

Selective* Pressures on Symbolic Systems[†]

Phillip Potamites

January 28, 2010

Abstract

As measured by *conditional relative entropy*, syntactic categories are more predictable in structural configurations than in sequences. This predictability of structural categories suggests their likely significance for learning, processing, and language robustness. However, Shannon’s fundamental insights that “information is entropy” and “more entropy is more efficient” suggest that evolutionary pressures on language must balance between efficiency and robustness. Words and categories manifest contrasting upper and lower limits on predictability, while the way that low entropy carriers can facilitate higher entropy transfer demonstrates an even deeper way that the spontaneous order found in the evolution of life and culture can facilitate a universal tendency to maximize entropy.

Contents

1	introduction	2
2	method	5
3	data	7
3.1	POS	8
3.2	words	8
3.3	DP	11
3.4	roots	12
3.5	Chinese	14
4	natural	17
5	to be false	19
6	outside evidence and new proposals	23
7	conclusion	26
A	additional arguments and supplementary statistics	28
A.1	mathematical motivations for Gibbs-Shannon entropy	28
A.2	data reductions	29
A.3	data sizes	30
A.3.1	PTB	30
A.3.2	NYT	30
A.3.3	STB	30
A.3.4	GGG	30
A.3.5	C (mysql)	31
A.4	sample extraction	31
A.5	DP recategorization	31
A.6	null items of the PTB	33
A.7	PTB POS raw entropy	33
A.8	C raw entropy	33
A.9	auxiliary redundancy	36
A.10	top 10s	36
A.11	Zipf: whiter than noise.	37

*This word is ambiguous across its evolutionary and linguistic senses. In evolution, it means the success of certain strategies over others, whereas in linguistics it means the influence of a word in restricting its relations. These meanings are freely conflated here, because this paper is about the evolutionary selection of linguistic selection.

[†]The initials of this title are an unfortunate accident. For statistical software, I’m not a fan of (proprietary, gui) SPSS. I like (free, cmd line) R. Any data processing in this paper that wasn’t done in R was done with (free) Python, and, as may be apparent, the document was set with the maddening but incomparable (and free) L^AT_EX. It’s remarkable how accessible this stuff has become.

“Die Entropie der Welt strebt einem Maximum zu.”
The Entropy of the World tends to a Maximum.
(‘the Second Law of Thermodynamics’)
-Clausius 1865

1 introduction

The above principle is already well-established in physics, but can seem to fundamentally contradict the order found in life, society, culture, and language.

Coffee cooling, balls falling, ice melting, dams breaking are all entropy maximizing. All are minimizing differentials. Insofar as life or society seem to maximize differentials, they seem to fundamentally oppose the second law of thermodynamics. This paper, however, is about phenomena that reconcile this seeming contradiction between the organic soft and the inorganic hard sciences. Language does this because information is entropy and because it organizes itself to benefit information transfer.

When colors mix, their entropy (what J.W. Gibbs called “mixedupness”) increases. A smooth purple may be called more uniform than strips of red and blue both because it is viewed as a single color, and because the likely location of any given particle is more uniformly distributed over the observed space. The possible locations of a particle become more diverse. That is even supposed to be the reason *why* the universe *naturally*, i.e. spontaneously, mixes things up. The second law of thermodynamics tells us that spontaneous change must produce positive entropy.

Similarly, the division of labor in society implies a greater diversity of occupations, but, as Durkheim and Marx both lament, typically leads to greater monotony for each individual. The entropy of occupations has increased even if the entropy of one’s day has decreased. One’s day has become quite predictable, while the job of any randomly encountered stranger has become much harder to predict. The greater variety of occupations makes jobs harder to predict – especially if they are more uniformly likely. A more uniform distribution is less redundant on particular outcomes. The entropy is higher because the outcome is harder to predict.¹ The division of labor is thus an immediate example of how maximizing entropy contributes to efficiency in social development, but predictable disparities in life and society can still lead to the impression that evolution opposes the unpredictability entailed by entropy.

Some words appear in a wide variety of contexts and other words have more restricted distributions. Conditioning on a given word, the entropy of its contexts will be higher where there is a more uniform distribution over more possibilities and lower where fewer specific possibilities are more likely. This quantity can just as well be measured for any various syntactic categories and relationships. Insofar as linguistics is interested in contextual restrictions, conditional entropy provides an (inverse) quantitative measure.

Entropy is just a scalar value calculated from a probability distribution, and it is defined to be higher for less concentrated distributions and lower for more highly concentrated distributions. Information sciences have converged with thermodynamics in measuring this quantity as below:

$$(1) H = -K \sum_{i=1}^n p_i \log p_i$$

-Shannon (1948), Theorem 2, p. 11

¹In a sense, the second law tells us that it is easy to predict that things will get harder to predict. However, it makes *useful* predictions insofar as macroscopic outcomes are likely in proportion to *sums* of microscopic probabilities.

H is (one of) the typical symbols for entropy; K and the base of the log are constants which merely effect our units of measurement; when the base of the log is 2 the units are called bits, whereas if e , they are called nats. Bits will be used throughout this paper. $p_1 - p_n$ are the probabilities of each possible state²; thus,

$$(2) \sum_{i=1}^n p_i = 1$$

Notice that, in eq. (1), it's really the negative sign that makes H go up for lots of little p 's and down for less higher p 's; since for all p_i , $p_i < 1$, the log is always some negative number of larger absolute magnitude for lesser p .

Shannon also calls it, 'how much information is "produced"', 'how much "choice" is involved', and 'how uncertain we are of the outcome' (p. 10).³ It may seem odd that uncertainty is equated with information,⁴ but the more predictable a signal, the less informative its reception. The more states a signal can transmit, the more complex the messages it can communicate. Similarly, insofar as more probable messages can be made more concise, the more efficient communication will be.

Shannon gives the name *relative entropy* to "the ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols" (here, it will be abbreviated 'rH'). Relative entropy is 0 where there's just one symbol, one outcome, for everything, and 1 where symbols are uniformly distributed over all possibilities. He defines *redundancy* as 1 minus this value. Redundancy is thus 1 where everything is the same, and 0 where everything is as different as possible. Shannon also makes the remarkable observation that relative entropy also defines "the maximum compression possible when we encode into the same alphabet" (p. 14). Insofar as some 'alphabet' can take advantage of shorter representations for more frequent items, that alphabet can compress the average size of such representations. In fact, the maximum possible compressibility of a message is a ratio exactly equivalent to the proportion of its observed entropy to its potential maximum entropy (see Shannon (1948), p. 18-19 for an extremely concise example or related literature for more extensive explanation). Insofar as higher relative entropy is less redundant and less compressible, it is more efficient.

This paper particularly concentrates on relative entropy with the intention of compensating for variations in raw frequencies. Thus, it limits maximum entropy by *raw count*, if that is less than vocabulary size. When we look at each word as a system of contexts, the maximum diversity is limited both by the possible vocabulary and that word's raw frequency, whichever is lower. An event cannot display more different outcomes than it has occurrences.⁵ Where frequencies are lower

²The above formulation of the equation also assumes discrete states, which will suit us here, but in principle the sum could instead be integrated.

³Shannon barely but cites the thermodynamic relevance in a footnote acknowledging Boltzmann, instead listing independent mathematical reasons which are given in appendix A.1, but Jaynes (1965) stresses that Gibbs was responsible for the more generally correct formula stated above. It's said that John Von Neumann, who was responsible for its quantum mechanics application, told Shannon to call it entropy both because it was already used in statistical mechanics under that name, but also because "nobody knows what entropy really is, so in a debate you will always have the advantage." The emphasis can hopefully remain on the first reason (though the wiki page on "Entropy" uncouthly claims that Von Neumann said the latter was more important).

⁴Gibbs also fretted that the concept "will doubtless seem to many far-fetched, and may repel beginners as obscure and difficult of comprehension" (Graphical Methods in the Thermodynamics of Fluids (1873)).

⁵Where maximum entropy is a *uniform distribution over all possibilities*, it would be the same for each word, and thus the relative entropy would be monotonic with raw entropy. Where the vocabulary size is higher than the word count, that entropy is higher than just if all the instances all came out differently. This paper then uses the lower estimate as a *hopefully less biased estimator*, in factoring out raw counts, though then we may want to be wary of

than possible vocabulary size, the maximum entropy for a set of N events is given by the following formula.⁶

$$(3) \text{ maximum possible entropy} = -\sum_{i=1}^N \frac{1}{N} \log \frac{1}{N} = -\log \frac{1}{N} = \log N$$

The black line in figure 1 shows the limit (from 1-10000) that raw frequency (N) establishes on maximum entropy in black (i.e it's just a graph of $f(N) = \log_2(N)$). The red dots represent words (in this frequency range) from the Penn Treebank by their frequencies (on the x-axis) and entropies (on the y-axis) from the probability distributions that they specify for following words.

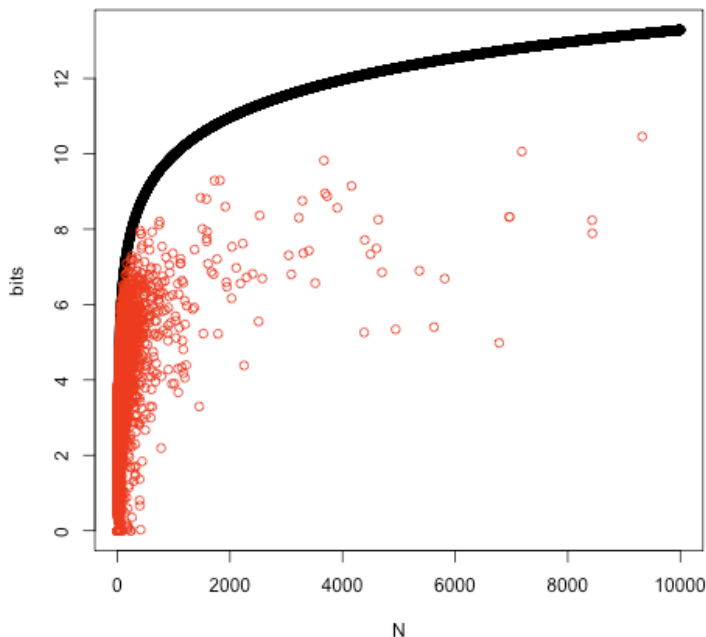


Figure 1: Sequential raw entropy of the PTB and its logical limits

This paper will typically concentrate on the ratio of the red dots (the observed entropy of words) to the black limiting line above it (the possible entropy). Thus, this paper uses *conditional relative entropy* to estimate *flexibility*, which is efficient, and the inverse of restrictiveness, redundancy, and predictability.

the lower denominator inflating ratios. An even lower, interesting maximum could be established, by calculating the entropy of a uniform distribution just over the set of types that the word was observed with, so that we were instead basically estimating, *within observed contexts*, how constrained are the distributions. The raw count limited maximum entropy is then also somewhat of a compromise between these 3.

⁶The reader should be aware that, though the basic concepts are the same, and precisely defined as such, the analysis of statistics used here makes no use of the kinds of techniques used to implement maximum entropy learning algorithms as advocated in such pivotal research as Jaynes (1957), Berger et al. (1996), or Ratnaparkhi (1998).

$$(4) \text{ relative entropy} = rH = \frac{\text{observed entropy}}{\text{maximum possible entropy}} = \frac{-\sum_{i=1}^n p_i \log p_i}{\log N}$$

In fact, almost all of the comparisons made in this paper hold good for both raw and proportional entropy calculations, but proportional entropy is typically presented as the fairer comparison. The reader is invited to contact the author if such comparisons are desired.

These comparisons will demonstrate that abstract categorical structures can improve predictability without necessarily sacrificing information transfer; it is hypothesized that these are strategies arrived at by an evolutionary negotiation of efficiency and robustness.

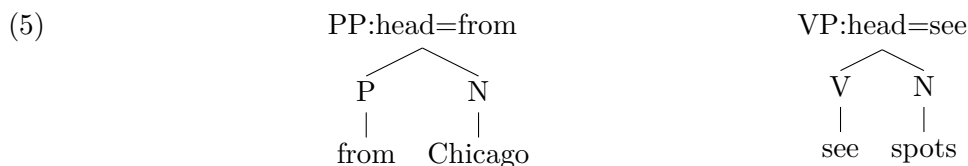
The following sections present the methodology and contrast the predictiveness of different syntactic and morphological theories. After that, issues of falsification are considered. Then outside experimental evidence for the psychological significance of contextual predictability will be presented and novel variable manipulations suggested.

2 method

This paper examines the “contextual entropy” of different linguistic elements in different relations. For a relation like ‘sequentially-following’, the probability of any following word, given the preceding word, has been calculated and the entropy for this preceding word, in this relationship, was calculated from this set of probabilities according to equation 1.

