The Rhythm of Lexical Stress in Prose

Doug Beeferman
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
dougb+@cs.cmu.edu

Abstract

"Prose rhythm" is a widely observed but scarcely quantified phenomenon. We describe an information-theoretic model for measuring the regularity of lexical stress in English texts, and use it in combination with trigram language models to demonstrate a relationship between the probability of word sequences in English and the amount of rhythm present in them. We find that the stream of lexical stress in text from the Wall Street Journal has an entropy rate of less than 0.75 bits per syllable for common sentences. We observe that the average number of syllables per word is greater for rarer word sequences, and to normalize for this effect we run control experiments to show that the choice of word order contributes significantly to stress regularity, and increasingly with lexical probability.

1 Introduction

Rhythm inheres in creative output, asserting itself as the meter in music, the iambs and trochees of poetry, and the uniformity in distances between objects in art and architecture. More subtly there is widely believed to be rhythm in English prose, reflecting the arrangement of words, whether deliberate or subconscious, to enhance the perceived acoustic signal or reduce the burden of remembrance for the reader or author.

In this paper we describe an information-theoretic model based on lexical stress that substantiates this common perception and relates stress regularity in written speech (which we shall equate with the intuitive notion of "rhythm") to the probability of the text itself. By computing the *stress entropy rate* for both a set of *Wall Street Journal* sentences and a version of the corpus with randomized intra-sentential word order, we also find that word order contributes significantly to rhythm, particularly within highly probable sentences. We regard this as a first step in quantifying the extent to which metrical properties influence syntactic choice in writing.

1.1 Basics

In speech production, syllables are emitted as pulses of sound synchronized with movements of the musculature in the rib cage. Degrees of stress arise from variations in the amount of energy expended by the speaker to contract these muscles, and from other factors such as intonation. Perceptually stress is more abstractly defined, and it is often associated with "peaks of prominence" in some representation of the acoustic input signal (Ochsner, 1989).

Stress as a lexical property, the primary concern of this paper, is a function that maps a word to a sequence of discrete levels of physical stress, approximating the relative emphasis given each syllable when the word is pronounced. Phonologists distinguish between three levels of lexical stress in English: *primary*, *secondary*, and what we shall call *weak* for lack of a better substitute for *unstressed.* For the purposes of this paper we shall regard stresses as symbols fused serially in time by the writer or speaker, with words acting as building blocks of predefined stress sequences that may be arranged arbitrarily but never broken apart.

The *culminative* property of stress states that every content word has exactly one primary-stressed syllable, and that whatever syllables remain are subordinate to it. Monosyllabic function words such as *the* and *of* usually receive weak stress, while content words get one strong stress and possibly many secondary and weak stresses.

It has been widely observed that strong and weak tend to alternate at "rhythmically ideal disyllabic distances" (Kager, 1989a). "Ideal" here is a complex function involving production, perception, and many unknowns. Our concern is not to pinpoint this ideal, nor to answer precisely why it is sought by speakers and writers, but to gauge *to what extent* it is sought.

We seek to investigate, for example, whether the avoidance of primary stress *clash*, the placement of two or more strongly stressed syllables in succession, influences syntactic choice. In the Wall Street Journal corpus we find such sentences as "The **fol**-low-ing **is**-sues **re**-cent-ly were **filed** with the Se-**cur**-i-ties and Ex-**change** Com-**mis**-sion". The phrase "recently were filed" can be syntactically permuted as "were filed recently", but this clashes *filed* with the first syllable of *recently.* The chosen sentence avoids consecutive primary stresses. Kager postulates with a decidedly information theoretic undertone that the resulting binary alternation is "simply the maximal degree of rhythmic organization compatible with the requirement that adjacent stresses are to be avoided." (Kager, 1989a)

Certainly we are not proposing that a hard decision based only on metrical properties of the output is made to resolve syntactic choice ambiguity, in the case above or in general. Clearly semantic emphasis has its say in the decision. But it is our belief that rhythm makes a nontrivial contribution, and that the tools of statistics and information theory will help us to estimate it formally. Words are the building blocks. How much do their selection (diction) and their arrangement (syntax) act to enhance rhythm?

1.2 Past models and quantifications

Lexical stress is a well-studied subject at the intraword level. Rules governing how to map a word's orthographic or phonetic transcription to a sequence of stress values have been searched for and studied from rules-based, statistical, and connectionist perspectives.

