Variation of Entropy and Parse Trees of Sentences as a Function of the Sentence Number

Dmitriy Genzel and **Eugene Charniak**
Brown Laboratory for Linguistic Information Processing
Department of Computer Science
Brown University
Providence, RI, USA, 02912
{dg,ec}@cs.brown.edu

Abstract

In this paper we explore the variation of sentences as a function of the sentence number. We demonstrate that while the entropy of the sentence increases with the sentence number, it decreases at the paragraph boundaries in accordance with the Entropy Rate Constancy principle (introduced in related work). We also demonstrate that the principle holds for different genres and languages and explore the role of genre informativeness. We investigate potential causes of entropy variation by looking at the tree depth, the branching factor, the size of constituents, and the occurrence of gapping.

1 Introduction and Related Work

In many natural language processing applications, such as parsing or language modeling, sentences are treated as natural self-contained units. Yet it is well-known that for interpreting the sentences the discourse context is often very important. The later sentences in the discourse contain references to the entities in the preceding sentences, and this fact is often useful, e.g., in caching for language modeling (Goodman, 2001). The indirect influence of the context, however, can be observed even when a sentence is taken as a stand-alone unit, i.e., without its context. It is possible to distinguish between a set of earlier sentences and a set of later sentences without any direct comparison by computing certain local statistics of individual sentences, such as their entropy (Genzel and Charniak, 2002). In this work we provide additional evidence for this hypothesis and investigate other sentence statistics.

1.1 Entropy Rate Constancy

Entropy, as a measure of information, is often used in the communication theory. If humans have evolved to communicate in the most efficient way (some evidence for that is provided by Plotkin and Nowak (2000)), then they would communicate in such a way that the entropy rate would be constant, namely, equal to the channel capacity (Shannon, 1948).

In our previous work (Genzel and Charniak, 2002) we propose that entropy rate is indeed constant in human communications. When read in context, each sentence would appear to contain roughly the same amount of information, per word, whether it is the first sentence or the tenth one. Thus the tenth sentence, when taken out of context, must appear significantly more informative (and therefore harder to process), since it implicitly assumes that the reader already knows all the information in the preceding nine sentences. Indeed, the greater the sentence number, the harder to process the sentence must appear, though for large sentence numbers this would be very difficult to detect. This makes intuitive sense: out-of-context sentences *are* harder to understand than in-context ones, and first sentences can never be out of context. It is also demonstrated empirically through estimating entropy rate of various sentences.

In the first part of the present paper (Sections 2 and 3) we extend and further verify these results. In the second part (Section 4), we investigate the potential reasons underlying this variation in complexity by looking at the parse trees of the sentences. We also discuss how genre and style affect the strength of this effect.

1.2 Limitations of Preceding Work

In our previous work we demonstrate that the word entropy rate increases with the sentence number; we do it by estimating entropy of Wall Street Journal articles in Penn Treebank in three different ways. It may be the case, however, that this effect is corpus- and language-specific. To show that the Entropy Rate Constancy Principle is universal, we need to confirm it for different genres and different languages. We will address this issue in Section 3.

Furthermore, if the principle is correct, it should also apply to the sentences numbered from the beginning of a paragraph, rather than from the beginning of the article, since in either case there is a shift of topic. We will discuss this in Section 2.

2 Within-Paragraph Effects

2.1 Implications of Entropy Rate Constancy Principle

We have previously demonstrated (see Genzel and Charniak (2002) for detailed derivation) that the conditional entropy of the *i*th word in the sentence (X_i), given its local context L_i (the preceding words in the same sentence) and global context C_i (the words in all preceding sentences) can be represented as

$$H(X_i|C_i, L_i) = H(X_i|L_i) - I(X_i, C_i|L_i)$$

where $H(X_i|L_i)$ is the conditional entropy of the *i*th word given local context, and $I(X_i, C_i|L_i)$ is the conditional mutual information between the *i*th word and out-of-sentence context, given the local context. Since C_i increases with the sentence number, we will assume that, normally, it will provide more and more information with each sentence. This would cause the second term on the right to increase with the sentence number, and since $H(X_i|C_i, L_i)$ must remain constant (by our assumption), the first term should increase with sentence number, and it had been shown to do so (Genzel and Charniak, 2002).

Our assumption about the increase of the mutual information term is, however, likely to break at the paragraph boundary. If there is a topic shift at the boundary, the context probably provides more information to the preceding sentence, than it does to the new one. Hence, the second term will decrease, and so must the first one.