For example, suppose we had a corpus where “the” appeared 10 times and was followed 5 times by “boy”, and 5 times by “girl”. Then the probability of “boy” given “the” would be .5, and the probability of “girl” given “the” would be .5, and the sequentially following entropy of “the” would be 1 bit ($-1 * ((.5 \log_2 .5) + (.5 \log_2 .5)) = 1$). The relative entropy would be also be 1, since the situation is a state of maximum possible entropy.

To compare syntactic predictability with sequential predictability, “head relations” were extracted from the PTB. The “head” of a phrase is considered the dominant word of a phrase, that is, the word which defines the nature of the phrase, and upon which all the other words of the phrase are thought to depend. For instance, prepositions are thought to be the heads of prepositional phrases just as verbs are for verb phrases⁷:

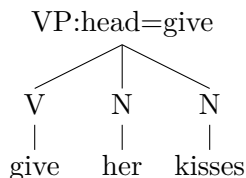


Syntactic predictability was estimated by determining the probability of different heads given any observed dependent. For instance, the above examples would show that ‘Chicago’ can be headed by ‘from’, and ‘spots’ can be headed by ‘see’. They also show that ‘N’ (nouns) can be headed by ‘P’ (prepositions) and ‘V’ (verbs). Such “part-of-speech categories” are also referred to as POS tags; the upcoming results contrast their behavior with that of words.

Because the PTB is not restricted to binary branching, phrases can also consist of multiple dependents for a single head:

⁷The reader may wish to note that the PTB uses more specific tags for specific kinds of nouns and verbs, even idiosyncratically tagging P, ‘IN’.

(6)



In the above example, both ‘her’ and ‘kisses’ are headed by ‘give’. This multiple dependency is the main reason why heads were predicted instead of dependents. One might expect there to be more psychological saliency in predicting dependents from heads than in predicting heads from dependents. However, dependents necessarily have one head, whereas heads can give rise to varying numbers of dependents, and this variation raises complex questions about how to treat multiple dependents. The simplest, most faithful method would treat a group of dependents as a single outcome, but such a method would miss the affinity between events which shared certain dependents but not others. Alternatively, counting each dependent as a separate outcome would likely result in sets of probabilities with sums larger than one (as if, assuming ex. 6 was the only known example of ‘give’, one concluded it had 2 100% likely outcomes; of course, one should not conclude that it has 2 50% likely outcomes either). These are questions which need to be addressed by future research, but for these preliminary investigations, this simplest, singular prediction was adopted. A simple, singular outcome for both structural and sequential relationships also presumably makes them more comparable.

Thus, words are basically guaranteed to have exactly as many ‘next word’ instances, as they have ‘selected by’ instances. For categories however, structural phrases introduce new titles so their counts can differ from the sequential category counts, which are limited to terminal tags. Appendix A.3 lists counts for all the corpora and extractions used.

The PTB is parsed and tagged with POS categories, but does not annotate head relations, so a simple ASCII comparison was performed between phrase and member categories to find the most similarly named categories, and thus choose the best candidate for “determining the nature of the phrase”. To be specific, the sum of the absolute difference of the ASCII value of each character in the mother and each daughter’s POS designations were collected, and the daughter with the least sum, i.e. the one with the most similar name, was selected as the best head. 0’s were assigned where names had shorter lengths. Ties went to the first daughter, which is almost certainly wrong for nouns but hopefully haven’t skewed the results too much. The entire method relies on the naming conventions of the corpus giving good indications of the correct heads. The PTB’s naming conventions guarantee that nouns head noun phrases and verbs head verb phrases, but its tag for prepositions, ‘IN’ had to be changed to ‘PH’ to guarantee that prepositions headed prepositional phrases (PP). Other modifications were also experimented with, and they are described below. Example extractions of both methods are provided in appendix A.4. There is most certainly a great deal of noise in this system, but our POS results will suggest that this head selection algorithm provided fairly systematic predictability. Furthermore, subsequent sections relying on a state-of-the-art, automated dependency parse (Minipar on the ACQUAINT corpora) and the even more convincing human annotations of the Sinica Treebank for Chinese will bear out the same results.

The PTB also includes 9 “null” elements, that is, elements with no audible manifestation, but presumed to be important for structural analysis and interpretation. These items include silent complementizers, ‘traces’ which indicate the interpreted relations of fronted items such as wh-words, and the understood silent subjects of non-finite verbs. Given their inaudible nature, they

were removed from the sequential analysis, but left in the syntactic analysis, so as to remain faithful to its theoretical stance. These 9 items account for 64057 tokens, and they are listed in appendix A.6.

Calculating raw entropy off of observed events makes such a statistic extremely susceptible to data scarcity issues. Note that any item which has only occurred once, has only one observed outcome, and thus would be estimated at an entropy of 0 and undefined rH. Any item which occurs twice will necessarily have 2 identical outcomes, or 2 different outcomes, each with a 50% chance, and thus a raw entropy of either 0 or 1.⁸ Under these these circumstances, it seems reasonable to insist that an item be sufficiently represented in a corpus before including it in the analysis.

For all of the analyses in this paper, items had to appear at least 100 times to be included. Of the 39352 word-types in the Penn Treebank (PTB), making up 949088 event-tokens, 17730 appear only once. Insisting on counts over 100 leaves only 1072 types, but still covers 710009 tokens. Though this reduction of data points (as types) may seem quite large, such requirements generally *reduce* the mean rH and provide a more well-rounded histogram, as demonstrated in Appendix A.2.⁹ Thus, this paper adopts these strict requirements as providing more cautious comparisons than more inclusive analyses would.

The populations presented are typically those of types. An analysis of tokens would provide a closer approximation to real-time entropy exposure. A token analysis would increase focus on high frequency items and generally reduce variance. The focus on types provides a more abstract index of linguistic populations.

For all the comparisons to follow, the results of T-tests are provided, though their reliability is debatable. T-tests require either normality or sufficient size of the test distributions. As just noted, sample sizes (for words) were greatly reduced for the sake of the more sufficient representation of their elements, though such reductions did not necessarily achieve plausible normality. Future research will be needed to determine what mathematical distributions most closely model these data, and what statistical tests should be applied. Until then, the reader should rest assured that T-tests on the larger samples typically sustained the main claims.

Thus, this paper adopted strict requirements for its analysis which are intended to improve the persuasiveness of its arguments; significantly, its claims are generally just as true without them. That is to say, the basic patterns presented below generally show up with both raw and relative entropy and for reduced and complete data sets. The reader is invited to contact the author for any desired evidence not already provided.

3 data

Sequential relations of natural language should probably be considered *contaminated by noise*. That is to say, when one seeks to interpret language, (relevant) sequential relations are constrained by crucial *constituent* relations. For instance, hopefully there is no question about who kissed who in the following sentence.

(7) The girl with the dog kissed the boy.

⁸See appendix A.1 for a quick example of the low-end discontinuities.

⁹In fact, this number was also adopted because the computer language data passes a Kolmogorov-Smirnov test for normality at this cut-off point, and that provides somewhat more legitimacy to the T-tests used for mean comparisons. However, even that data doesn't pass a Shapiro-Wilks normality test, and the natural language data typically doesn't, so this motivation is far short of sufficient.

Chomsky is, in fact, frequently credited with the claim that language learners *do not* consider sequential hypotheses. As Steven Pinker puts it, “One of Chomsky’s classic illustrations of the logic of language involves the process of moving words around to form questions,” (Pinker (1994), p. 40, citing Chomsky’s *Reflections on Language* (1975)). He then contrasts the following examples:

- (8) a. *Is a unicorn that eating a flower is in the garden?
 b. Is a unicorn that is eating a flower in the garden?

The first sentence is problematic because it merely “fronts” the first auxiliary encountered (the ‘is’ related to ‘eating’), rather than the correct main clause auxiliary (the ‘is’ related to ‘in the garden’). Such examples are thought to provide evidence that hierarchal relations are more relevant to linguistic processes than sequential relations.

The evidence here will show that both kinds of relationships balance instance redundancy, while syntactic relations show better categorical predictability. Thus, learners can hypothesize abstract redundancies without sacrificing communicative efficiency.

3.1 POS

Figures 2-3 indicates that, for POS categories, syntactic relations are more predictive than sequential relations. A (weak) T-test is consistent with the conclusion.

	seq	syn
mean rH	.5525	.3787

t = 4.6201, df = 77.565, p-value = 1.500e-05

There are only 44 terminal POS tags in the PTB, but 70 tags including larger phrase categories. A cautious reader may fear that our normalizing denominator could be overcompensating for unrealized possible tag relations. However, as shown in appendix A.7, raw entropies bear out the same conclusion. After excluding tags not represented at least 100 times, 41 terminals covered 948897 tokens, while 54 categories covered 1012774 tokens (see appendix A.3 for tables of sample sizes).

3.2 words

Somewhat surprisingly, word instantiations do *not* follow a similar pattern of improved syntactic predictiveness. Figures 4-5 indicates that words are less predictive in syntactic than sequential relations. A (weak) T-test is consistent with this claim.

	seq	syn
mean rH	.6252	.7194

t = -14.4309, df = 2121.018, p-value < 2.2e-16

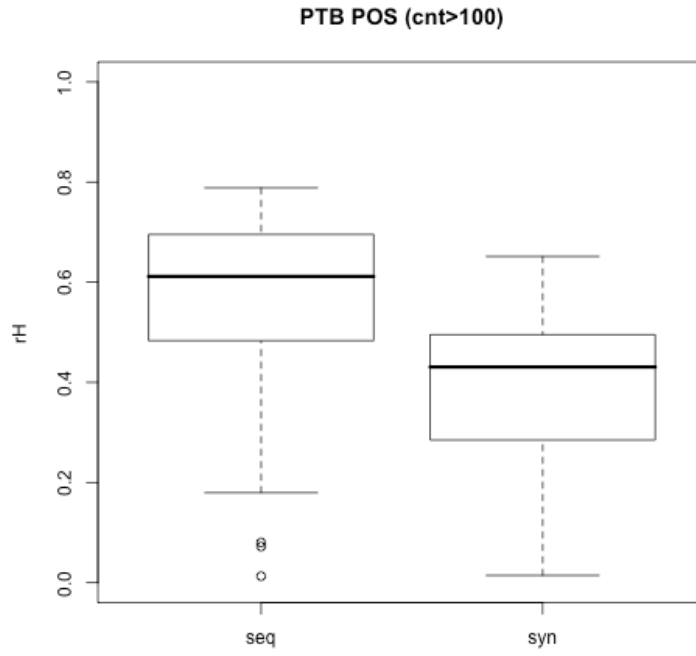


Figure 2: POS sequences are harder to predict.

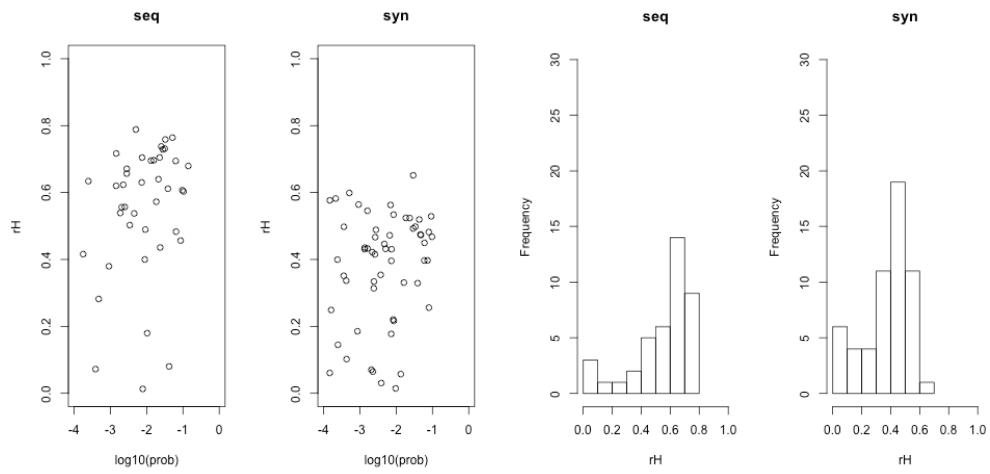


Figure 3: rH by log-frequency, and rH histograms

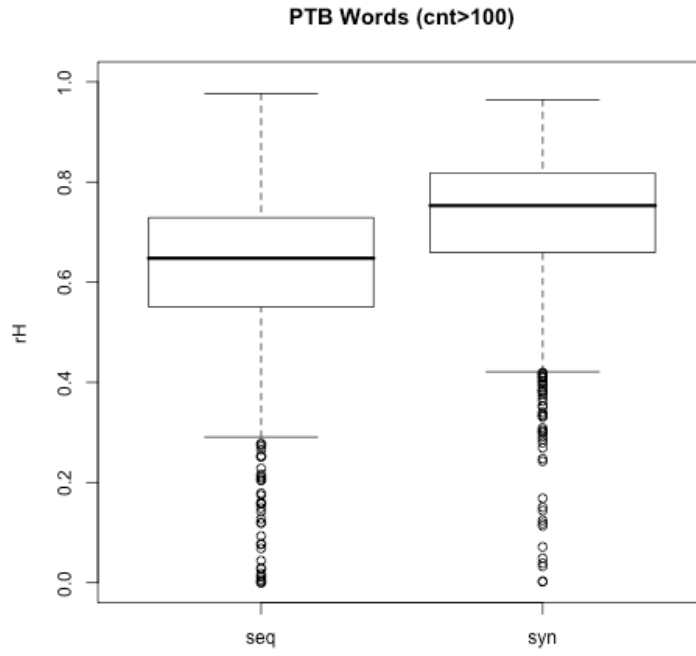


Figure 4: Word sequences are easier to predict.

