's work is , likewise, 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 equation (1), section 3, 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 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 1967, (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:

(2) Lambda (x^6)(y^4)(z^2)/(K^11)

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

(3) Lambda [x1^2n][x2^(2n-2)].....
.....[xn^(2n-(2(n-1)))]/(K^(n^2 + n - 1))

n represents the number of iconic markers in the sentence and the number of bracketed terms in equation (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^4)(y^2)]/K^5; [(x^6)(y^4)(z^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 equation (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, equation (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 very well serve as a model for mutual information in calculating the Entropy of English. By merely removing the K factor and simplifying equation (3) to [(p(x1))(p(x2)).....(p(xn))], we essentially convert equation (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 typically will have an overall frequency of occurrence of anywhere from 1/64 to 1/128,000 [1](a value of at least ¼ will be above the threshold value for well formedness in

[1] Reference is made here only to the declarative core sentence. 'Questions' , 'parenthetical phrases', and 'subordinating clauses' would require a separate treatment based upon frequency of

short sentences.) Let's assume we train on 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 (1984). 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. 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 equation (1) in section 3 would have to be applied. In accordance with equation (1), to accurately calculate the overall Entropy we must substract the Entropy representing the joint probability (grammaticality) from the Entropy representing the overall word based corpus frequencies calculated by one-way trigrams[2]. A calculation based upon equation (1) will be illustrated in the following section.

---

occurrence parameters in the 'question', 'parenthetical', and 'subordinate clause' portion of the corpus.

[2] An assumption that the tri-gram models provide us a pure measure of the non-conditional, dis-joint Entropy of the Corpus 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 paper, H(X)

5. Mutual Information-A Closer Look

Other approaches to calculating English entropy, Shannon (1993), Brown (1992) 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 (1996) 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[3]. This is because in any sentence or phrase an adjective

---

is operationally defined and referred to as a non-grammatical, disjoint, n-gram corpus based Entropy value when, in actuality, it does include a grammatical component.

[3] Phonetic stress serves as cues even in written English. The fact that any generic noun can 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

proceeding and describing a generic open class noun will always have a characteristic stress which, in the context of the sentence, is sentence dependent. The details of the concepts and axioms relied upon in FVG become rather complex and will not be fully described in this paper. Further explanation is provided in Stepak(2003). 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 I introduced in Stepak(2003). 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 mathematical analyses of the LOB corpus (1989), 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.[4]
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

---

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 no positional relation to the noun as in French.

[4] Values presented here are approximations based upon the frequency analysis of Johansson and Hofland. Any revision 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.

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 equation (1) of section 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 equation (1) The sentence dependent components comprise of the probabilities that result in H(X|Y) of equation (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.[5]

We can now proceed to calculate the overall Entropy based upon the mutual information equation (1) of section 3. We once again rely on the reasonably expected but hypothetical data described in section 4 and the Entropy calculation by Brown et al (1992) 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.[6] There are on average 15 words per sentence and 75 characters per sentence so the corpus based Entropy per sentence is 1.75 x 75 = 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 equation (1) in section 3, 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

[5] 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. 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.

[6] Typically, there are more than 5 iconic markers per average sentence and the sentence dependent frequency value is less than that which is indicated. But since we are merely interested in calculating an upper bound for the overall Entropy, these overstated values will suffice.

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 (1993) 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.

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 contradictary 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 so as to render impossible its efficient usage. This 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 for the reason probabilities of words are sentence 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 change as a function of sentence 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.

6. 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%.

(4) P(Det|Nword) >>> P(Nword|Det)

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 (2003), 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 directionality 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(2003). 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

(5) Det-Nword/Nword>>>Det-Nword/Det
where Det-Nword/Nword represents the backward analysis

Using the FVG values introduced in section 5,
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.

7. 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 equation(1) of section 3. 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.

8. 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 (2003). 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.

Conclusion

N-grams were relied upon by Shannon (1993) 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. I believe I have shown this convincingly and that my effort is the first to refute the forward directional n-gram approach adopted by Shannon and mimicked by others in the calculation of language Entropy. The calculation of language Entropy is an important metric that can serve a wide array of useful purposes. It is important that a correct model be used. There is strong implication, therefore, that probability syntax and other formalisms that have relied on the theoretical model framework of n-grams, also, require closer scrutiny.

FVG rests on a broad theoretical framework that can be used for purposes other than formal syntax or NLP such as serving as an improved model for calculating an upper bound for the Entropy of natural language.

References

1. Stepak, A 2003 A Proposed Mathematical Theory Explaining Word Order Typology, UCREL Technical Papers,Vol.16, Part 2, pp 744-753, LancasterUniv.,UK., Revised Version, TAAL Print Archive, Univ.of Edinburgh; Expanded Implementation Section, In Proceedings 4th International Conference on Cognitive Science, Vol. 2, pp 634-640, Sydney, Australia.

2.Manning, C 2003 Probabilistic Syntax. In Rens Bod, Jennifer Hay, and Stefanie Jannedy (eds), Probabilistic Linguistics, pp. 289-341. Cambridge, MA: MIT Press.

3. Cover-Thomas 1991 Elements of Information Theory, page 18, New York, N.Y:John Wiley & Sons.

4. Shannon, C 1993 Shannon Collected Papers, Sloane,Wyner (eds), Prediction and Entropy of Printed English, pp. 194-208, Hoboken, N.J.: John Wiley & Sons.

5. Brown et al 1992, An Estimate of Upper Bound for the Entropy of English. Computational Linguistics. 16(2), 79-85

6. Liberman, A. et al, 1967 Perception of the Speech Code. Psychological Review 74 (1967): 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(4), 859-861

8. Teahan W, Clearly J 1996 The Entropy of English Using PPM-Based Models, Proceedings of the 1996 Data Compression Conference
(DCC).

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