In the next section we will verify this experimentally.

2.2 Experimental Setup

We use the Wall Street Journal text (years 1987-1989) as our corpus. We take all articles that contain ten or more sentences, and extract the first ten sentences. Then we:

1. Group extracted sentences according to their sentence number into ten sets of 49559 sentences each.

2. Separate each set into two subsets, paragraph-starting and non-paragraph-starting sentences[1].

3. Combine first 45000 sentences from each set into the training set and keep all remaining data as 10 testing sets (19 testing subsets).

We use a simple smoothed trigram language model:

$$\begin{aligned} P(x_i|x_1 \ldots x_{i-1}) &\approx P(x_i|x_{i-2}x_{i-1}) \\ &= \lambda_1 \hat{P}(x_i|x_{i-2}x_{i-1}) \\ &+ \lambda_2 \hat{P}(x_i|x_{i-1}) \\ &+ (1 - \lambda_1 - \lambda_2)\hat{P}(x_i) \end{aligned}$$

where λ_1 and λ_2 are the smoothing coefficients[2], and $\hat{P}$ is a maximum likelihood estimate of the corresponding probability, e.g.,

$$\hat{P}(x_i|x_{i-2}x_{i-1}) = \frac{C(x_{i-2}x_{i-1}x_i)}{C(x_{i-2}x_{i-1})}$$

where $C(x_i \ldots x_j)$ is the number of times this sequence appears in the training data.

We then evaluate the resulting model on each of the testing sets, computing per-word entropy of the set:

$$\hat{H}(X) = \frac{1}{|X|} \sum_{x_i \in X} \log P(x_i|x_{i-2}x_{i-1})$$

[1] First sentences are, of course, all paragraph-starting.

[2] We have arbitrarily chosen the smoothing coefficients to be 0.5 and 0.3, correspondingly.

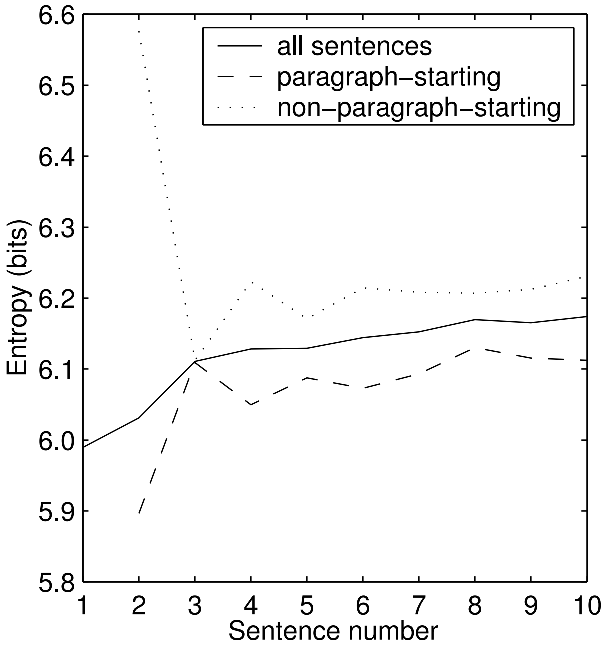


Figure 1: Entropy vs. Sentence number

2.3 Results and Discussion

As outlined above, we have ten testing sets, one for each sentence number; each set (except for the first) is split into two subsets: sentences that start a paragraph, and sentences that do not. The results for full sets, paragraph-starting subsets, and non-paragraph-starting subsets are presented in Figure 1.

First, we can see that the the entropy for full sets (solid line) is generally increasing. This result corresponds to the previously discussed effect of entropy increasing with the sentence number. We also see that for all sentence numbers the paragraph-starting sentences have lower entropy than the non-paragraph-starting ones, which is what we intended to demonstrate. In such a way, the paragraph-starting sentences are similar to the first sentences, which makes intuitive sense.

All the lines roughly show that entropy increases with the sentence number, but the behavior at the second and the third sentences is somewhat strange. We do not yet have a good explanation of this phenomenon, except to point out that paragraphs that start at the second or third sentences are probably not “normal” because they most likely do not indicate a topic shift. Another possible explanation is that this effect is an artifact of the corpus used.