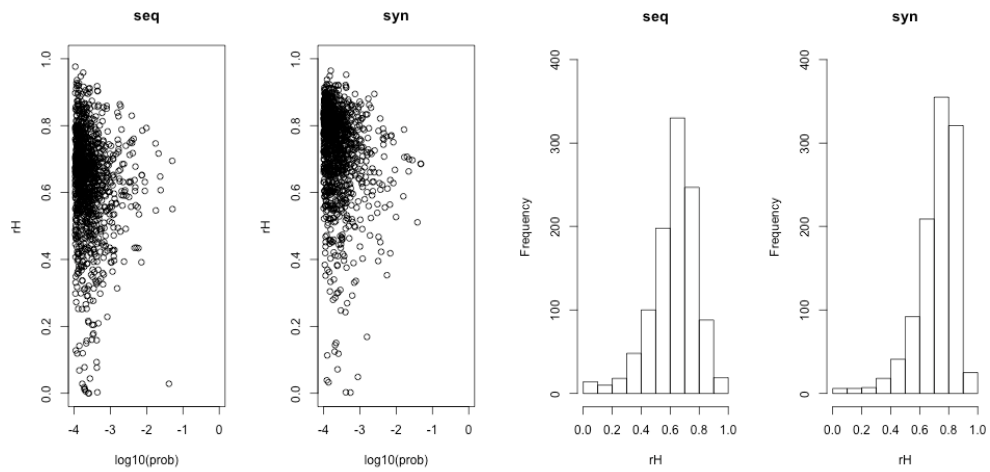


Figure 5: rH by log-frequency, and rH histograms

Excepting the “null items” noted above, the word counts between sequential and syntactic analyses are identical (see appendix A.3 for tables of sample sizes).

It also should be noted that though the sequential word series is entirely determined without any reference to the POS tags considered above, the syntactic relations are *read off* of tags’ similarities and entirely dependent on the same hypothesized structural configurations of the more predictive categories compared above. Thus, the higher word diversity is taking place entirely within the confines of the same systematic categories that offered better predictability above.

The evidence thus bears witness to a system which both places a premium on complexity for its efficiency and implements macroscopic order to facilitate it.

3.3 DP

There is, in fact, a current syntactic perspective which can significantly improve word predictiveness. Taking determiners (such as ‘the’, ‘a’, ‘this’, ‘that’, etc.) as the heads of “determiner-phrases”, rather than nouns heading “noun-phrases”, means that, in predicting the “head-of” relation, determiners must be predicted from nouns, rather than vice versa.¹⁰ Since determiners are a rather small class, they are much easier to predict than the whole open set of nouns. A vast number of phrases now become represented by “the” and “a”.¹¹ However, since this system does *not* provide any improved predictability at the abstract (categorical) level, the increase in instance (word) redundancy could be interpreted as information loss.

The complete set of category changes are shown in appendix A.5.

Here, in fact, a T-test is not able to distinguish the DP-headed syntax’s POS predictions from that of the N-headed syntax, though the DP-syntax seems lightly less predictive (i.e. has slightly higher relative entropy):

POS	DP-syn	NP-syn
mean rH	.3937	.3787

t = 0.4713, df = 108.85, p-value = 0.6384

The word heads, for the DP-syntax, have however become much easier to predict:

words	DP-syn	NP-syn
mean rH	.5914	.7194

t = -19.6051, df = 2130.779, p-value < 2.2e-16

They are even easier to predict than the word sequences, and though quite close, a T-test even suggests significance.

words	Seq	DP-syn
mean rH	.6252	.5914

t = 4.903, df = 2149.968, p-value = 1.015e-06

¹⁰In fact, the flat, multi-braching tendencies of the PTB (as in “(NP (DT a) (JJ level) (NN playing) (NN field))”) means that most noun modifiers also become subordinates of DT.

¹¹Nouns like generics don’t necessarily require determiners, in which case nouns can still end up heading DPs (in the present analysis, cf. fn. 12).

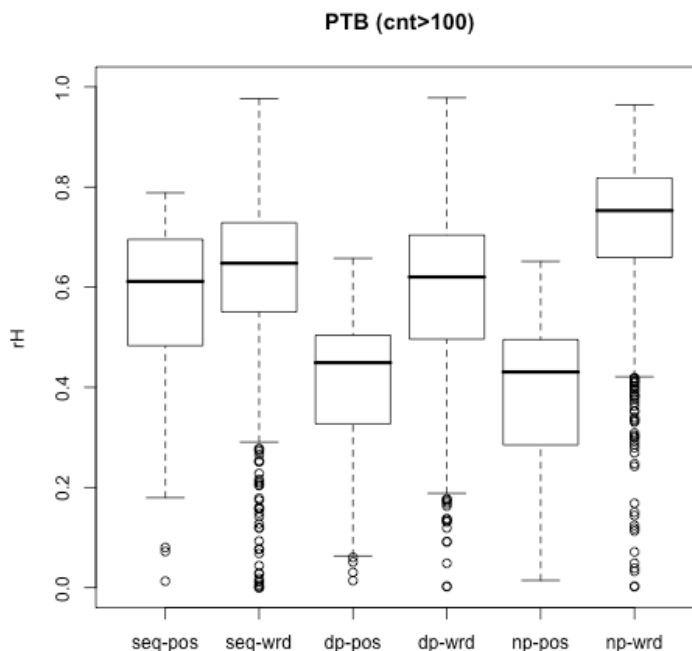


Figure 6: DP words are easier to predict.

So particular head decisions can make syntactic word relations even (slightly) more redundant than sequential strings. However, the lack of improved POS predictions is a bit disappointing. In that sense, the DP-theory is no better, no more predictive than the NP-theory.¹² However, the NP-theory is just as predictive at the abstract level, and it is also “packing more information” into the embedded word relations that it attempts to predict.

Dependency parses, like the one used in the following section, also tend to subordinate determiners to nouns. This decision facilitates their use in thesaurus extraction and other “semantically”-focused applications where properties of animacy, edibility, imageability, stativity, and so on, may be even more relevant than syntactic categories. Although predictivity can be improved by a function word focus, without corresponding abstract improvements, that focus might also be viewed as detrimental information loss.

3.4 roots

Linguistic analysis can also distinguish word manifestations from their morphological roots. For instance, the root of “had” is ‘have’ and the root of “is” is ‘be’. Root behavior might be expected to tend towards that of categories, and morphological alternations could be justified as providing

¹²The (words of the) DP-theory could probably be made even more redundant by inserting null determiners for generics, so that there wouldn’t even be the remaining noun headed DPs.

requisite informativeness.¹³ This section, however, suggests that stems are just as informative as roots, and thus, for the moment, fails to provide a functionalist explanation for such variation.

The data in this section came from the ACQUAINT corpus of New York Times articles analyzed with MINIPAR, a state-of-the-art dependency parser. Rather than postulating constituents as in the PTB data above, dependency parses directly represent exactly the ‘head-of’ relation which we have been studying. This data also provides root information. The data thus indicates the category, root, and head for every word of every sentence of the corpus. Please see appendix A.3 for information on the samples sizes.

The data of this section thus provides an independent head analysis which generally patterns as the data of the last section. It further provides morphological root information for analysis, but the root pattern entirely parallels the word pattern.

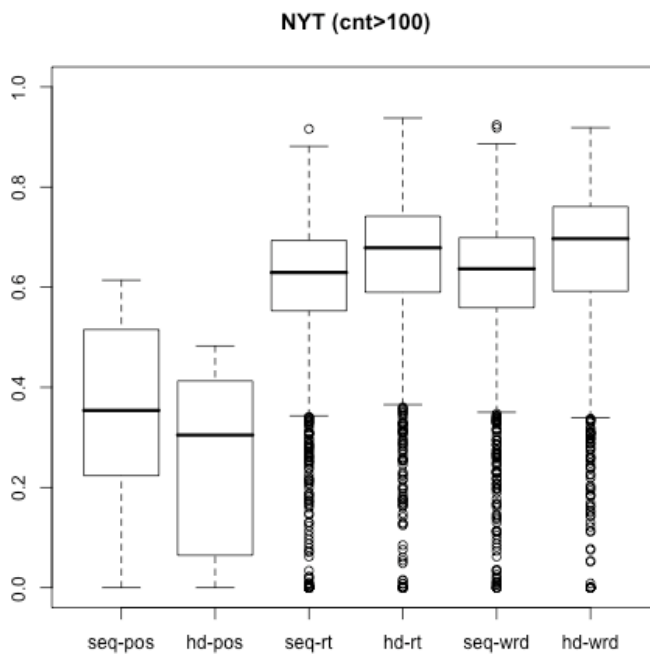


Figure 7: Roots pattern like words.

The following T-test suggests, as above, that syntactic word relationships are more complex than sequential word relationships.

words	seq	syn
mean rH	.61	.66

$t = -7.8991, df = 2939.606, p\text{-value} = 3.935e-15$

¹³However, testing the efficiency hypothesis would also seem to demand examining how much various morphological markers are constrained by context. I hope to more precisely examine this question in future research.

Root relationships are also significant in the same direction:

roots	seq	syn
mean rH	.60	.64

t = -7.1692, df = 2751.512, p-value = 9.649e-13

As with our previous findings, POS categories pattern in an opposite direction:

POS	seq	syn
mean rH	.36	.24

t = 2.1013, df = 39.925, p-value = 0.04197

Thus, this new corpora, with specific head annotations, further substantiates our original findings for the PTB. It also provides additional data demonstrating a similar pattern of complexity for morphological roots. This finding is particularly interesting insofar as morphological variation might have been explained as a kind of requisite diversification of overly redundant items. Since the roots appear to be as informative as the words, that explanation is not well-motivated by the data. If such variation is really *not* motivated by information requirements, they might even be interpreted as arbitrary, architectural manifestations lacking functional evolutionary explanations (cf. Pinker and Bloom (1990)’s interpretation of Chomsky’s attitude towards evolution). However, more precise analysis of morphological variation and its contexts will certainly be required.

3.5 Chinese

Surely one should not hastily generalize conclusions about “language” on the basis of a single example, so this section provides additional evidence from 2 additional corpora available for Chinese. In a certain sense, Chinese could even be expected to be comparable to the root behavior demonstrated above. Insofar as Chinese eschews agreement processes (“you walk” vs. “he walks”) and plural marking, and markers of aspect (e.g. 了 /le/) and nominalization (e.g. 者 /zhe/) are analyzed as separate words, Chinese can be thought to display less morphological variation. Though such an analysis is probably an oversimplification (especially insofar as it relies on the unreliable intuition of a ‘word’), we should still inquire whether Chinese can bare out the identical patterning of the English roots and words, as shown above.

In fact, the pattern for Chinese looks approximately the same (see figure 8), though the word differences are no longer significant (p(t)=.56 (i.e. >.05) with the word means in the T-test even patterning opposite the medians displayed in the graph). Thus, though the syntactic word relations aren’t necessarily more complex than their sequences, they at least maintain roughly equivalent complexity derived from simpler abstract hypotheses.