Word-external stress regularity has been denied this level of attention. Patterns in phrases and compound words have been studied by Halle (Halle and Vergnaud, 1987) and others, who observe and reformulate such phenomena as the emphasis of the penultimate constituent in a compound noun (*National Center for* Supercomputing *Applications*, for example.) Treatment of *lexical* stress across word boundaries is scarce in the literature, however. Though prose rhythm inquiry is more than a hundred years old (Ochsner, 1989), it has largely been dismissed by the linguistic community as irrelevant to formal models, as a mere curiosity for literary analysis. This is partly because formal methods of inquiry have failed to present a compelling case for the existence of regularity (Harding, 1976).

Past attempts to quantify prose rhythm may be classified as perception-oriented or signal-oriented. In both cases the studies have typically focussed on regularities in the distance between peaks of prominence, or interstress intervals, either perceived by a human subject or measured in the signal. The former class of experiments relies on the subjective segmentation of utterances by a necessarily limited number of participants—subjects tapping out the rhythms they perceive in a waveform on a recording device, for example (Kager, 1989b). To say nothing of the psychoacoustic biases this methodology introduces, it relies on too little data for anything but a sterile set of means and variances.

Signal analysis, too, has not yet been applied to very large speech corpora for the purpose of investigating prose rhythm, though the technology now exists to lend efficiency to such studies. The experiments have been of smaller scope and geared toward detecting isochrony, regularity in absolute *time.* Jassem *et al.*(Jassem, Hill, and Witten, 1984) use statistical techniques such as regression to analyze the duration of what they term *rhythm units.* Jassem postulates that speech is composed of extrasyllable *narrow* rhythm units with roughly fixed duration independent of the number of syllable constituents, surrounded by variable-length *anacruses.* Abercrombie (Abercrombie, 1967) views speech as composed of metrical *feet* of variable length that begin with and are conceptually highlighted by a single stressed syllable.

Many experiments lead to the common conclusion that English is *stress-timed*, that there is some regularity in the absolute duration between strong stress events. In contrast to postulated syllable-timed languages like French in which we find exactly the inverse effect, speakers of English tend to expand and to contract syllable streams so that the duration between bounding primary stresses matches the other intervals in the utterance. It is unpleasant for production and perception alike, however, when too many weak-stressed syllables are forced into such an interval, or when this amount of "padding" varies wildly from one interval to the next. Prose rhythm analysts so far have not considered the syllable stream independent from syllabic, phonemic, or interstress duration. In particular they haven't measured the regularity of the purely lexical stream. They have instead continually re-answered questions concerning isochrony.

Given that speech can be divided into interstress units of roughly equal duration, we believe the more interesting question is whether a speaker or writer modifies his diction and syntax to fit a regular number of *syllables* into each unit. This question can only be answered by a lexical approach, an approach that pleasingly lends itself to efficient experimentation with very large amounts of data.

2 Stress entropy rate

We regard every syllable as having either strong or weak stress, and we employ a purely lexical, context independent mapping, a pronunciation dictionary [1], to tell us which syllables in a word receive which level of stress. We base our experiments on a binary-valued symbol set $\Sigma_1 = \{W, S\}$ and on a ternary-valued symbol set $\Sigma_2 = \{W, S, P\}$, where 'W' indicates weak stress, 'S' indicates strong stress,

[1] We use the 116,000-entry *CMU Pronouncing Dictionary version 0.4* for all experiments in this paper.

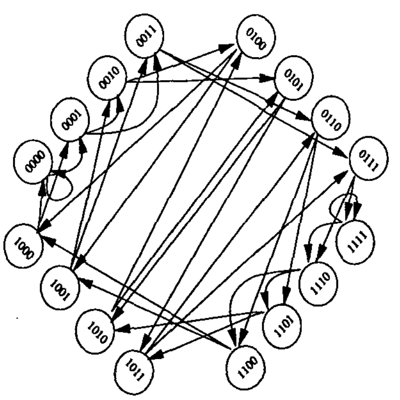

Figure 2: A 5-gram model viewed as a first-order Markov chain

and 'P' indicates a pause. Abstractly the dictionary maps words to sequences of symbols from {primary, secondary, unstressed}, which we interpret by downsampling to our binary system—primary stress is strong, non-stress is weak, and secondary stress ('2') we allow to be either weak or strong depending on the experiment we are conducting.

We represent a sentence as the concatenation of the stress sequences of its constituent words, with 'P' symbols (for the Σ_2 experiments) breaking the stream where natural pauses occur.