We have also tried to group sentences based on their sentence number within paragraph, but were unable to observe a significant effect. This may be due to the decrease of this effect in the later sentences of large articles, or perhaps due to the relative weakness of the effect[3].

3 Different Genres and Languages

3.1 Experiments on Fiction

3.1.1 Introduction

All the work on this problem so far has focused on the Wall Street Journal articles. The results are thus naturally suspect; perhaps the observed effect is simply an artifact of the journalistic writing style. To address this criticism, we need to perform comparable experiments on another genre.

Wall Street Journal is a fairly prototypical example of a news article, or, more generally, a writing with a primarily informative purpose. One obvious counterpart of such a genre is fiction[4]. Another alternative might be to use transcripts of spoken dialogue.

Unfortunately, works of fiction, are either non-homogeneous (collections of works) or relatively short with relatively long subdivisions (chapters). This is crucial, since in the sentence number experiments we obtain one data point per article, therefore it is impossible to use book chapters in place of articles.

3.1.2 Experimental Setup and Results

For our experiments we use *War and Peace* (Tolstoy, 1869), since it is rather large and publicly available. It contains only about 365 rather long chapters[5]. Unlike WSJ articles, each chapter is not written on a single topic, but usually has multiple topic shifts. These shifts, however, are marked only as paragraph breaks. We, therefore, have to assume that each paragraph break represents a topic shift,

[3] We combine into one set very heterogeneous data: both 1st and 51st sentence might be in the same set, if they both start a paragraph. The experiment in Section 2.2 groups only the paragraph-starting sentences with the same sentence number.

[4] We use prose rather than poetry, which presumably is even less informative, because poetry often has superficial constraints (meter); also, it is hard to find a large *homogeneous* poetry collection.

[5] For comparison, Penn Treebank contains over 2400 (much shorter) WSJ articles.

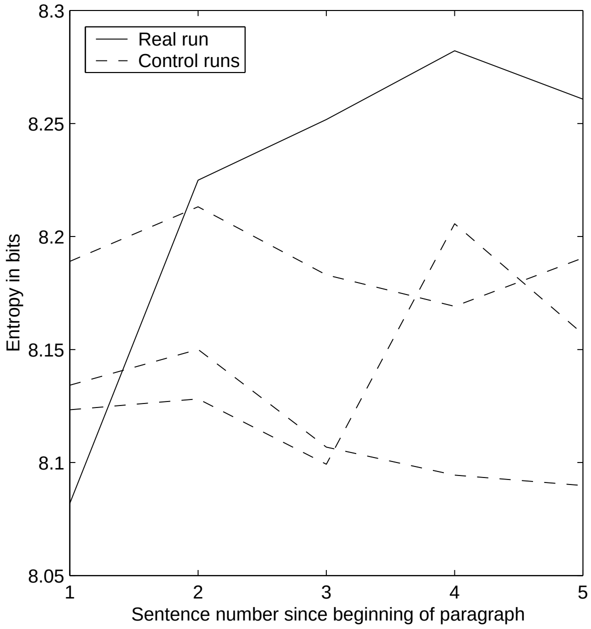


Figure 2: *War and Peace*: English

and treat each paragraph as being an equivalent of a WSJ article, even though this is obviously suboptimal.

The experimental setup is very similar to the one used in Section 2.2. We use roughly half of the data for training purposes and split the rest into testing sets, one per each sentence number, counted from the beginning of a paragraph.

We then evaluate the results using the same method as in Section 2.2. We expect that the entropy would increase with the sentence number, just as in the case of the sentences numbered from the article boundary. This effect is present, but is not very pronounced. To make sure that it is statistically significant, we also do 1000 control runs for comparison, with paragraph breaks inserted randomly at the appropriate rate. The results (including 3 random runs) can be seen in Figure 2. To make sure our results are significant we compare the correlation coefficient between entropy and sentence number to ones from simulated runs, and find them to be significant (P=0.016).

It is fairly clear that the variation, especially between the first and the later sentences, is greater than it would be expected for a purely random occurrence. We will see further evidence for this in the next section.

3.2 Experiments on Other Languages

To further verify that this effect is significant and universal, it is necessary to do similar experiments in other languages. Luckily, *War and Peace* is also digitally available in other languages, of which we pick Russian and Spanish for our experiments.

We follow the same experimental procedure as in Section 3.1.2 and obtain the results for Russian (Figure 3(a)) and Spanish (Figure 3(b)). We see that results are very similar to the ones we obtained for English. The results are again significant for both Russian (P=0.004) and Spanish (P=0.028).