POS	seq	syn
mean rH	.57	.49

t = 4.9225, df = 280.103, p-value = 1.461e-06

words	seq	syn
mean rH	.74	.73

t = 0.5812, df = 758.523, p-value = 0.5613

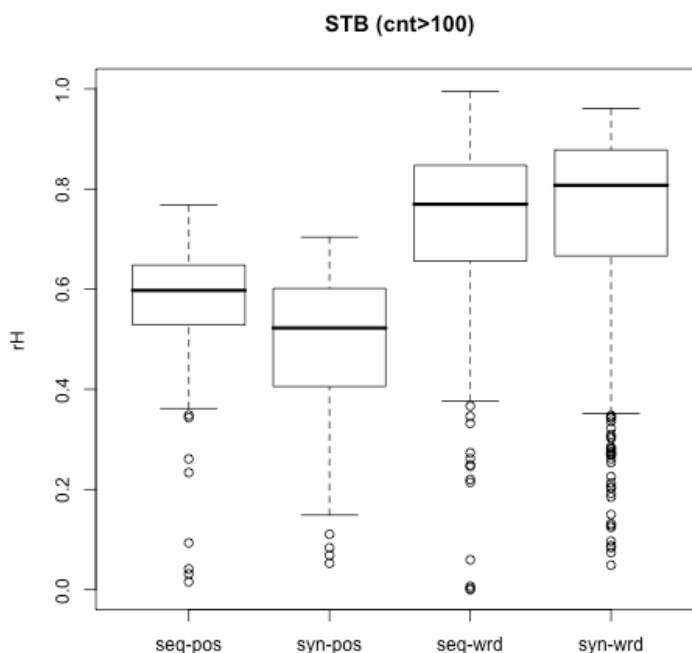


Figure 8: Syntax promotes more predictive categories without information loss.

The corpora used in the above analysis came from Taiwan’s (Academia) Sinica Treebank, which was manually annotated by humans for both constituent and head information. Thus, it provides yet a third independent – and *essential human* – analysis of head relations approximating the results established in preceding sections.

A bilingual corpus can provide an even tighter comparison of Chinese and English. The researchers at ISI have a file that they call the GoldGoldGold file because it is part of a manually translated, aligned, and parsed set of Chinese-English files in the Penn format. It is a “subset of the intersection” of LDC2006E86, LDC2006E93, and LDC2007T02.

Please see appendix A.3 for sample sizes of all these corpora.

These head relationships were again extracted with the ASCII method applied to the PTB above. The results shown in figure 9 display again the same relative pattern. Remarkably, as shown by the following T-test tables, they agree with the human annotations of the Sinica Treebank in finding no significant difference between Chinese word sequences and structures ($rH \approx .7$, $p(t) = .4799$). On the other hand, the English words, as above, display more redundancy in word sequences. Most importantly, in all cases, structures simplify categorical predictions without undermining instance (i.e. word) relation complexity.¹⁴

¹⁴The major observable difference, between these 2 Chinese corpora, lies in the Sinica tags also appearing generally more complex (mean & median above .5), where Penn tags seem simpler (mean & median below .5). The reader should be aware that the Sinica Treebank uses about 200 tags, where Penn uses only about 50. This difference again raises the question if our normalizing denominator is fairly compensating for such differences. Notably, the difference

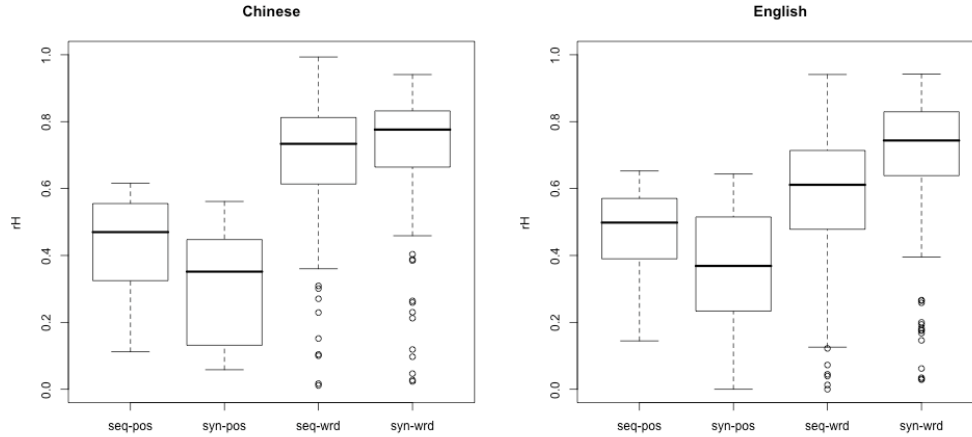


Figure 9: (GGG) Chinese and English pattern pretty much the same.

CHINESE:

POS	seq	syn
mean rH	.43	.31

$t = 2.8365, df = 48.779, p\text{-value} = 0.006624$

words	seq	syn
mean rH	.69	.71

$t = -0.7076, df = 217.887, p\text{-value} = 0.4799$

ENGLISH:

POS	seq	syn
mean rH	.46	.35

$t = 3.0284, df = 77.883, p\text{-value} = 0.003333$

words	seq	syn
mean rH	.57	.69

$t = -4.7, df = 275.942, p\text{-value} = 4.111e-06$

Thus, like the English morphological roots, Chinese words sustain the same pattern as English words, while also displaying the greater predictability of POS categories in syntactic relations. This section has therefore provided additional evidence of opposing patterning of words and categories for entirely new annotations in a completely unrelated language.

implies that the Sinica stats are divided by larger denominators, and thus their larger magnitude is all the more surprising.

4 natural

The above sections have contrasted word and category sequences and structures, but all these relationships have been assumed to be properties or hypotheses of properties of natural language. This section now sets out to show that there are functioning languages which *do not* sustain the high entropy of naturally evolving language.

Contemporary computer languages seem likely candidates to be currently less evolved and more redundant than human languages. This section uses C code utilized in MySQL as an example. Specifically, 83430 tokens of 1585 types were extracted from 19 files in the ‘btree’ folder of the macports distribution of MySQL-5.0.75. This file can typically be found, on a Mac, when using a macports installation of MySQL-5.0.75 at:

```
(9) /opt/local/var/macports/build/  
_opt_local_var_macports_sources_rsync.macports.org_release_ports_databases_mysql5/  
work/mysql-5.0.75/bdb/btree/
```

The above numbers resulted from treating punctuation as separate words, as do all the natural language corpora worked with in this paper. Unseparated data sustains the same conclusions, and can be provided if the reader so desires. The entropy measured in this section is strictly sequential.

For counts over 100, just 83 types covered 68947 tokens. Given this small set, the smaller sample of English from the GoldGoldGold file was adopted for comparison.

Notice that we might almost immediately conclude that the entire English corpus is less redundant in so far as it originally consisted of a rough average of 13.5 tokens/type (113986/8440, see appendix A.3), where the C code has 52.6 (83430/1585). For our well-represented items, the English is also less redundant since, for these items, there is a rough average of 472.9 tokens/type (65727/139), where the C code has 830.7 (68947/83). Crucially, the main concern of this research is word *relationships*, their conditional dependencies, their contextual predictability, and not simple frequency. Such raw frequency should, however, be expected to bear upon our conditional relationships. As expected, the naturally evolving collocations of English display much less entropic redundancy than C code, as shown in figures 10-11.¹⁵

A (debatable) T-test agrees with this conclusion:

	C	Eng
mean rH	.2892	.6536

t = -29.9372, df = 686.07, p-value < 2.2e-16

So, if the reader has accepted rH as a reasonable measure of (anti-)redundancy, then the evidence supports the claim that the sequential relationships of English words, within these corpora, are generally less redundant than those of the C code. Thus, there are useful, but less evolved, languages which display more redundancy than that of English. Or to be more precise, there is at least one example of one language that manifests more redundancy, and so it would be indisputably incorrect to think that the behavior demonstrated by this statistic in English was a mere mathematical artifact which any string of symbols would possess.

¹⁵A clever reader might suggest that higher frequencies for the C types lead to larger denominators in the entropy ratio, thus dragging down the ratio for more frequent items. I would reassure this reader that the comparisons are sustained by raw entropy calculations, as shown in appendix A.8.

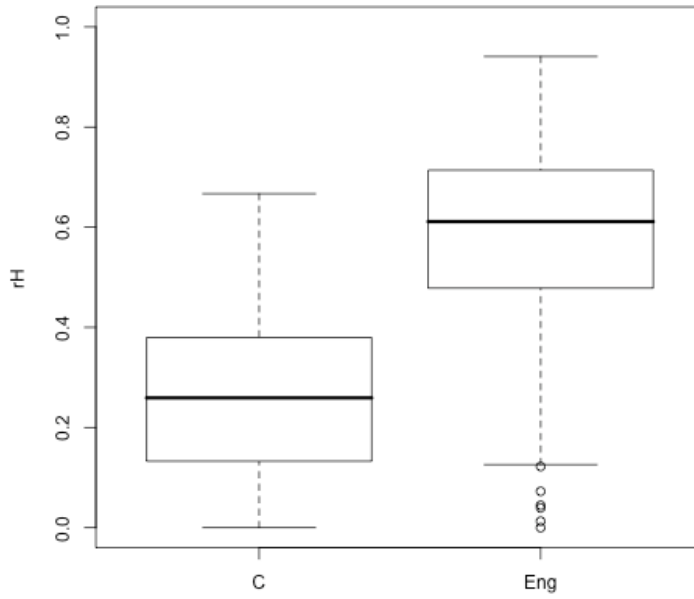


Figure 10: C is more redundant than English

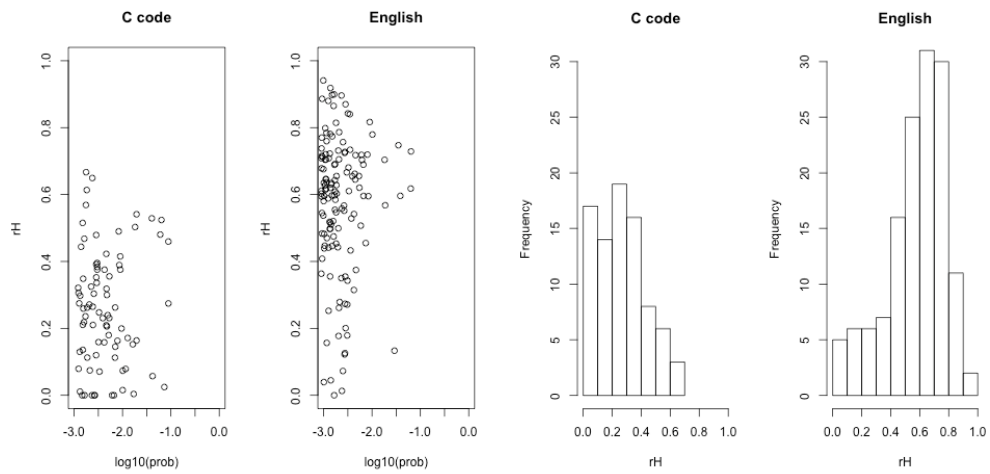


Figure 11: pH by frequency, and pH histograms

A language may have less chance to evolve because of youth or because it is provided less opportunities for change. Both might be derived from having less generations from some initial arbitrary state, either because it has only been a short time since the initial state, or because a single generation may last quite a long time. For a computer language, it is rather easy to imagine fixed distinctive generations of a language, but human language is rather much more like some open source composite with new functions constantly becoming available and old ones either being discontinued or remaining available though rarely used. Thus, where the presumed absence of acquired traits in biological evolution focuses selectional pressure specifically on generational succession, cultural evolution is yet more complex and adaptable.

5 to be false

When one observes *stasis*, one has the opportunity to ask *what forces are in equilibrium* (or there are no forces at work, usually a less likely possibility). Since with language it can be difficult to know if we are in stasis or not, we are less certain if forces are actually at balance, but here, where items above a certain frequency threshold all fall within limited entropy ranges, there is certainly cause to speak of *neither too much nor too little repetition*.

The scatterplots previously presented have all sustained the impression that both redundancy and anti-redundancy are increasingly limited with increasing frequency. The hypothesis here has associated anti-redundancy, that is unpredictability, with efficiency, with informativeness, with more distinctive meaning contributions. Redundancy has alternatively been associated with robustness and learnability. Thus, though “entropy is maximized”, there are opposing forces working to contain it.¹⁶ The scatterplots have suggested that words are constrained by upper and lower limits that increasingly narrow with frequency.

Figure 12 shows how the boundaries of observed values can be explicitly represented for English¹⁷ (excluding a single exception, to be explained below). Figure 13 shows that the limits evidenced by the Chinese data are quite similar, though not exactly the same.¹⁸ The 3 tables in figure 14 show that the Chinese and English limits aren’t so bad for each other, but the Chinese limits are even worse for the C code than the English.

Statistical populations like these deserve *graded* attributions of properties. To be completely accurate, it is necessary to say that the PTB has only violated these limits in 1/39352 of its types.

¹⁶G.K. Zipf also saw his results in terms of an evolutionary negotiation between ‘forces of unification’ and ‘forces of diversification’. He has been unjustly dismissed for having a supposedly over-general application. Appendix A.11 shows how this generality has been overestimated.

¹⁷The fact that these limits display opposing slopes also suggests that their intersection should provide a limit on raw word frequency. However, for English, this limit reaches the trivial and unsurprising result of 100%, as indicated by their meeting on the y-intercept (of a log-plot), at .7.