Traditional approaches to lexical language modeling provide insight on our analogous problem, in which the input is a stream of syllables rather than words and the values are drawn from a vocabulary Σ of stress levels. We wish to create a model that yields approximate values for probabilities of the form $p(s_k|s_0, s_1, \ldots, s_{k-1})$, where $s_i \in \Sigma$ is the stress symbol at syllable i in the text. A model with separate parameters for each history is prohibitively large, as the number of possible histories grows exponentially with the length of the input; and for the same reason it is impossible to train on limited data. Consequently we partition the history space into equivalence classes, and the stochastic n-gram approach that has served lexical language modeling so well treats two histories as equivalent if they end in the same $n-1$ symbols.

As Figure 2 demonstrates, an n-gram model is simply a stationary Markov chain of order $k = n - 1$, or equivalently a first-order Markov chain whose states are labeled with tuples from Σ^k.

To gauge the regularity and compressibility of the training data we can calculate the *entropy rate* of the stochastic process as approximated by our model, an upper bound on the expected number of bits needed to encode each symbol in the best possible encoding. Techniques for computing the entropy rate of a stationary Markov chain are well known in information theory (Cover and Thomas, 1991). If $\{X_i\}$ is a Markov chain with stationary distribution μ and transition matrix P, then its entropy rate is $H(\mathcal{X}) = -\sum_{i,j} \mu_i p_{ij} \log p_{ij}$.

The probabilities in P can be trained by accumulating, for each $(s_1, s_2, \ldots, s_k) \in \Sigma^k$, the k-gram count in $C(s_1, s_2, \ldots, s_k)$ in the training data, and normalizing by the $(k-1)$-gram count $C(s_1, s_2, \ldots, s_{k-1})$.

The stationary distribution μ satisfies $\mu P = \mu$, or equivalently $\mu_k = \sum_j \mu_j p_{j,k}$ (Parzen, 1962). In general finding μ for a large state space requires an eigenvector computation, but in the special case of an n-gram model it can be shown that the value in μ corresponding to the state $(s_1, s_2, \ldots, s_k)$ is simply the k-gram frequency $C(s_1, s_2, \ldots, s_k)/N$, where N is the number of symbols in the data.[2] We therefore can compute the entropy rate of a stress sequence in time linear in both the amount of data and the size of the state space. This efficiency will enable us to experiment with values of n as large as seven; for larger values the amount of training data, not time, is the limiting factor.

3 Methodology

The training procedure entails simply counting the number of occurrences of each n-gram for the training data and computing the stress entropy rate by the method described. As we treat each sentence as an independent event, no cross-sentence n-grams are kept: only those that fit between sentence boundaries are counted.

3.1 The meaning of stress entropy rate

We regard these experiments as computing the entropy rate of a Markov chain, estimated from training data, that approximately models the emission of symbols from a random source. The entropy rate bounds how compressible the training sequence is, and not precisely how predictable unseen sequences from the same source would be. To measure the efficacy of these models in prediction it would be necessary to divide the corpus, train a model on one subset, and measure the entropy rate of the other with respect to the trained model. Compression can take place off-line, after the entire training set is read, while prediction cannot "cheat" in this manner.

But we claim that our results predict how effective prediction would be, for the small state space in our Markov model and the huge amount of training data translate to very good state coverage. In language modeling, unseen words and unseen n-grams are a serious problem, and are typically combatted with smoothing techniques such as the backoff model and the discounting formula offered by Good and Turing. In our case, unseen "words" never occur, for

[2] This ignores edge effects, for $\sum_s C(s_1, s_2, \ldots, s_k) = N - k + 1$, but this discrepancy is negligible when N is very large.

Lis	ten	to	me	close	ly	I'll	en	deav	or	to	ex	plain	/
S	W	S	S	S	W	S	W	S	W	S	W	S	P
what	sep	ar	ates	a	char	la	tan	from	a	Char	le	magne	.
W	S	W	2	W	S	W	W	S	W	S	W	2	P

Figure 1: A song lyric exemplifies a highly regular stress stream (from the musical *Pippin* by Stephen Schwartz.)

the tiniest of realistic training sets will cover the binary or ternary vocabulary. Coverage of the n-gram set is complete for our prose training texts for n as high as eight; nor do singleton states (counts that occur only once), which are the bases of Turing's estimate of the frequency of untrained states in new data, occur until $n = 7$.

