The Entropy Rate Principle as a Predictor of Processing Effort: An Evaluation against Eye-tracking Data

Frank Keller
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
keller@inf.ed.ac.uk

Abstract

This paper provides evidence for Genzel and Charniak's (2002) entropy rate principle, which predicts that the entropy of a sentence increases with its position in the text. We show that this principle holds for individual sentences (not just for averages), but we also find that the entropy rate effect is partly an artifact of sentence length, which also correlates with sentence position. Secondly, we evaluate a set of predictions that the entropy rate principle makes for human language processing; using a corpus of eye-tracking data, we show that entropy and processing effort are correlated, and that processing effort is constant throughout a text.

1 Introduction

Genzel and Charniak (2002, 2003) introduce the entropy rate principle, which states that speakers produce language whose entropy rate is on average constant. The motivation for this comes from information theory: the most efficient way of transmitting information through a noisy channel is at a constant rate. If human communication has evolved to be optimal in this sense, then we would expect humans to produce text and speech with approximately constant entropy. There is some evidence that this is true for speech (Aylett, 1999).

For text, the entropy rate principle predicts that the entropy of an individual sentence increases with its position in the text, if entropy is measured out of context. Genzel and Charniak (2002) show that this prediction is true for the Wall Street Journal corpus, for both function words and for content words. They estimate entropy either using a language model or using a probabilistic parser; the effect can be observed in both cases. Genzel and Charniak (2003) extend this results in several ways: they show that the effect holds for different genres (but the effect size varies across genres), and also applies within paragraphs, not only within whole texts. Furthermore, they show that the effect can also be obtained for language other than English (Russian and

Spanish). The entropy rate principle also predicts that a language model that takes context into account should yield lower entropy estimates compared to an out of context language model. Genzel and Charniak (2002) show that this prediction holds for caching language models such as the ones proposed by Kuhn and de Mori (1990).

The aim of the present paper is to shed further light on the entropy rate effect discovered by Genzel and Charniak (2002, 2003) (henceforth G&C) by providing new evidence in two areas.

In Experiment 1, we replicate G&C's entropy rate effect and investigate the source of the effect. The results show that the correlation coefficients that G&C report are inflated by averaging over sentences with the same position, and by restricting the range of the sentence position considered. Once these restrictions are removed the effect is smaller, but still significant. We also show that the effect is to a large extend due to a confound with sentence length: longer sentences tend to occur later in the text. However, we are able to demonstrate that the entropy rate effect still holds once this confound has been removed.

In Experiment 2, we test the psycholinguistic predictions of the entropy rate principle. This experiment uses a subset of the British National Corpus as training data and tests on the Embra corpus, a set of newspaper articles annotated with eye-movement data. We find that there is a correlation between the entropy of a sentence and the processing effort it causes, as measured by reading times in eye-tracking data. We also show that there is no correlation between processing effort and sentence position, which indicates that processing effort in context is constant through a text, which is one of the assumptions underlying the entropy rate principle.

2 Predictions for Human Language Processing

Let us examine the psycholinguistic predictions of G&C's entropy rate principle in more detail. We need to distinguish two types of predictions: in-

context predictions and out-of-context predictions. The principle states that the entropy rate in a text is constant, i.e., that speakers produce sentences so that on average, all sentences in a text have the same entropy. In other words, communication is optimal in the sense that all sentences in the text are equally easy to understand, as they all have the same entropy.

This constancy principle is claimed to hold for connected text: all sentences in a text should be equally easy to process if they are presented in context. If we take reading time as a measure of processing effort, then the principle predicts that there should be no significant correlation between sentence position and reading time in context. We will test this prediction in Experiment 2 using an eye-tracking corpus consisting of connected text.

The entropy rate principle also makes the following prediction: if the entropy of a sentence is measured out of context (i.e., without taking the preceding sentences into account), then entropy will increase with sentence position. This prediction was tested extensively by G&C, whose results will be replicated in Experiment 1. With respect to processing difficulty, the entropy rate principle also predicts that processing difficulty out of context (i.e., if isolated sentences are presented to experimental subjects) should increase with sentence position. We could not test this prediction, as we only had in-context reading time data available for the present study.

However, there is another important prediction that can be derived from the entropy rate principle: sentences with a higher entropy should have higher reading times. This is an important precondition for the entropy rate principle, whose claims about the relationship between entropy and sentence position are only meaningful if entropy and processing effort are correlated. If there was no such correlation, then there would be no reason to assume that the out-of-context entropy of a sentence increases with sentence position. G&C explicitly refer to this relationship i.e., they assume that a sentence that is more informative is harder to process (Genzel and Charniak, 2003, p. 65). Experiment 1 will try to demonstrate the validity of this important prerequisite of the entropy rate principle.