¹⁸The parameters for the English and Chinese limiting lines are compared in the following table:

	slope	y-intercept
Eng. Red. Req.	-.1	.7
Eng. Anti-Red. Req.	.25	.7
Ch. Red. Req.	-.08	.8
Ch. Anti-Red. Req.	.4	1.1

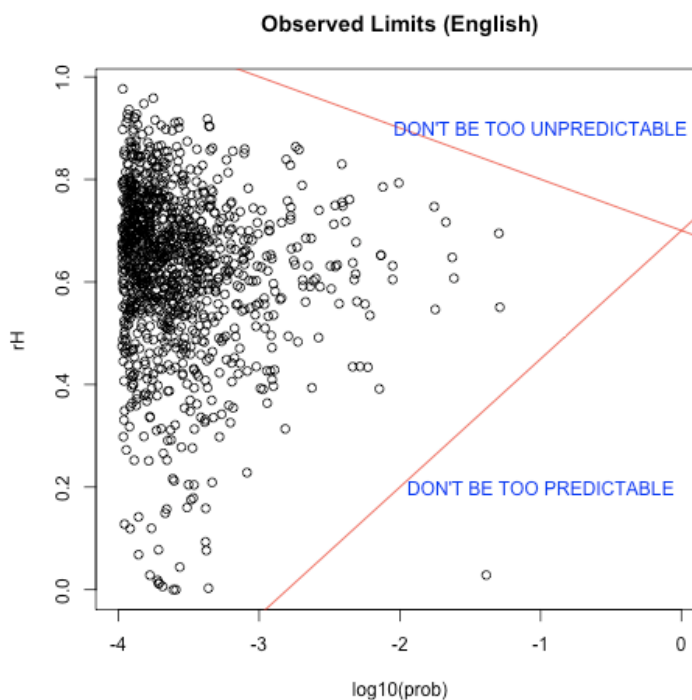


Figure 12: Limits on redundancy

On the other hand, since 20 out of 83 types from the C code fall below the bottom line, as shown in figure 15, we should suggest that C violates natural limits for 24% of its (observed) types.

Restricting our attention to even higher frequency requirements would indicate a larger falsification. The C items that appear at least $\log_{10}(\text{freq}) > -2$ still account for 59% of the token events (49379 out of 83430).¹⁹ Only 6 of these 16 items fall within the English limits indicating 10/16=62.5% violation of the limits.

A curious reader might also appreciate knowing that the 4 most frequent C items, above .4 pH, are ‘(’, ‘;’, ‘=’, ‘>’, in that order of frequency. It hardly seems surprising that their flexibility in C rivals that in natural language texts. Conversely, the single outlying point of English, at a frequency of roughly .3, and pH of about .13 (appendix 3 shows the top ten most frequent words of English and Chinese along with their sequential and structural relative entropies) represents a period. Sentence-final periods appear with such ridiculously low entropy because the data extraction method used here attached a special ‘@end@’ symbol to every sentence. Thus, periods were inordinately followed by ‘@end@’. This method was used to guarantee that every word was counted in every position it occurred, whether followed by other words or sentence-final. A period symbol itself is hardly a part of natural speech, and its relationship with this extraction artifact, which provided it with so much redundancy, should even less be considered a part of natural language. In that sense, it provides a good example of exactly the kind of symbolic redundancy which a naturally evolving communicative system will not engage in.

¹⁹Which again raises the point that token – rather than type-based – populations would better represent real-time entropy exposure.

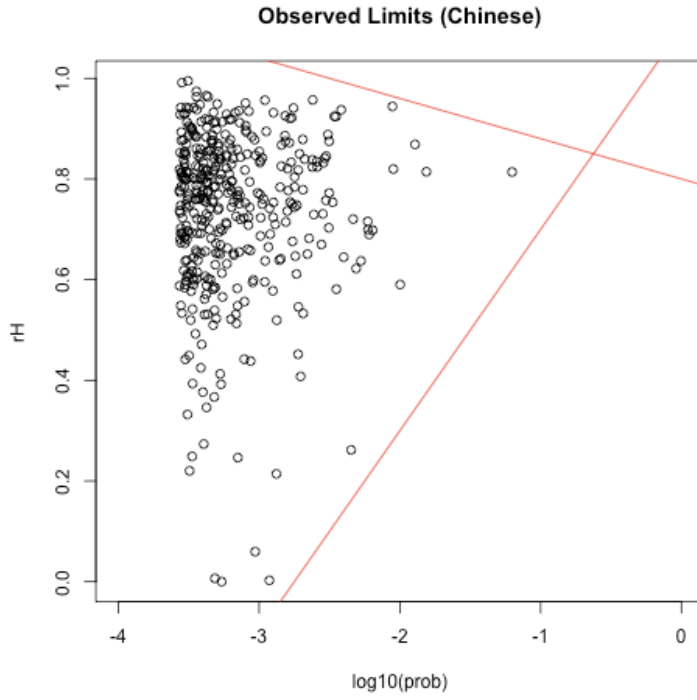


Figure 13: Chinese intersects a tad higher.

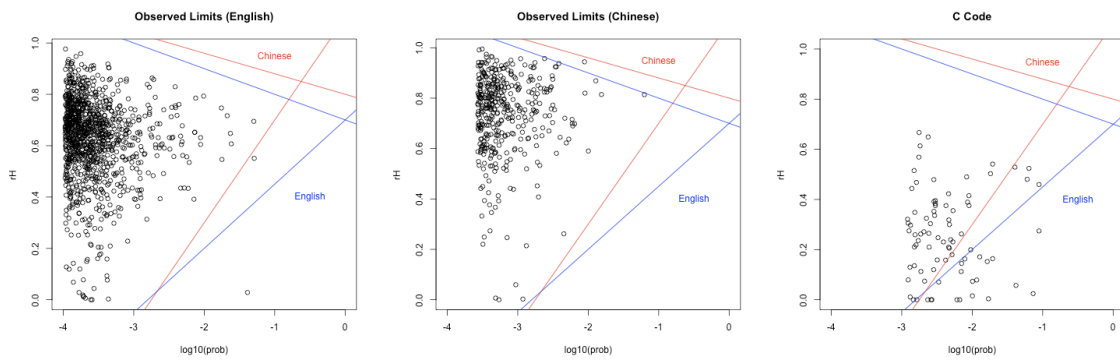


Figure 14: Chinese and English aren't as far off as C is.

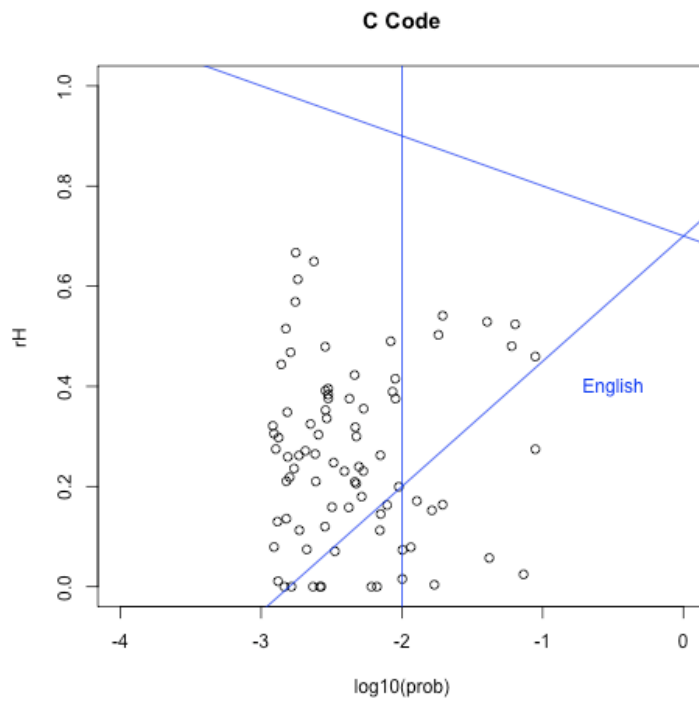


Figure 15: C has many exceptions.

At least since Karl Popper, there has been a philosophy of science which requires *falsifiability through negative predictions*.²⁰ Negative predictions can be interpreted as categorical impossibilities, but this section has attempted to show how they can also be deployed as gradable attributions of properties.²¹

6 outside evidence and new proposals

Recent psycholinguistic evidence is already demonstrating the psychological significance of (gradable) contextual predictiveness. At least since Solomon and Howes (1951), word frequency has been recognized as a crucial variable which any stimulus-response experiment must control for. However, new research has begun to suggest that (item-specific) contextual relations may be even more significant.

Levy and Jaeger (2007) demonstrate that optional complementizers introducing relative clauses (“the student (who/that) the teacher saw”) are predictable from “information optimization”. The complementizers are more likely where the relative clauses are less predictable and less likely where the clauses are more likely. Their calculation of “information” apparently doesn’t sum over multiple possibilities; it is simply the log inverse of a single probability conditioned on a comprehensive Bayesian model. Still, it is essentially the same concept of hypothesizing balanced management of context predictiveness.²²

Data collected from the CHILDES database (<http://childes.psy.cmu.edu/>; the data summaries are available at <http://www-scf.usc.edu/~potamite/dev.pdf>) also finds simple Bayesian bigram diversity to be *one of the best* indicators of child development. The data collected there gets rather patchy for ages over 30 months, but up to that point, there seem to be clear developmental trends for utterance length, vocabulary, and type/token (type/token is a better estimate of the *breadth* of one’s vocabulary than the raw type count which is heavily document size dependent). However, concluding that this data provides clear evidence for obvious conclusions is hampered by the fact that the adult interlocutors in the data (usually the mother) almost invariably display *identical developmental trends*. Though the adults typically maintain higher absolute length, vocab, and vocab variation, their apparently parallel development remains unaccounted for. Such evidence could suggest a hypothesis of *dynamic motherese*, but it also raises the possibility that both trends are confounded artifacts of the data. Crucially, the one variable which remains roughly constant for the adults, at about $p=.4$, while the kids veer from $p=.7$ onto $p=.4$ is *the average conditional probability of the next word*. Thus, contextual diversity, as measured by (lower) conditional probabilities, is a good indicator of children’s linguistic development. If one interprets entropy as a (log) inverse of (this single) probability (as Levy and Jaeger (2007) do, though this is being a bit liberal with the definition), then it would be fair to say that child language development from about 10 to 30 months is typically an act of *maximizing (conditional) entropy* (i.e. decreasing conditional

²⁰Note that such a philosophy in no way guarantees an escape from *explanation as abduction*: one’s theory is assumed safe as long as experimental outcomes are *not inconsistent* with it.

²¹This effort is reminiscent of the project of “fuzzy logic”. The main difference between probability theory and fuzzy logic can be exemplified by understanding that having a 50% chance of being (100%) tall is different than being 50% tall (perhaps in 100% of the cases).

²²They also examine “surface” and “structural conditioning”, finding both independently significant and concluding “they evidently exhibit enough differences to contribute non-redundant information in the ... model. We interpret this as evidence that speakers may be using both surface and structural cues for phrasal predictability estimation in utterance structuring” (p. 7).

probabilities). That alone should be sufficient to prove the general hypothesis here that there are observable tendencies to maximize entropy in the domains of language and evolution.

But furthermore, when one hypothesizes a tendency to maximize entropy as evolutionary selection for efficiency, one might also like to see evidence of compression acting to resolve potential inefficiencies. That is to say, nearing acceptable limits on redundancy, one should see variation arising to avoid redundancy, and if these variations are more efficient then they should be shorter. *Auxiliary contraction*, in fact, seems to be exactly such a phenomenon, which furthermore provides evidence of redundancy effects *at a categorical level*. Auxiliaries can morphologically merge with both subjects and negation (“she’s smart”; “he isn’t dumb”). Krug (1998) offers a “frequentist” explanation, but really seems to primarily show that *among items that contract, the more frequent contract more*. In fact, if one simply looks at frequent bigrams, there are many items which appear together more frequently than auxiliaries with ‘not’, without obvious coalescence (e.g. “of the”, “in the”, etc.). One can also look at a statistic such as mutual information or even the entropy statistic used here: individual auxiliaries do not cluster at any particular extreme. However, auxiliaries *as a category* do rank as *one of the most redundant categories*, for both sequential and structural relations, as shown in appendix A.9. However, to develop this hypothesis seriously, we need better quantitative measures of coalescence across all words and categories.