3.2 Lexicalizing stress

Lexical stress is the "backbone of speech rhythm" and the primary tool for its analysis. (Baum, 1952) While the precise acoustical prominences of syllables within an utterance are subject to certain word-external hierarchical constraints observed by Halle (Halle and Vergnaud, 1987) and others, lexical stress is a local property. The stress patterns of individual words within a phrase or sentence are generally context independent.

One source of error in our method is the ambiguity for words with multiple phonetic transcriptions that differ in stress assignment. Highly accurate techniques for part-of-speech labeling could be used for stress pattern disambiguation when the ambiguity is purely lexical, but often the choice, in both production and perception, is dialectal. It would be straightforward to divide among all alternatives the count for each n-gram that includes a word with multiple stress patterns, but in the absence of reliable frequency information to weight each pattern we chose simply to use the pronunciation listed first in the dictionary, which is judged by the lexicographer to be the most popular. Very little accuracy is lost in making this assumption. Of the 115,966 words in the dictionary, 4635 have more than one pronunciation; of these, 1269 have more than one distinct stress pattern; of these, 525 have different primary stress placements. This smallest class has a few common words (such as "refuse" used as a noun and as a verb), but most either occur infrequently in text (obscure proper nouns, for example), or have a primary pronunciation that is overwhelmingly more common than the rest.

4 Experiments

The efficiency of the n-gram training procedure allowed us to exploit a wealth of data—over 60 million syllables—from 38 million words of *Wall Street Journal* text. We discarded sentences not completely covered by the pronunciation dictionary, leaving 36.1 million words and 60.7 million syllables for experimentation.

Our first experiments used the binary Σ_1 alphabet. The maximum entropy rate possible for this process is one bit per syllable, and given the unigram distribution of stress values in the data (55.2% are primary), an upper bound of slightly over 0.99 bits can be computed. Examining the 4-gram frequencies for the entire corpus (Figure 3a) sharpens this substantially, yielding an entropy rate estimate of 0.846 bits per syllable. Most frequent among the 4-grams are the patterns WSWS and SWSW, consistent with the principle of binary alternation mentioned in section 1.

The 4-gram estimate matches quite closely with the estimate of 0.852 bits that can be derived from the distribution of word stress patterns excerpted in Figure 3b. But both measures overestimate the entropy rate by ignoring longer-range dependencies that become evident when we use larger values of n. For $n = 6$ we obtain a rate of 0.795 bits per syllable over the entire corpus.

Since we had several thousand times more data than is needed to make reliable estimates of stress entropy rate for values of n less than 7, it was practical to subdivide the corpus according to some criterion, and calculate the stress entropy rate for each subset as well as for the whole. We chose to divide at the sentence level and to partition the 1.59 million sentences in the data based on a likelihood measure suitable for testing the hypothesis from section 1.

A lexical trigram backoff-smoothed language model was trained on separate data to estimate the language *perplexity* of each sentence in the corpus. Sentence perplexity $PP(S)$ is the inverse of sentence probability normalized for length, $1/P(S)^{\frac{1}{|S|}}$, where $P(S)$ is the probability of the sentence according to the language model and $|S|$ is its word count. This measure gauges the average "surprise" after revealing each word in the sentence as judged by the trigram model. The question of whether more probable word sequences are also more rhythmic can be approximated by asking whether sentences with lower perplexity have lower stress entropy rate.

Each sentence in the corpus was assigned to one of one hundred bins according to its perplexity—sentences with perplexity between 0 and 10 were assigned to the first bin; between 10 and 20, the sec-

(a)

WWWW:	0.78%	WSWW:	6.91%	SWWW:	2.96%	SSWW:	3.94%
WWWS:	2.94%	WSWS:	11.00%	SWWS:	7.80%	SSWS:	8.59%
WWSW:	6.97%	WSSW:	6.16%	SWSW:	11.21%	SSSW:	6.25%
WWSS:	3.71%	WSSS:	6.06%	SWSS:	8.48%	SSSS:	6.27%

(b)

S	45.87%
SW	18.94%
W	9.54%
SWW	5.74%
WS	5.14%
WSW	4.54%

Figure 3: (a) The corpus frequencies of all binary stress 4-grams (based on 60.7 million syllables), with secondary stress mapped to "weak" (W). (b) The corpus frequencies of the top six lexical stress patterns.