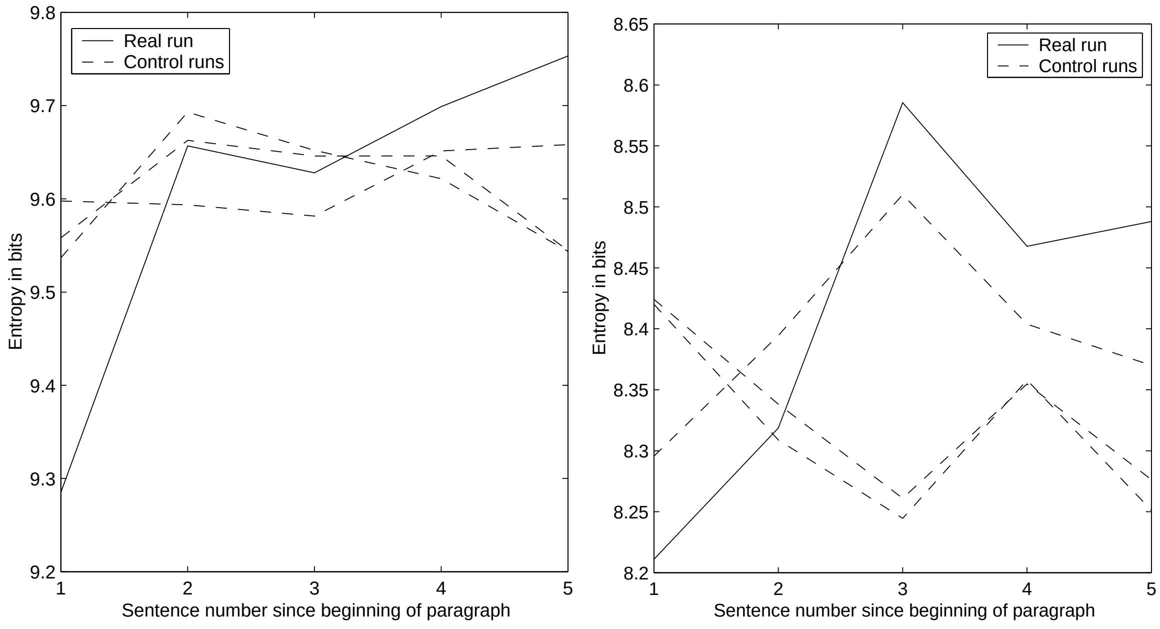
3.3 Influence of Genre on the Strength of the Effect

We have established that entropy increases with the sentence number in the works of fiction. We observe, however, that the effect is smaller than reported in our previous work (Genzel and Charniak, 2002) for Wall Street Journal articles. This is to be expected, since business and news writing tends to be more structured and informative in nature, gradually introducing the reader to the topic. Context, therefore, plays greater role in this style of writing.

To further investigate the influence of genre and style on the strength of the effect we perform experiments on data from British National Corpus (Leech, 1992) which is marked by genre.

For each genre, we extract first ten sentences of each genre subdivision of ten or more sentences. 90% of this data is used as training data and 10% as testing data. Testing data is separated into ten sets: all the first sentences, all the second sentences, and so on. We then use a trigram model trained on the training data set to find the average per-word entropy for each set. We obtain ten numbers, which in general tend to increase with the sentence number. To find the degree to which they increase, we compute the correlation coefficient between the entropy estimates and the sentence numbers. We report these coefficients for some genres in Table 1. To ensure reliability of results we performed the described process 400 times for each genre, sampling different testing sets.

The results are very interesting and strongly support our assumption that informative and structured (and perhaps better-written) genres will have



(a) Russian

(b) Spanish

Figure 3: *War and Peace*

stronger correlations between entropy and sentence number. There is only one genre, tabloid newspapers[6], that has negative correlation. The four genres with the smallest correlation are all quite non-informative: tabloids, popular magazines, advertisements[7] and poetry. Academic writing has higher correlation coefficients than non-academic. Also, humanities and social sciences writing is probably more structured and better stylistically than science and engineering writing. At the bottom of the table we have genres which tend to be produced by professional writers (biography), are very informative (TV news feed) or persuasive and rhetorical (parliamentary proceedings).