Other researchers have been estimating “predictability” from cloze tests (Miellet et al. (2007)), or merely from document counts (“contextual diversity”, Adelman et al. (2006)). Bien et al. (2005) use the exact same entropy calculation as here, but without proceeding to its *relative* value. That is one possible reason why they, like the other research mentioned in this paragraph, find *faster response times for higher entropy items*. Both Adelman et al. (2006) and Bien et al. (2005) argue that their context predictability factors may account for more response time variation than frequency, but their statistic is seriously confounded with frequency. Their results do point to an exciting role of variation in reinforcement, but they are at odds with the notion of lower entropy as more simple and certain and easier. Hopefully, considering the relative proportion of an item’s entropy would be able to reaffirm that the more predictable should be easier to process.

In fact, since the research above calculates response times off the item *whose entropy is measured*, it also makes more sense that higher entropy would facilitate response times. An item that goes with a lot of contexts is very useful (and will almost certainly be frequent). Higher entropy may be *predictable*, but it is not *predictive*. If we want to look for processing difficulties imposed by information complexity (i.e. high entropy that slows response times), we seemingly have to look at the context effects that *some variable context imposes on some other (controlled) target area*. Thus, the experimental variables proposed here must be interpreted *relationally* – as in priming experiments.

context: variable	⇒	target: controlled
high ent		medium freq¢
low ent		medium freq¢
RT?		

As with all priming experiments, the design should vary the properties of the context while controlling properties of the target word where the response time needs to be measured. Thus, it would measure the effect context has on response times.

The properties of the target word almost certainly provide yet another crucial variable, one immediately suggestive of “reanalysis theory”. The *surprisal* of a word should depend both on the expectations that its context sets up and how much that word disappoints or satisfies those

expectations. The conditional probability of the target should give us an estimate of the satisfaction, and the entropy of the context is a kind of estimate of the strength of expectations. It should be harder to recover from incorrect expectations where those expectations are stronger (at least we know that those expectations *matter* if they affect response). Low-expectation contexts (i.e. diverse, i.e. high entropy) should show less variation in response times over different outcome probabilities, where high expectation contexts would hopefully show greater differences in more or less likely outcomes. In essence, the more confident we are in our expectations, the more annoying should be their disappointment. Even if this “reanalysis hypothesis” fails, such experiments should be able to unveil effects of context on at least some of the various categories and relations encoded as variables.

Since Adelman et al. (2006) and Bien et al. (2005) both showed *faster* response times for higher entropy items (“less predictive”), it appears that it may be challenging to show that *predictable contexts facilitate processing*. Future research will hopefully be able to show specifically that *processing is facilitated by predictive structural categories*.

A simple experimental design examining response times, in say a lexical decision task, on pairs of randomly selected words should be able to demonstrate a number of the effects hypothesized here. For one, responses would hopefully replicate the effect mentioned above of entropy facilitation, as evidence of reinforcement through diversity. Conditional probabilities also seem fair to expect to affect response times. Reanalysis theory further predicts an interaction between the entropy of the first word and conditional probability of the second.²³ Categorical effects could presumably be sought in groups of words which would hopefully parallel entropy predictions. Studying structural effects would be somewhat more complicated, since our 1-dimensional perception of time means context is essentially experienced in sequential order. Note that the statistic collected here merely measured restrictiveness of selectors without concern for sequential order, and thus may not be a great estimate of real-time predictability, except where they correlate. Common sequences that lack direct structural relations, such as determiner - adjective sequences would be one place to look for a *weakening* of predictive effects.²⁴ Regrettably, the linear perception of time could make isolating structural from sequential effects tricky.

Making use of such structural hypotheses, however, seems essential to developing artificial agents with the capacity to learn language. Maximum likelihood modeling already adopts the method of *selecting hypotheses which make the data more likely*. Maximum entropy modeling assumes as little as possible, by adopting *the most uniform possible model given (observed) constraints*. The research here has suggested that attractive models have low entropy abstractions with high entropy instantiations. Presumably, when AI-systems are able to automatically acquire the languages of the world, from simple exposure, just as children do, they will also arrive at *abstract, compositional* hypotheses which provide more robust predictability without sacrificing information transfer.

Since conditional probabilities and entropy can be calculated for any distribution of whatever

²³Conditional probability effects are likely easier to observe than entropy effects since they are more specific to the exact stimuli presented. While it would be nice to show the above mentioned interaction of conditional probability and entropy, many specific probabilities may be more significant than *potential entropy*. That would however not take away from the general systemic distributions presented here. The effectiveness of context predictability would also still be presupposed, and would still need to be sought by better estimators. Experiments like these would not necessarily be experiments of *falsification* to the larger hypotheses which motivate this paper. Rather, failure to find correlations would incriminate either the specific relations or their observed estimates. The larger hypotheses of contextual effects would then be all the more in need of estimates of factors which did model observed reactions.

²⁴This suggestion assumes a typical contemporary analysis of noun phrases (DT (ADJ NOUN)), which is *not* how the PTB is structured. The NYT corpus would perhaps then be more relevant.

relation, the variables they make available for experimental comparison are theoretically endless. The data presented so far has essentially assumed that it is important to compare the structural and sequential effects of words and categories.

7 conclusion

Since Boltzmann, entropy has been understood as the tendency of systems to disorder. Macroscopic uniformity arises because of underlying chaos. Energy differentials are minimized because such differentials are compatible with fewer random possibilities (and randomness – ‘all else held equal’ – always increases).

Rod Swenson describes Boltzmann’s ideas as follows:

‘Because there are so many more possible disordered states than ordered ones, he concluded, a system will almost always be found either in the state of maximum disorder or moving towards it. As a consequence, a dynamically ordered state, one with molecules moving “at the same speed and in the same direction,” Boltzmann (1974/1886, p. 20) asserted, is thus “the most improbable case conceivable...an infinitely improbable configuration of energy.”’

(<http://www.entropylaw.com/entropydisorder.html>)

New systems theory, by Rod Swenson in particular, is replacing Boltzmann’s interpretation of entropy as disorder, with an appreciation of how order arises from its maximization. Swenson claims that order arises according to a “Law of Maximum Entropy Production”, which he states as follows:

(10) “the Law of Maximum Entropy Production” (MEP):

“A system will select the path or assemblage of paths out of available paths that minimizes the potential or maximizes the entropy at the fastest rate given the constraints.”

He suggests the reader imagine a warm cabin in a winter woods maximizing entropy as it dissipates heat into the surrounding environment. Now imagine opening a window in this cabin and imagine the focused flow that arises as the heat is able to escape at this much faster rate (Swenson and Turvey (1991)). Alternatively, imagine how a crack in a dyke will concentrate flow towards the crack leading to further cracking and further flow. Directed focused motion arises from maximizing the rate at which these given differentials are minimized.

Swenson also likes to cite the classic example of “Bénard cells”, first presented by Henri Bénard in 1900. The so-called cells arise in a layer of liquid such as water when heated from below. The pictures in figures 16 and 17 originally appeared in Swenson (1989a) and Swenson (1989b) respectively, though they can be found in many of his articles. Figure 16 shows a typical example of these “cells” arising, and figure 17 shows how such ordering transformations cause sudden jumps in the rate of heat transport.

Swenson’s major thesis then is that the appearance of such predictable regularities at the macroscopic level are not contrary to the law of entropy but, in fact, arise exactly in so far as they facilitate it.

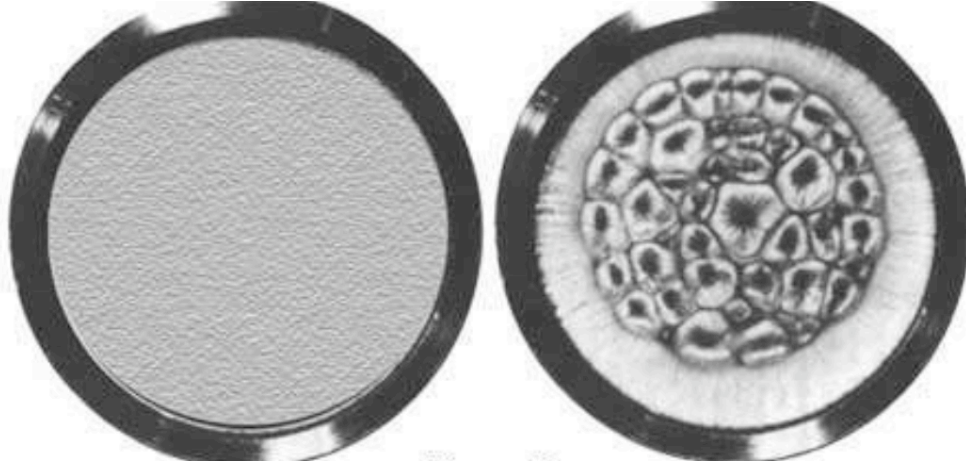


Figure 16: Typical Bénard cells

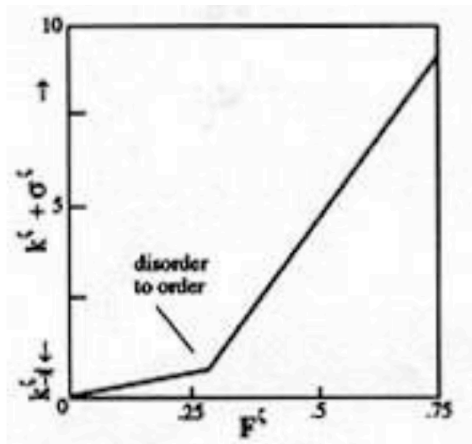


Figure 17: Rate of heat transport

Thus, syntax is like that crack in the dyke, or that open window. It is like the Bénard cells. It is a conventionalized set of predictable protocols that provide for a higher rate of information-entropy transfer. It has been asked why learners should converge on a hierarchal constituent system of language, rather than say a sequence-based system. The answer put forth here is because, in comparing alternatives, evolution selects for lower entropy macroscopic states without sacrificing microscopic complexity. The part-of-speech abstractions act as macroscopic states, while their word instantiations are like the microscopic states. The conventionalized configurational states of syntactic abstractions are like the symmetrically ordered cells, while the words are the faster heat flowing through them.

Thus, rather than seeking categorical, deterministic, universal laws of language, this research has suggested that language has universal laws of a dynamic, fluid, and gradable nature.

Theoretically, entropy management seems to be evidence of instinctive selection between ef-

efficiency and robustness. Practically speaking, consciously manipulating entropy exposure should help us control experiments better and advance more effective learning.

So both in the way that entropy is information, and in the way that information organizes abstract regularities to the advantage of efficient complexity, the useful compatibility of entropy with evolution and culture is manifested.

A additional arguments and supplementary statistics

A.1 mathematical motivations for Gibbs-Shannon entropy

Shannon proposes his formula for predictability because it is the “only H” satisfying 3 intuitively desirable conditions (see *his* Appendix 2 for the proof):

(11) Shannon (1948), p. 10:

- a. “ H should be continuous in the p_i .”
- b. “If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n . With equally likely events there is more choice, or uncertainty, when there are more possible events.”
- c. “If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.”

Shannon’s example of the third property is provided below:

In so far as the predictability of the distributions represented by the following trees are the same, we should require that $H(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}) = H(\frac{1}{2}, \frac{1}{2}) + \frac{1}{2}H(\frac{2}{3}, \frac{1}{3})$:



Note that the following is a special case of the above requirement:

$$(12) \quad p(X) * p(Y) = p(Z) \rightarrow H(X) + H(Y) = H(Z)$$

Some simple examples are also provided below. Assuming discrete situations resulting in discrete outcomes makes low frequency entropies easy to enumerate. The square brackets refer to a situation listing the counts (c) of different outcomes ($p_i = \frac{c_i}{\sum_i c_i}$; in other words, the notation adopts a minor reinterpretation for the arguments of H, as counts instead of probabilities):

- (13) a. $H([1])=0$
- b. $H([2])=0, H([1,1])=1$
- c. $H([3])=0, H([2,1])=0.9182958, H([1,1,1])=1.584963$
- d. $H([4])=0, H([3,1])=0.8112781, H([2,2])=1, H([2,1,1])=1.5, H([1,1,1,1])=2$
- e. $H([5])=0, H([4,1])=0.6500224, H([3,2])=0.9709506, H([3,1,1])=1.370951, H([2,1,1,1])=1.921928, H([1,1,1,1,1])=2.321928$

A.2 data reductions

These statistics refer to the Penn Treebank (PTB). Figure 18 shows how mean rH is affected as words with insufficient counts are dropped from the analysis. Figure 19 shows that requiring counts of at least 100 removes the irregularity and cut-off appearance of the histograms which include lower count items. Presumably such higher-count statistics are less deformed by underrepresentation.