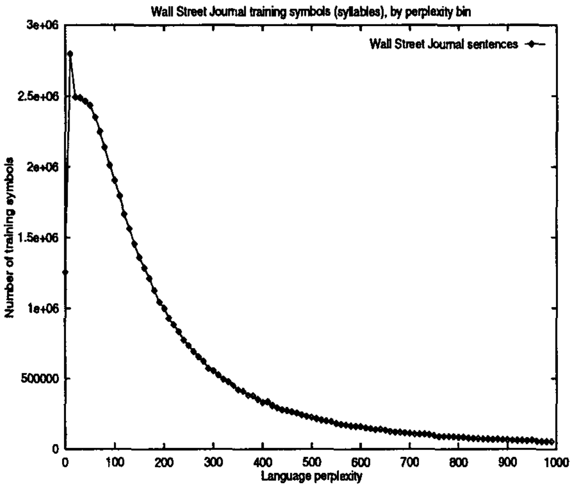

Figure 4: The amount of training data, in syllables, in each perplexity bin. The bin at perplexity level *pp* contains all sentences in the corpus with perplexity no less than *pp* and no greater than *pp* + 10. The smallest count (at bin 990) is 50662.

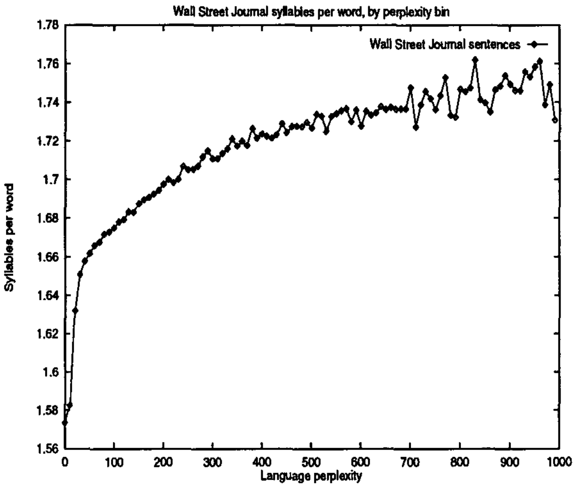

Figure 5: The average number of syllables per word for each perplexity bin.

ond; and so on. Sentences with perplexity greater than 1000, which numbered roughly 106 thousand out of 1.59 million, were discarded from all experiments, as 10-unit bins at that level captured too little data for statistical significance. A histogram showing the amount of training data (in syllables) per perplexity bin is given in Figure 4.

It is crucial to detect and understand potential sources of bias in the methodology so far. It is clear that the perplexity bins are well trained, but not yet that they are comparable with each other. Figure 5 shows the average number of syllables per word in sentences that appear in each bin. That this function is roughly increasing agrees with our intuition that sequences with longer words are rarer. But it biases our perplexity bins at the extremes. Early bins, with sequences that have a small syllable rate per word (1.57 in the 0 bin, for example), are predisposed to a lower stress entropy rate since primary stresses, which occur roughly once per word, are more frequent. Later bins are also likely to be prejudiced in that direction, for the inverse reason: The increasing frequency of multisyllabic words makes it more and more fashionable to transit to a weak-stressed syllable following a primary stress, sharpening the probability distribution and decreasing entropy.

This is verified when we run the stress entropy rate computation for each bin. The results for n-gram models of orders 3 through 7, for the case in which secondary lexical stress is mapped to the "weak" level, are shown in Figure 6.

All of the rates calculated are substantially less than a bit, but this only reflects the stress regularity inherent in the vocabulary and in word selection, and says nothing about word arrangement. The atomic elements in the text stream, the words, contribute regularity independently. To determine how much is contributed by the way they are glued together, we need to remove the bias of word choice.

For this reason we settled on a model size, $n = 6$, and performed a variety of experiments with both the original corpus and with a control set that contained exactly the same bins with exactly the same sentences, but mixed up. Each sentence in the control set was permuted with a pseudorandom sequence of swaps based on an insensitive function of the original; that is to say, identical sentences in the