3.4 Conclusions

We have demonstrated that paragraph boundaries often cause the entropy to decrease, which seems to support the Entropy Rate Constancy principle. The effects are not very large, perhaps due to the fact that each new paragraph does not necessarily represent a shift of topic. This is especially true in a medium like the Wall Street Journal, where articles are very focused and tend to stay on one topic. In fiction, paragraphs are often used to mark a topic shift, but probably only a small proportion of paragraph breaks in fact represents topic shifts. We also observed that more informative and structured writing is subject to stronger effect than speculative and imaginative one, but the effect is present in almost all writing.

In the next section we will discuss the potential causes of the entropy results presented both in the preceding and this work.

4 Investigating Non-Lexical Causes

In our previous work we discuss potential causes of the entropy increase. We find that both lexical (*which* words are used) and non-lexical (*how* the words are used) causes are present. In this section we will discuss possible non-lexical causes.

We know that some non-lexical causes are present. The most natural way to find these causes is to examine the parse trees of the sentences. Therefore, we collect a number of statistics on the parse

[6] Perhaps, in this case the readers are only expected to look at the headlines.

[7] Advertisements could be called informative, but they tend to be sets of loosely related sentences describing various features, often in no particular order.

BNC genre	Corr. coef.
Tabloid newspapers	-0.342 ± 0.014
Popular magazines	0.073 ± 0.016
Print advertisements	0.175 ± 0.015
Fiction: poetry	0.261 ± 0.013
Religious texts	0.328 ± 0.012
Newspapers: commerce/finance	0.365 ± 0.013
Non-acad: natural sciences	0.371 ± 0.012
Official documents	0.391 ± 0.012
Fiction: prose	0.409 ± 0.011
Non-acad: medicine	0.411 ± 0.013
Newspapers: sports	0.433 ± 0.047
Acad: natural sciences	0.445 ± 0.010
Non-acad: tech, engineering	0.478 ± 0.011
Non-acad: politics, law, educ.	0.512 ± 0.004
Acad: medicine	0.517 ± 0.007
Acad: tech, engineering	0.521 ± 0.010
Newspapers: news reportage	0.541 ± 0.009
Non-acad: social sciences	0.541 ± 0.008
Non-acad: humanities	0.598 ± 0.007
Acad: politics, laws, educ.	0.619 ± 0.006
Newspapers: miscellaneous	0.622 ± 0.009
Acad: humanities	0.676 ± 0.007
Commerce/finance, economics	0.678 ± 0.007
Acad: social sciences	0.688 ± 0.004
Parliamentary proceedings	0.774 ± 0.002
TV news script	0.850 ± 0.002
Biographies	0.894 ± 0.001

Table 1: Correlation coefficient for different genres

trees and investigate if any statistics show a significant change with the sentence number.

4.1 Experimental Setup

We use the whole Penn Treebank corpus (Marcus et al., 1993) as our data set. This corpus contains about 50000 parsed sentences.

Many of the statistics we wish to compute are very sensitive to the length of the sentence. For example, the depth of the tree is almost linearly related to the sentence length. This is important because the average length of the sentence varies with the sentence number. To make sure we exclude the effect of the sentence length, we need to normalize for it.

We proceed in the following way. Let T be the set of trees, and $f : T \to \mathbb{R}$ be some statistic of a tree. Let $l(t)$ be the length of the underlying sentence for tree t. Let $L(n) = \{t|l(t) = n\}$ be the set of trees of size n. Let $L_f(n)$ be defined as $\frac{1}{|L(n)|}\sum_{t \in L(n)} f(t)$, the average value of the statistic f on all sentences of length n. We then define the sentence-length-adjusted statistic, for all t, as

$$f'(t) = \frac{f(t)}{L_f(l(t))}$$

The average value of the adjusted statistic is now equal to 1, and it is independent of the sentence length.