3 Experiment 1: Entropy Rate and Sentence Length

The main aim of this experiment was to replicate G&C's entropy rate effect. A second aim was to test the generality of their result by determining if the relationship between sentence position and en-

tropy also holds for individual sentences (rather than for averages over sentences of a given position, as tested by G&C). We also investigated the effect of two parameters that G&C did not explore: the cut-off for article position (G&C only deal with sentences up to position 25), and the size of the n-gram used for estimating sentence probability. Finally, we include sentence length as a baseline that entropy-based models should be evaluated against.

3.1 Method

3.1.1 Materials

This experiment used the same corpus as Genzel and Charniak (2002), viz., the Wall Street Journal part of the Penn Treebank, divided into a training set (section 0–20) and a test set (sections 21–24). Each article was treated as a separate text, and sentence positions were computed by counting the sentences from the beginning of the text. The training set contained 42,075 sentences, the test set 7,133 sentences. The sentence positions in the test set varied between one and 149.

3.1.2 Procedure

The per-word entropy was computed using an n-gram language model, as proposed by G&C:[1]

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

Here, $\hat{H}(X)$ is the estimate of the per-word entropy of the sentence X, consisting of the words x_i, and n is the size of the n-gram. The n-gram probabilities were computed using the CMU-Cambridge language modeling toolkit (Clarkson and Rosenfeld, 1997), with the following parameters: vocabulary size 50,000; smoothing by absolute discounting; sentence beginning and sentence end as context cues (default values were used for all other parameters).

G&C use $n = 3$, i.e., a trigram model. We experimented with this parameter and used $n = 1, \ldots, 5$. For $n = 1$, equation (1) reduces to $\hat{H}(X) = -\frac{1}{|X|} \sum_{x_i \in X} \log P(x_i)$, i.e., a model based on word frequency.

The experiment also includes a simple model that does not take any probabilistic information into account, but simply uses the sentence length $|X|$ to predict sentence position. This model will serve as the baseline.

---

[1]Note that the original definition given by Genzel and Charniak (2002, 2003) does not include the minus sign. However, all their graphs display entropy as a positive quantity, hence we conclude that this is the definition they are using.
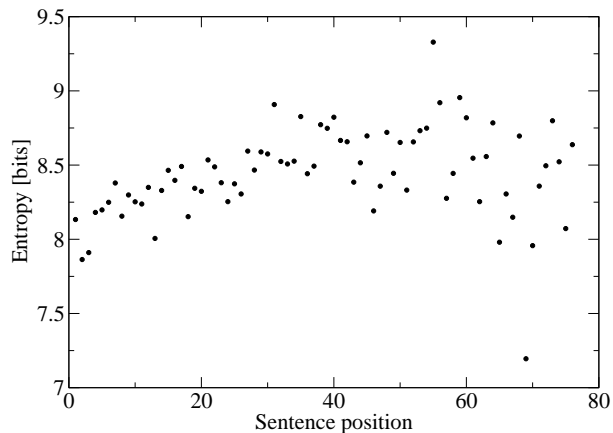
Figure 1: Experiment 1: correlation of sentence entropy and sentence position (bins, 3-grams, cut-off 76)

Figure 2: Experiment 1: correlation of sentence length and sentence position (bins, cut-off 76)

We also vary another parameter: c, the cut-off for the position. Genzel and Charniak (2002) use $c = 25$, i.e., only sentences with a position of 25 or lower are considered. In Genzel and Charniak (2003), an even smaller cut-off of $c = 10$ is used. This severely restricts the generality of the results obtained. We will therefore report results not only for $c = 25$, but also for $c = 76$. This cut-off has been set so that there are at least 10 items in the test set for each position. Furthermore, we also repeated the experiment without a cut-off for sentence length.

3.2 Results

Table 1 shows the results for the replication of Genzel and Charniak's (2002) entropy rate effect. The results at the top of the table were obtained using binning, i.e., we computed the mean entropy of all sentences of a given position, and then correlated these mean entropies with the sentence positions. The parameters n (n-gram size) and c (cut-off value) were varied as indicated in the previous section.

The bottom of Table 1 gives the correlation coefficients computed on the raw data, i.e., without binning: here, we correlated the entropy of a given sentence directly with its position. The graphs in Figure 1 and Figure 2 illustrate the relationship between position and entropy and between position and length, respectively.