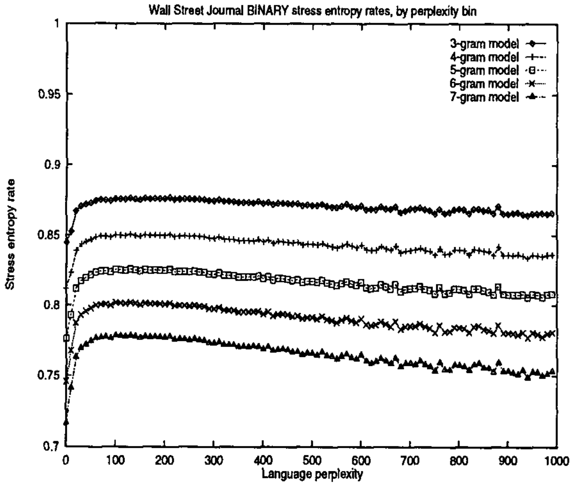

Figure 6: n-gram stress entropy rates for Σ_1, weak secondary stress

corpus were shuffled the same way and sentences differing by only one word were shuffled similarly. This allowed us to keep steady the effects of multiple copies of the same sentence in the same perplexity bin. More importantly, these tests hold everything constant—diction, syllable count, syllable rate per word—except for syntax, the arrangement of the chosen words within the sentence. Comparing the unrandomized results with this control experiment allows us, therefore, to factor out everything *but* word order. In particular, subtracting the stress entropy rates of the original sentences from the rates of the randomized sentences gives us a figure, relative entropy, that estimates how many bits we save by knowing the proper word order given the word choice. The results for these tests for weak and strong secondary stress are shown in Figures 7 and 8, including the difference curves between the randomized-word and original entropy rates.

The consistently positive difference function demonstrates that there *is* some extra stress regularity to be had with proper word order, about a hundredth of a bit on average. The difference is small indeed, but its consistency over hundreds of well-trained data points puts the observation on statistically solid ground.

The negative slopes of the difference curves suggests a more interesting conclusion: As sentence perplexity increases, the gap in stress entropy rate between syntactic sentences and randomly permuted sentences narrows. Restated inversely, using entropy rates for randomly permuted sentences as a baseline, sentences with higher sequence probability are relatively more rhythmical in the sense of our definition from section 1.

To supplement the Σ_1 binary vocabulary tests we ran the same experiments with $\Sigma_2 = \{0, 1, P\}$, introducing a pause symbol to examine how stress behaves near phrase boundaries. Commas, dashes, semicolons, colons, ellipses, and all sentence-terminating punctuation in the text, which were removed in the Σ_1 tests, were mapped to a single pause symbol for Σ_2. Pauses in the text arise not only from semantic constraints but also from physiological limitations. These include the "breath groups" of syllables that influence both vocalized and written production. (Ochsner, 1989). The results for these experiments are shown in Figures 9 and 10. Expectedly, adding the symbol increases the confusion and hence the entropy, but the rates remain less than a bit. The maximum possible rate for a ternary sequence is $\log_2 3 \approx 1.58$.

The experiments in this section were repeated with a larger perplexity interval that partitioned the corpus into 20 bins, each covering 50 units of perplexity. The resulting curves mirrored the finer-grain curves presented here.

5 Conclusions and future work

We have quantified lexical stress regularity, measured it in a large sample of written English prose, and shown there to be a significant contribution from word order that increases with lexical perplexity.

This contribution was measured by comparing the entropy rate of lexical stress in natural sentences with randomly permuted versions of the same. Randomizing the word order in this way yields a fairly crude baseline, as it produces asyntactic sequences in which, for example, single-syllable function words can unnaturally clash. To correct for this we modified the randomization algorithm to permute only open-class words and to fix in place determiners, particles, pronouns, and other closed-class words. We found the entropy rates to be consistently midway between the fully randomized and unrandomized values. But even this constrained randomization is weaker than what we'd like. Ideally we should factor out semantics as well as word choice, comparing each sentence in the corpus with its grammatical variations. While this is a difficult experiment to do automatically, we're hoping to approximate it using a natural language generation system based on link grammar under development by the author.

Also, we're currently testing other data sources such as the Switchboard corpus of telephone speech (Godfrey, Holliman, and McDaniel, 1992) to measure the effects of rhythm in more spontaneous and grammatically relaxed texts.

6 Acknowledgments

Comments from John Lafferty, Georg Niklfeld, and Frank Dellaert contributed greatly to this paper. The work was supported in part by an ARPA AASERT award, number DAAH04-95-1-0475.