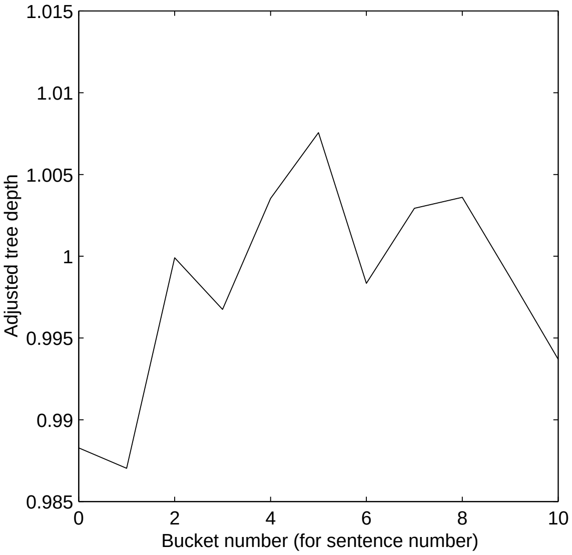


Figure 4: Tree Depth

We can now report the average value of each statistic for each sentence number, as we have done before, but instead we will group the sentence numbers into a small number of “buckets” of exponentially increasing length[8]. We do so to capture the behavior for all the sentence numbers, and not just for the first ten (as before), as well as to lump together sentences with similar sentence numbers, for which we do not expect much variation.

4.2 Tree Depth

The first statistic we consider is also the most natural: tree depth. The results can be seen in Figure 4.

In the first part of the graph we observe an increase in tree depth, which is consistent with the increasing complexity of the sentences. In the later

[8] For sentence number n we compute the bucket number as $\lfloor \log_{1.5} n \rfloor$

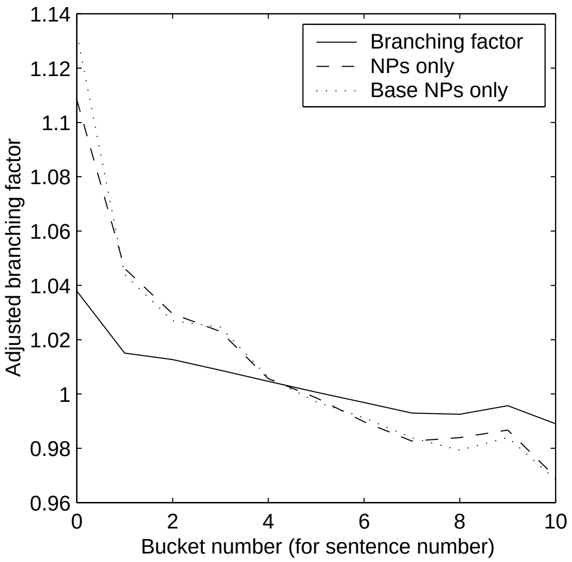


Figure 5: Branching factor

sentences, the depth decreases slightly, but still stays above the depth of the first few sentences.

4.3 Branching Factor and NP Size

Another statistic we investigate is the average branching factor, defined as the average number of children of all non-leaf nodes in the tree. It does not appear to be directly correlated with the sentence length, but we normalize it to make sure it is on the same scale, so we can compare the strength of resulting effect.

Again, we expect lower entropy to correspond to flatter trees, which corresponds to large branching factor. Therefore we expect the branching factor to decrease with the sentence number, which is indeed what we observe (Figure 5, solid line).

Each non-leaf node contributes to the average branching factor. It is likely, however, that the branching factor changes with the sentence number for certain types of nodes only. The most obvious contributors for this effect seem to be NP (noun phrase) nodes. Indeed, one is likely to use several words to refer to an object for the first time, but only a few words (even one, e.g., a pronoun) when referring to it later. We verify this intuitive suggestion, by computing the branching factor for NP, VP (verb phrase) and PP (prepositional phrase) nodes. Only NP nodes show the effect, and it is much stronger (Figure 5, dashed line) than the effect for the branching factor.

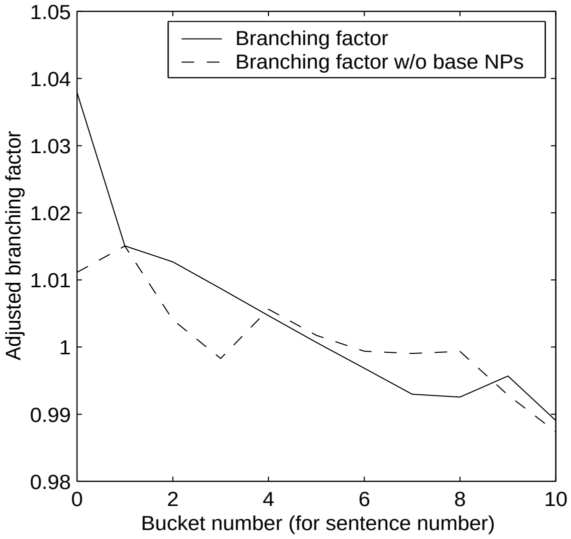


Figure 6: Branching Factor without Base NPs

Furthermore, it is natural to expect that most of this effect arises from base NPs, which are defined as the NP nodes whose children are all leaf nodes. Indeed, base NPs show a slightly more pronounced effect, at least with regard to the first sentence (Figure 5, dotted line).