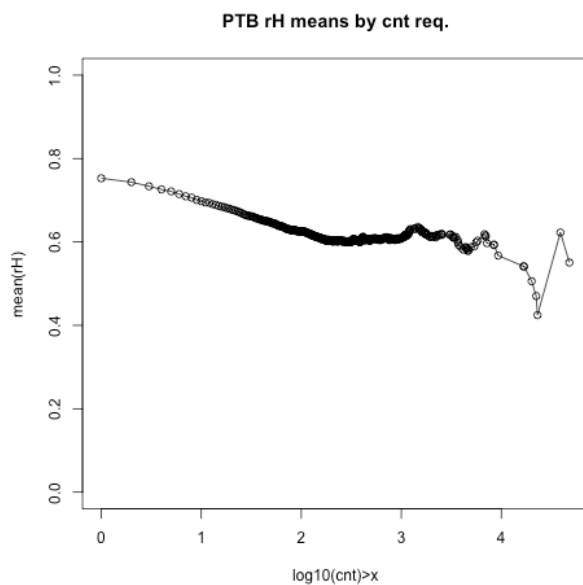


Figure 18: The effect of excluding low count data on mean rH

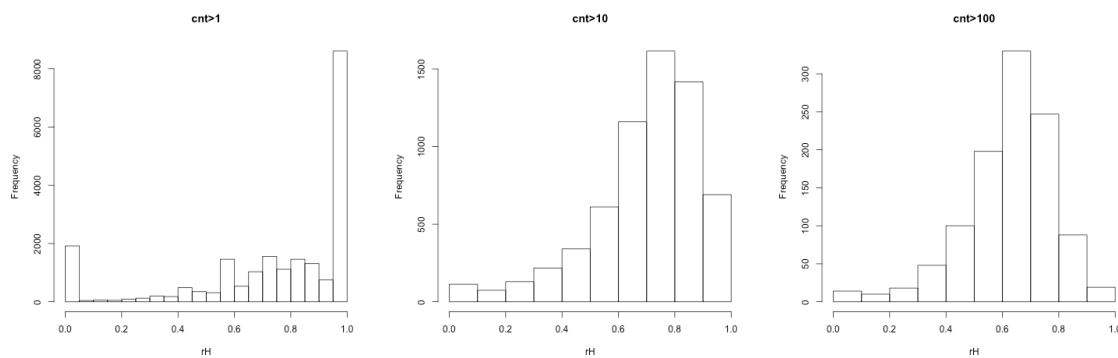


Figure 19: Higher count requirements provide more well-rounded histograms.

A.3 data sizes

A.3.1 PTB

39803 sentences of English.

cnt>100 words	tokens	types	cnt>100 POS	tokens	types
next	710009	1072	next	948897	41
sele	774029	1080	sele	1012759	54

all words	tokens	types	all POS	tokens	types
next	949088	39352	next	949088	44
sele	1013145	39361	sele	1013145	70

A.3.2 NYT

NYT automatic, dependency corpus: 76180 sentences.

cnt>100 words	tokens	types	cnt>100 ROOT	tokens	types
next	1075533	1474	next	1076453	1377
sele	1075533	1474	sele	1076453	1377

	cnt>100 POS	tokens	types
	next	1400108	21
	sele	1400108	21

all words	tokens	types	all ROOT	tokens	types	all POS	tokens	types
next	1400174	45523	next	1400174	62847	next	1400174	24
sele	1400174	45523	sele	1400174	62847	sele	1400174	24

A.3.3 STB

61793 ‘lines’ of Chinese.

cnt>100 words	tokens	types	cnt>100 POS	tokens	types
next	171346	402	next	369559	139
sele	171330	402	sele	369570	146

all words	tokens	types	all POS	tokens	types
next	371126	42819	next	371125	196
sele	371093	42816	sele	371093	208

A.3.4 GGG

3401 sentences of Chinese, and 3342 lines of English (since apparently titles were translated “NULL”, though bylines and other semi-grammatical constructions were included).

	all words	tokens	types	all POS	tokens	types
ENGLISH:	next	113986	8440	next	113987	44
	sele	113986	8440	sele	113987	63

	cnt>100 words	tokens	types	cnt>100 POS	tokens	types
ENGLISH:	next	65727	139	next	113663	36
	sele	65727	139	sele	113497	44
	all words	tokens	types	all POS	tokens	types
CHINESE:	next	84550	9595	next	84550	31
	sele	84543	9595	sele	84543	50
	cnt>100 words	tokens	types	cnt>100 POS	tokens	types
CHINESE:	next	39021	110	next	84249	23
	sele	39017	110	sele	83887	28

A.3.5 C (mysql)

	tokens	types
all	83430	1585
cnt>100 words	68947	83

A.4 sample extraction

Table 1 shows an example of both noun and determiner oriented syntactic extractions for the following sentence:

- (14) ((S(NP-SBJ (DT That))(VP (MD could)(VP (VB cost)(NP (PRP him))(NP (DT the) (NN chance)(S(NP-SBJ (-NONE- *))(VP (TO to)(VP(VP (VB influence)(NP (DT the) (NN outcome))))(CC and)(VP(ADVP (RB perhaps))(VB join)(NP (DT the) (VBG winning) (NN bidder))))))))))(. .))

A.5 DP recategorization

The full set of tag changes are listed below. The asterisk is intended as a ‘kleene star’, including (and retaining) any characters suffixed where it stand below.

- (15) a. NP⇒DP
b. CLP⇒QF
c. PRP@usd@⇒DPS
d. ADJP⇒JP
e. PRP*⇒Dpro*
f. POS*⇒DPOS*
g. NN*⇒EN*
h. IN⇒PH
i. TO⇒PH, *if dominated by PP*²⁵

²⁵The PTB bizarrely distinguishes ‘to’ from other prepositions, *without* distinguishing whether it is appearing in a PP or a VP.

Table 1: Head extractions; selector indicated by h; interlaced POS and word extractions.

("noun phrases")		("determiner phrases")	
the	h:outcome	h:the	outcome
DT	h:NN	h:DT	EN
the	h:bidder	h:the	winning
DT	h:NN	h:DT	VBG
winning	h:bidder	h:the	bidder
VBG	h:NN	h:DT	EN
h:influence	outcome	h:influence	the
h:VB	NN	h:VB	DT
perhaps	h:join	perhaps	h:join
RB	h:VB	RB	h:VB
h:join	bidder	h:join	the
h:VB	NN	h:VB	DT
h:influence	and	h:influence	and
h:VP	CC	h:VP	CC
h:influence	join	h:influence	join
h:VP	VB	h:VP	VB
to	h:influence	to	h:influence
TO	h:VP	TO	h:VP
@ast@	h:influence	@ast@	h:influence
-NONE-	h:VP	-NONE-	h:VP
the	h:chance	h:the	chance
DT	h:NN	h:DT	EN
h:chance	influence	h:the	influence
h:NN	VP	h:DT	VP
h:cost	him	h:cost	him
h:VB	PRP	h:VB	Dpro
h:cost	chance	h:cost	the
h:VB	NN	h:VB	DT
could	h:cost	could	h:cost
MD	h:VP	MD	h:VP
That	h:cost	That	h:cost
DT	h:VP	DT	h:VP
h:cost	.	h:cost	.
h:VP	.	h:VP	.
h:@ROOT@	cost	h:@ROOT@	cost
h:@ROOT@	VP	h:@ROOT@	VP

A.6 null items of the PTB

Table 2 shows the 10 silent items of the PTB, with their corresponding counts, and totals, included in the syntactic, but not in the sequential, analysis. The most common of them are the plain asterisk, which refers to silent non-finite verb subjects, ‘*t*’ which refers to movement traces, ‘0’, which refers to null complementizers, and ‘*u*’, which refers to unrealized units of measurement. Please refer to the literature on the PTB for a more complete description of all these items.

Table 2: Null items

word	head cnt
*	28242
t	15926
0	9997
u	7478
ich	998
exp	556
?	480
rnr	343
ppa	20
not	18
=10	=64057

A.7 PTB POS raw entropy

The raw entropy of the PTB’s POS tags also suggests that syntactic relations are easier to predict than sequential strings, as shown in figure 20.

A (questionable) T-test is in agreement with this conclusion:

	seq	syn
mean H	2.95	2.31

t = 2.8433, df = 81.939, p-value = 0.005634

A.8 C raw entropy

Figure 21 shows that C collocations present much lower raw entropy than English (of the GGG, >100, word sequences).

A (questionable) T-test is in agreement with this conclusion:

	C	Eng
mean H	2.25	4.57

t = -9.8331, df = 188.889, p-value < 2.2e-16

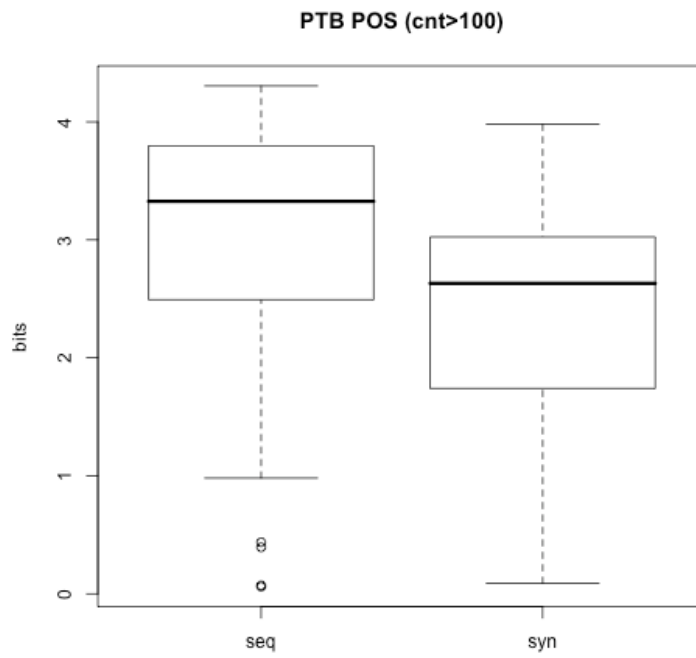


Figure 20: POS sequences are harder to predict.

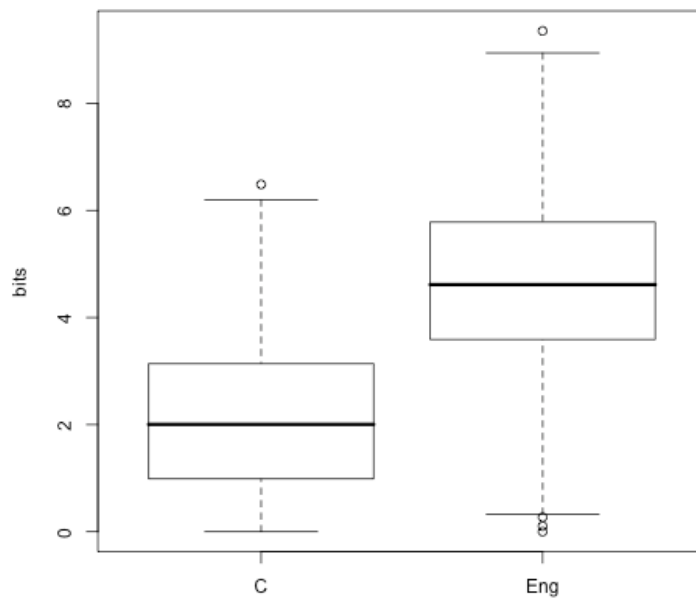


Figure 21: C sequences are easier to predict.

A.9 auxiliary redundancy

The following tables, of the 10 most redundant categories, show that auxiliaries, tagged “md”, in the PTB, take either first or second place for redundancy (while ‘pdt’ is a tag specific to preceding determiners, such as “all” in “all the tips” or “such” in “such a plan”). “md” is headed by “vp” 9708 times out of 9789 and followed by “vb” 7809 times out of 9793 times.

SEQ			SYN		
POS	count	rH	POS	count	rH
pdt	369	0.07223621	md	9789	0.01453077
md	9793	0.17961992	wdt	3921	0.03057409
rbs	450	0.28186068	to	13559	0.05735850
ex	863	0.37972883	wp@usd@	149	0.06083448
prp@usd@	8404	0.39993898	wp	2302	0.06472455
wp@usd@	168	0.41586265	wrb	2093	0.07036479
to	22339	0.43569998	whnp	434	0.10210527
dt	81756	0.45648856	prn	250	0.14485216
jj	61158	0.48318889	ex	863	0.18528823
pos	8700	0.48935646	pos	8649	0.21643449

A.10 top 10s

Table 3: 10 most frequent words of English and Chinese

wd	cnt	seq-rH	(cnt increasing downward)		wd	cnt	seq-rH	hd-rH
			hd-rH	wd				
for	8436	0.60486312	0.7505648	我	2192	0.7157656	0.7875979	
's	9322	0.79308094	0.6678321	他	2202	0.6994732	0.8073140	
and	16648	0.74685066	0.7883133	也	2245	0.6898114	0.7797553	
in	16919	0.54640114	0.7056652	不	2380	0.6983695	0.7276540	
a	20134	0.71676673	0.7071155	、	3287	0.9442705	0.8485279	
to	22334	0.64789417	0.6632975	有	3335	0.8198131	0.2613128	
of	22987	0.60718174	0.6996027	了	3730	0.5903567	0.7980473	
.	38991	0.02845874	0.5111096	在	4756	0.8688121	0.6664678	
the	47922	0.69478991	0.6863739	是	5723	0.8144799	0.2817151	
,	48678	0.55069506	0.6853835	的	23264	0.8139230	0.8260881	
all tokens=	949088			all tokens=	371126			

A.11 Zipf: whiter than noise.