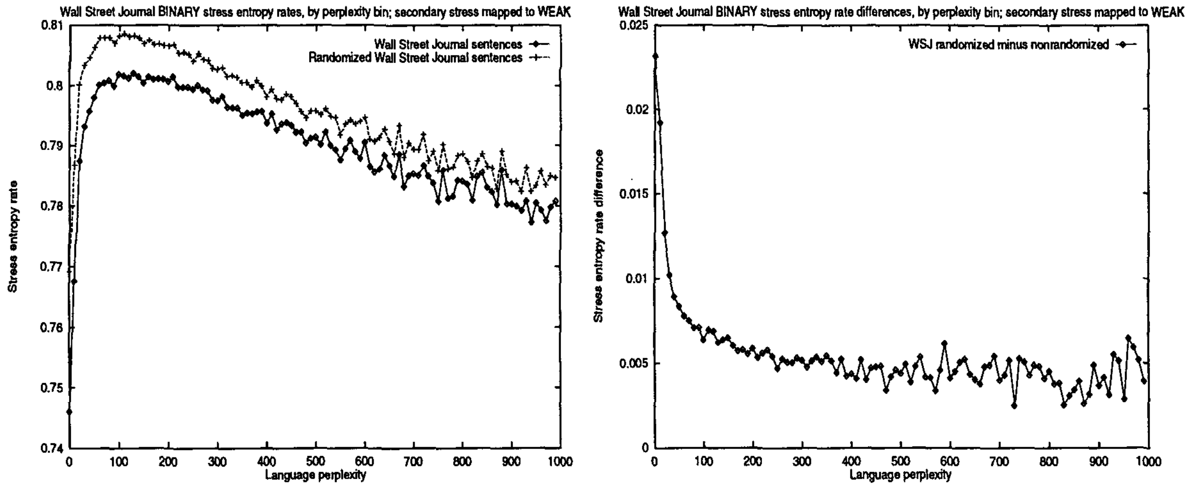

Figure 7: 6-gram stress entropy rates and difference curve for Σ_1, weak secondary stress

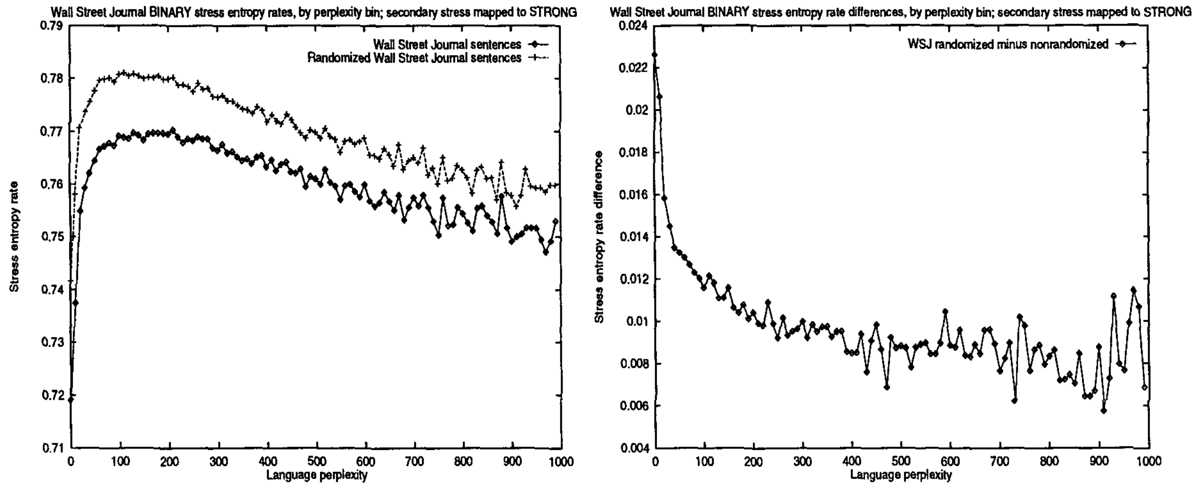

Figure 8: 6-gram entropy rates and difference curve for Σ_1, strong secondary stress