3.3 Discussion

3.3.1 Entropy Rate and Sentence Length

The results displayed in Table 1 confirm G&C's main finding, i.e., that entropy increases with sentence length. For a cut-off of $c = 25$ (as used by G&C), a maximum correlation of 0.6480 is obtained (for the 4-gram model). The correlations for the other n-gram models are lower. All correlations
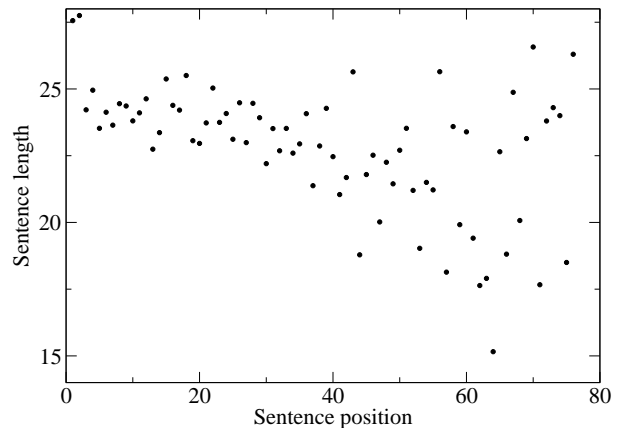
are significant (with the exception of the unigram model). However, we also find that a substantial correlation of -0.4607 is obtained even for the baseline model: there is a negative correlation between sentence length and sentence position, i.e., longer sentences tend to occur earlier in the text. This finding potentially undermines the entropy rate effect, as it raises the possibility that this effect is simply an effect of sentence length, rather than of sentence entropy. Note that the correlation coefficient for the none of the n-gram models is significantly higher than the baseline (significance was computed on the absolute values of the correlation coefficients).

The second finding concerns the question whether the entropy rate effect generalizes to sentences with a position of greater than 25. The results in Table 1 show that the effect generalizes to a cut-off of $c = 76$ (recall that this value was chosen so that each position is represented at least ten times in the test data). Again, we find a significant correlation between entropy and sentence position for all values of n. This is illustrated in Figure 1. However, none of the n-gram models is able to beat the baseline of simple sentence position; in fact, now all models (with the exception of the unigram model) perform significantly worse than the baseline. The correlation obtained by the baseline model is graphed in Figure 2.

Finally, we tried to generalize the entropy rate effect to sentences with arbitrary position (no cut-off). Here, we find that there is no significant positive correlation between entropy and position for any of the n-gram models. Only sentence length yields a reliable correlation, though it is smaller than if a cut-off is applied. This result is perhaps not surprising, as a lot of the data is very sparse: for positions between 77 and 149, less than ten data points are

	c = 25		c = 76		c = ∞	
Binned data	r	p	r	p	r	p
Entropy 1-gram	0.0593*	0.7784	0.3583	0.0015	−0.3486*	0.0000
Entropy 2-gram	0.4916	0.0126	0.2849*	0.0126	0.0723	0.3808
Entropy 3-gram	0.6387	0.0006	0.2427*	0.0346	0.1350	0.1006
Entropy 4-gram	0.6480	0.0005	0.2378*	0.0386	0.1354	0.0996
Entropy 5-gram	0.6326	0.0007	0.2281*	0.0475	0.1311	0.1111
Sentence length	−0.4607	0.0205	−0.4943	0.0000	−0.1676	0.0410
	c = 25		c = 76		c = ∞	
Raw data	r	p	r	p	r	p
Entropy 1-gram	−0.0023*	0.8781	0.0598*	0.0000	0.0301*	0.0110
Entropy 2-gram	0.0414	0.0056	0.0755*	0.0000	0.0615*	0.0000
Entropy 3-gram	0.0598	0.0001	0.0814*	0.0000	0.0706*	0.0000
Entropy 4-gram	0.0625	0.0000	0.0830*	0.0000	0.0712*	0.0000
Entropy 5-gram	0.0600	0.0001	0.0812*	0.0000	0.0695*	0.0000
Sentence length	−0.0635	0.0000	−0.1099	0.0000	−0.1038	0.0000

Table 1: Results of Experiment 1: correlation of sentence entropy and sentence position, on binned data; c: cut-off, r: correlation coefficient, p: significance level, *: correlation significantly different from baseline (sentence length)

available per position. Based on data this sparse, no reliable correlation coefficients can be expected.

Let us now turn to Table 1, which displays the results that were obtained by computing correlation coefficients on the raw data, i.e., without computing the mean entropy for all sentences with the same position. We find that for all parameter settings a significant correlation between sentence entropy and sentence position is obtained (with the exception of $n = 1$, $c = 25$). The correlation coefficients are significantly lower than the ones obtained using binning, the highest coefficient is 0.0830. This means that a small but reliable entropy effect can be observed even on the raw data, i.e., for individual sentences rather than for bins of sentences with the same position.