4.4 Further Investigations

We need to determine whether we have accounted for all of the branching factor effect, by proposing that it is simply due to decrease in the size of the base NPs. To check, we compute the average branching factor, excluding base NP nodes.

By comparing the solid line in Figure 6 (the original average branching factor result) with the dashed line (base NPs excluded), you can see that base NPs account for most, though not all of the effect. It seems, then, that this problem requires further investigation.

4.5 Gapping

Another potential reason for the increase in the sentence complexity might be the increase in the use of gapping. We investigate whether the number of the ellipsis constructions varies with the sentence number. We again use Penn Treebank for this experi-

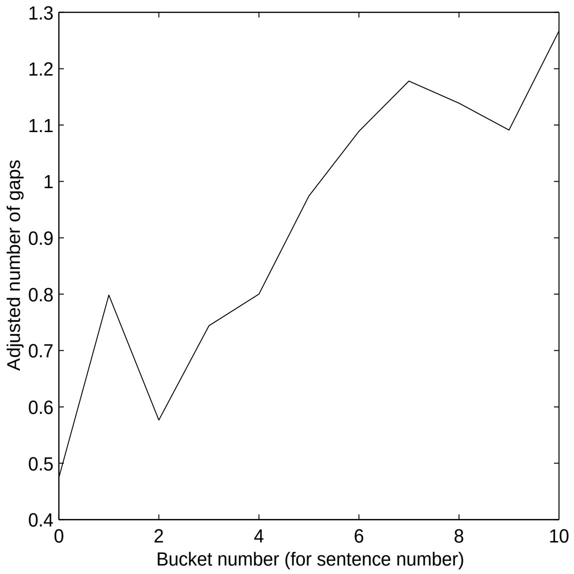


Figure 7: Number of ellipsis nodes

ment[9].

As we can see from Figure 7, there is indeed a significant increase in the use of ellipsis as the sentence number increases, which presumably makes the sentences more complex. Only about 1.5% of all the sentences, however, have gaps.

5 Future Work and Conclusions

We have discovered a number of interesting facts about the variation of sentences with the sentence number. It has been previously known that the complexity of the sentences increases with the sentence number. We have shown here that the complexity tends to decrease at the paragraph breaks in accordance with the Entropy Rate Constancy principle. We have verified that entropy also increases with the sentence number outside of Wall Street Journal domain by testing it on a work of fiction. We have also verified that it holds for languages other than English. We have found that the strength of the effect depends on the informativeness of a genre.

We also looked at the various statistics that show a significant change with the sentence number, such as the tree depth, the branching factor, the size of noun phrases, and the occurrence of gapping.

Unfortunately, we have been unable to apply these results successfully to any practical problem so far, primarily because the effects are significant on average and not in any individual instances. Finding applications of these results is the most important direction for future research.

Also, since this paper essentially makes statements about human processing, it would be very appropriate to to verify the Entropy Rate Constancy principle by doing reading time experiments on human subjects.

6 Acknowledgments

We would like to acknowledge the members of the Brown Laboratory for Linguistic Information Processing and particularly Mark Johnson for many useful discussions. This research has been supported in part by NSF grants IIS 0085940, IIS 0112435, and DGE 9870676.

References

A. Bies, M. Ferguson, K. Katz, and R. MacIntyre, 1995. *Bracketing Guidelines for Treebank II Style Penn Treebank Project*. Penn Treebank Project, University of Pennsylvania.

D. Genzel and E. Charniak. 2002. Entropy rate constancy in text. In *Proceedings of ACL–2002, Philadelphia*.

J. T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech and Language*, 15:403–434.

G. Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19:313–330.

J. B. Plotkin and M. A. Nowak. 2000. Language evolution and information theory. *Journal of Theoretical Biology*, pages 147–159.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October.

L. Tolstoy. 1869. *War and Peace*. Available online, in 4 languages (Russian, English, Spanish, Italian): http://www.magister.msk.ru/library/tolstoy/wp/wp00.htm.

[9] Ellipsis nodes in Penn Treebank are marked with *?*. See Bies et al. (1995) for details.