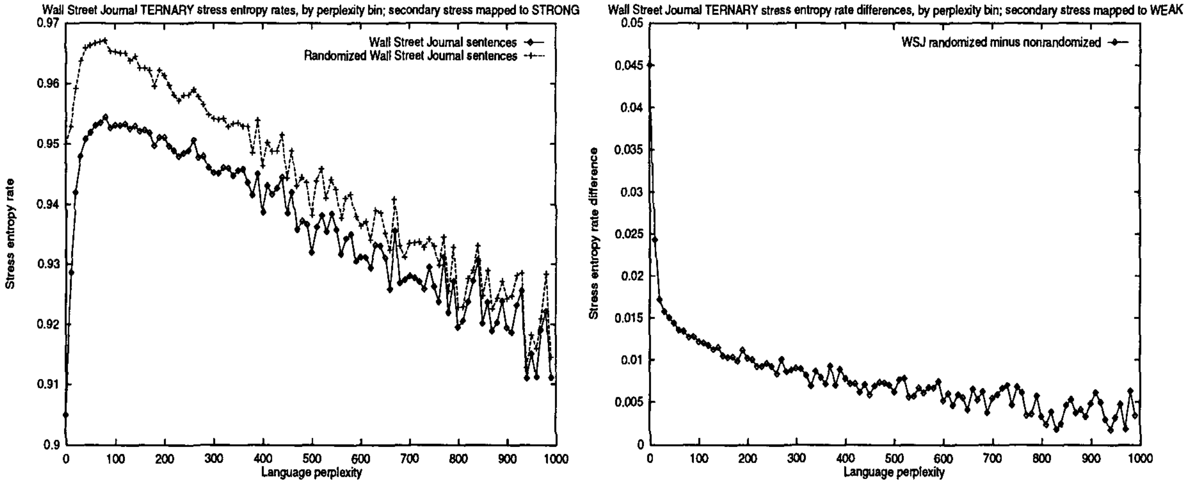

Figure 9: 6-gram entropy rates and difference curve for Σ_2, weak secondary stress

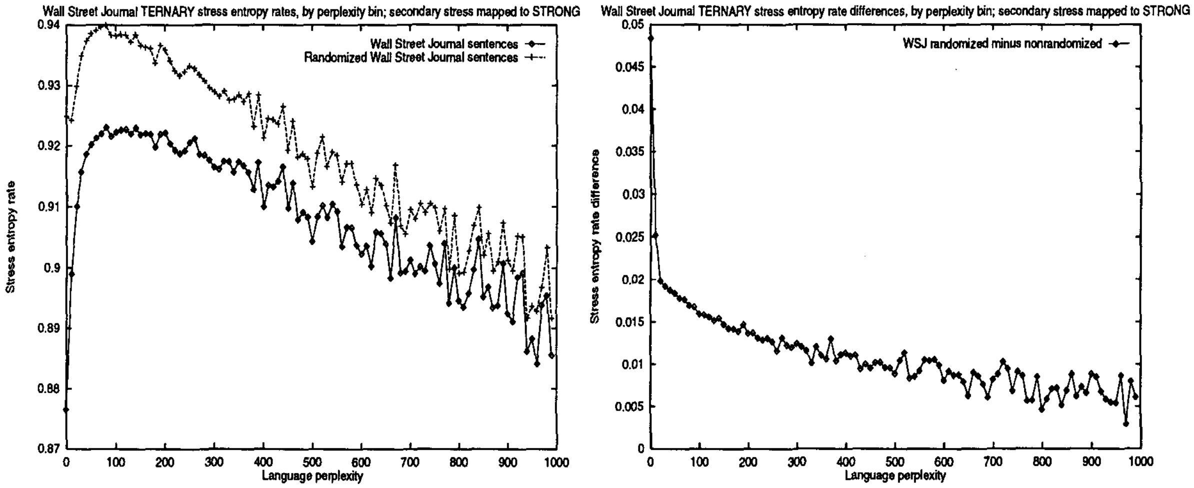

Figure 10: 6-gram entropy rates and difference curve for Σ_2, strong secondary stress

References

Abercrombie, D. 1967. *Elements of general phonetics.* Edinburgh University Press.

Baum, P. F. 1952. *The Other Harmony of Prose.* Duke University Press.

Cover, T. M. and J. A. Thomas. 1991. *Elements of information theory.* John Wiley & Sons, Inc.

Godfrey, J., E. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research development. In *Proc. ICASSP-92*, pages I-517-520.

Halle, M. and J. Vergnaud. 1987. *An essay on stress.* The MIT Press.

Harding, D. W. 1976. *Words into rhythm: English speech rhythm in verse and prose.* Cambridge University Press.

Jassem, W., D. R. Hill, and I. H. Witten. 1984. Isochrony in English speech: its statistical validity and linguistic relevance. In D. Gibbon and H. Richter, editors, *Intonation, rhythm, and accent: Studies in Discourse Phonology.* Walter de Gruyter, pages 203–225.

Kager, R. 1989a. *A metrical theory of stress and destressing in English and Dutch.* Foris Publications.

Kager, R. 1989b. *The rhythm of English prose.* Foris Publications.

Ochsner, R. S. 1989. *Rhythm and writing.* The Whitson Publishing Company.

Parzen, E. 1962. *Stochastic processes.* Holden-Day.