George Miller, in his introduction to the last known (1965) printing of Zipf (1935), glibly associates Zipf’s famous law with typing monkeys and argues that it does not “distinguish rational from random behavior”, and that “Zipf was wrong”, and “others were right” that his pattern “represents some necessary consequence of the laws of probability” rather than reflecting “some universal property of the human mind” (though Zipf was very clear that he had no intention of limiting his theory to the mere human mind).

Li (1992) notes that Miller’s accusation comes with no proof, and thus sets out to fill the gap. He assumes random, independent generation of characters, with anyone of them privileged as a “word boundary”, basically as Miller seems to assume. Just these assumptions insure that longer strings are less probable by a factor of $V^{(n-m)}$ (the vocab to the power of the length difference). Various inequalities then suggest that this generates the Zipf property.

The following tables show that the accuracy of this proof has been overestimated. The first graph shows a classic Zipfian log-rank-frequency graph for the (roughly a million) words of the PTB. The second shows a sample of a million random words generated as described above.

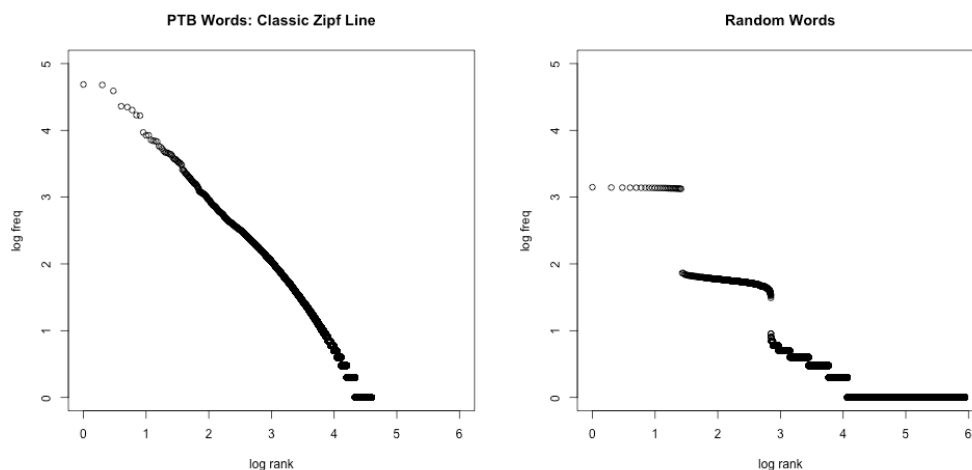


Figure 22: Random strings frequencies are more discontinuous.

The random output manifests blatant discontinuities lacking from the natural distribution. These discontinuities can be understood as arising directly from the roughly uniform sharing of probability at each length. Natural language is apparently subject to pressures which grade probability even within length classes. The following length table (up to just 100) for English would probably be more significant if replaced with syllable units, but the letter length makes it more comparable to the subsequent table which is provided to demonstrate that these noise sequences are adhering to Miller and Li’s assumptions. Those assumptions just don’t provide an equally smooth (log) rank-frequency table.

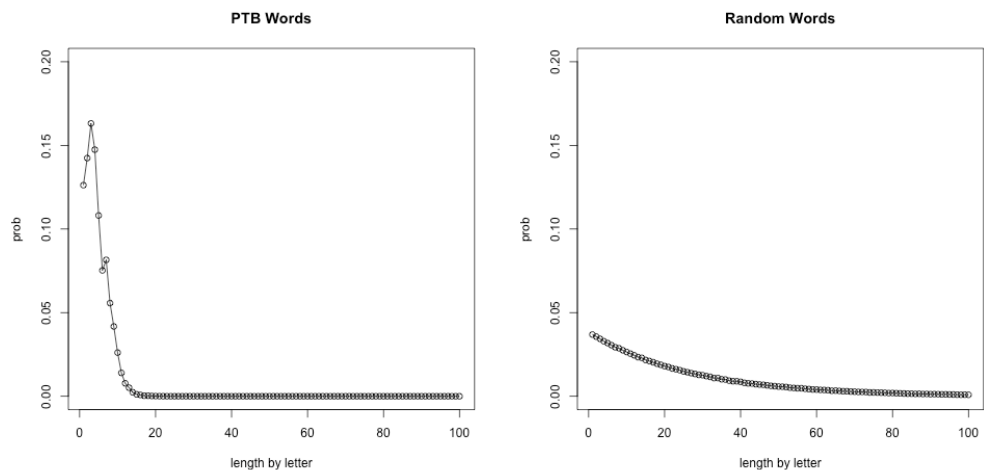


Figure 23: Length is balanced, though not identically.

However, it's worth noting that the more language approximates noise, the more it demonstrates that “entropy is maximized”. However, the noise manifests discontinuities because it uniformly distributes within (length) classes, where natural language distributes frequency even more smoothly.

Plus, since word parts are not independently selected, how could such an assumption be part of explaining such behavior?

Though somewhat tangential to the original observations of this paper, it would be this paper's greatest honor if it could but acquit its noble predecessor from the unfair accusations that have been leveled against him.

References

- Adelman, James S., Gordon D.A. Brown, and José F. Quesada. 2006. Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science* 17:814 – 823.
- Berger, Adam L., Stephen Della Pietra, and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22:39–71. URL citeseer.ist.psu.edu/berger96maximum.html.
- Bien, Heidrun, Willem J. M. Levelt, and R. Harald Baayen. 2005. Frequency effects in compound production. *PNAS* 102:17876–17881.
- Brent, Michael R. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Meeting of the Association for Computational Linguistics*, 209–214. URL citeseer.ist.psu.edu/brent91automatic.html.
- Brown, Peter F., John Cockey, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach

- to machine translation. *Computational Linguistics* 16:79–85. URL citeseer.ist.psu.edu/brown90statistical.html.
- Charniak, Eugene, Kevin Knight, and Kenji Yamada. 2003. Syntax-based language models for statistical machine translation. In *MT Summit IX. Intl. Assoc. for Machine Translation*.
- Cheng, Lisa Lai-Shen, and Rint Sybesma. 1999. Bare and not-so-bare nouns and the structure of NP. *Linguistic Inquiry* 30:509–542.
- Chiang, David, and Daniel M. Bikel. 2002. Recovering latent information in treebanks. In *In Proceedings of COLING 2002*, 183–189.
- Crain, Stephan, Takuya Goro, and Rosalind Thornton. 2006. Language acquisition is language change. *Journal of Psycholinguistic Research* 35.
- Ferreirra, Victor S. 1996. Is it better to give than to donate? syntactic flexibility in language production. *Journal of Memory and Language* 35:724–755.
- Frisch, Stefan. 1997. The change in negation in middle english. *Lingua* 101:21–64.
- Harner, Lorraine. 1981. Children talk about time and aspect of actions. *Child Development* 52:498–506.
- Hobbs, Jerry R. 2006. The origin and evolution of language: a plausible, strong-ai account. In *Action to language via the mirror neuron system*, ed. Michael A. Arbib, 48–88. Cambridge: Cambridge.
- Huang, C.-T. James, Y.-H. Audrey Li, and Yafei Li. 2009. *The syntax of Chinese*. Cambridge: University Press.
- Huang, Shi-Zhe. 2006. Property theory, adjectives, and modification in chinese. *Journal of East Asian Linguistics* 15:343–369.
- Hurford, James R. 1989. Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua* 77:187–222.
- Jaynes, E. T. 1957. Information theory and statistical mechanics. *The Physical Review* 106:620–30.
- Jaynes, E. T. 1965. Gibbs vs. Boltzmann entropies. *American Journal of Physics* 33:391–8.
- Jeffrey Podos, Stephen Nowicki, and Susan Peters. 1999. Permissiveness in the learning and development of song syntax in swamp sparrows. *Animal Behaviour* 58:93–103.
- Johnson, Mark. 1998. Pcfg models of linguistic tree representations. *Computational Linguistics* 24:613–632.
- Jurafsky, Dan, Christopher Manning, Pi-Chuan Chang, Michel Galley, Dan Cer, and Huihsin Tseng. 2007. Why is Chinese so hard to translate to english? In *Rosetta*.
- Kearns, Kate. 2007. Regional variation in the syntactic distribution of null finite complementizer. *Language Variation and Change* 19:295–336.

- Krug, Manfred. 1998. String Frequency: A Cognitive Motivating Factor in Coalescence, Language Processing, and Linguistic Change. *Journal of English Linguistics* 26:286–320. URL <http://eng.sagepub.com>.
- Lauer, Mark. 1995. Designing statistical language learners. Doctoral Dissertation, Macquarie University, Australia.
- Levy, Roger, and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems 19*, ed. B. Schölkopf, J. Platt, and T. Hoffman, 849–856. Cambridge, MA: MIT Press.
- Levy, Roger, and Christopher Manning. 2003. Is it harder to parse chinese, or the chinese tree-bank? In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 439–446. Morristown, NJ, USA: Association for Computational Linguistics.
- Li, Wentian. 1992. Random texts exhibit Zipf’s-law-like word frequency distribution. *IEEE Transactions on Information Theory* 38:1842–1845.
- Miellet, Sebastien, Laurent Sparrow, and Sara C. Sereno. 2007. Word frequency and predictability effects in reading French: An evaluation of the E-Z reader model. *Psychonomic Bulletin & Review* 14:762–769.
- Norris, Dennis. 2006. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychol Rev.* 113:327–57.
- Pinker, Steven. 1994. *The language instinct*. New York: William Morrow.
- Pinker, Steven, and Paul Bloom. 1990. Natural language and natural selection. *Behavioral and Brain Sciences* 13:707–784. URL <http://www.isrl.uiuc.edu/~amag/langev/paper/pinker90naturalLanguage.html>.
- Pullum, G.K., and B.C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19:9–50.
- Pustejovsky, James. 1991. The generative lexicon. *Computational Linguistics* 17:409–441.
- Ratnaparkhi, A. 1998. Maximum entropy models for natural language ambiguity resolution. URL citeseer.ist.psu.edu/ratnaparkhi98maximum.html, ph.D Thesis, University of Pennsylvania.
- Shannon, Claude E. 1948. *A mathematical theory of communication*. CSLI Publications. URL <http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html>.
- Smith, Madorah E. 1942. The english of Hawaiian children. *American Speech* 17:16–24.
- Solomon, Richard L., and Davis H. Howes. 1951. Word frequency, personal values, and visual duration thresholds. *Psychological Review* 58:256–270.
- Swenson, R. 1989a. Emergent attractors and the law of maximum entropy production: Foundations to a theory of general evolution. *Systems Research* 6:187–198.

- Swenson, R. 1989b. Engineering initial conditions in a self-producing environment. In *A delicate balance: Technics, culture, and consequences*, 68–73. Los Angeles: IEEE.
- Swenson, R., and M.T. Turvey. 1991. Thermodynamic reasons for perception-action cycles. *Ecological Psychology* 3:317–348.
- Yang, C.D. 1999. A selectionist theory of language development. In *Proceedings of 37th Meeting of the Association for Computational Linguistics*, 429–435. East Stroudsburg, PA: Association for Computational Linguistics. URL <http://www.isrl.uiuc.edu/~amag/langev/paper/yang99ACL.html>.
- Yang, Charles D. 2000. Internal and external forces in language change. *Language Variation and Change* 12:231–250. URL <http://www.isrl.uiuc.edu/~amag/langev/paper/yang00internalAnd.html>.
- Yang, Charles D. 2003. *Knowledge and learning in natural language*. USA: Oxford.
- Zipf, G.K. 1935. *The psycho-biology of language: an introduction to dynamic philology*. Houghton Mifflin. Also published 1965, MIT Press.
- Zipf, G.K. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison–Wesley. Also published 1968, MIT Press.
- Zwicky, Arnold M. 1970. Auxiliary reduction in english. *Linguistics Inquiry* 1:323–336.
- Zwicky, Arnold M., and Geoffrey K. Pullum. 1983. Cliticization vs. inflection: English n't. *Language* 59:502–513.