However, the results in Table 1 also confirm our findings regarding the baseline model (simple sentence length): in all cases the correlation coefficient achieved for the baseline is higher than the one achieved by the entropy models, in some cases even significantly so.

3.3.2 Disconfounding Entropy and Sentence Length

Taken together, the results in Table 1 seem to indicate that the entropy rate effect reported by G&C is not really an effect of entropy, but just an effect of sentence length. The effect seems to be due to the fact that G&C compute entropy rate by dividing the entropy of a sentence by its length: sentence length is correlated with sentence position, hence entropy rate will be correlated with position as well.

It is therefore necessary to conduct additional analyses that remove the confound of sentence length. This can be achieved by computing partial correlations; the partial correlation coefficient between a factor 1 and a factor 2 expresses the degree of association between the factors that is left once the influence of a third factor has been removed from both factors. For example, we can compute the correlation of position and entropy, with sentence length partialled out. This will tell us use the amount of association between position and entropy that is left once the influence of length has been removed from both position and entropy.

Table 2 shows the results of partial correlation analyses for length and entropy. Note that these results were obtained using total entropy, not per-word entropy, i.e., the normalizing term $\frac{1}{|X|}$ was dropped from (1). The partial correlations are only reported for the trigram model.

The results indicate that entropy is a significant predictor sentence position, even once sentence length has been partialled out. This result holds for both the binned data and the raw data, and for all cut-offs (with the exception of $c = 76$ for the binned data). Note however, that entropy is always a worse predictor than sentence length; the absolute value of the correlation coefficient is always lower. This indicates that the entropy rate effect is a much weaker effect than the results presented by G&C suggest.

4 Entropy Rate Effect and Processing Effort

The previous experiment confirmed the validity of the entropy rate effect: it demonstrated a signifi-

	$c = 25$		$c = 76$		$c = \infty$	
Binned data	r	p	r	p	r	p
Entropy 3-gram	0.6708	0.0000	0.1473	0.2067	0.1703	0.0383
Sentence length	-0.7435	0.0000	-0.3020	0.0084	-0.2131	0.0093
Raw data	r	p	r	p	r	p
Entropy 3-gram	0.0784	0.0000	0.0929	0.0000	0.0810	0.0000
Sentence length	-0.0983	0.0000	-0.1311	0.0000	-0.1176	0.0000

Table 2: Results of Experiment 1: correlation of entropy and sentence length with sentence position, with the other factor partialled out

cant correlation between sentence entropy and sentence position, even when sentence length, which was shown to be a confounding factor, was controlled for. The effect, however, was smaller than claimed by G&C, in particular when applied to individual sentences, as opposed to means obtained for sentences at the same position.

In the present experiment, we will test a crucial aspect of the entropy rate principle, viz., that entropy should correlate with processing effort. We will test this using a corpus of newspaper text that is annotated with eye-tracking data. Eye-tracking measures of reading time are generally thought to reflect the amount of cognitive effort that is required for the processing of a given word or sentence.

A second prediction of the entropy rate principle is that sentences with higher position should be harder to process than sentences with lower position. This relationship should hold out of context, but not in context (see Section 2).

4.1 Method

4.1.1 Materials

As a test corpus, we used the Embra corpus (Mc-Donald and Shillcock, 2003). This corpus consists of 10 articles from Scottish and UK national broadsheet newspapers. The excerpts cover a wide range of topics; they are slightly edited to make them compatible with eye-tracking.[2] The length of the articles varies between 97 and 405 words, the total size of the corpus is 2,262 words (125 sentences). Twenty-three native speakers of English read all 10 articles while their eye-movements were recorded using a Dual-Purkinje Image eye-tracker. To make sure that subjects read the texts carefully, comprehension questions were also administered. For details on method used to create the Embra corpus, see McDonald and Shillcock (2003).

The training and development sets for this experiment were compiled so as to match the test corpus in terms of genre. This was achieved by selecting

---

[2]This includes, e.g., the removal of quotation marks and brackets, which can disrupt the eye-movement record.

all files from the British National Corpus (Burnard, 1995) that originate from UK national or regional broadsheet newspapers. This subset of the BNC was divided into a 90% training set and a 10% development set. This resulted in a training set consisting of 6,729,104 words (30,284 sentences), and a development set consisting of 746,717 words (34,269 sentences). The development set will be used to test if the entropy rate effect holds on this new corpus.

The sentence positions in the test set varied between one and 24, in the development, they varied between one and 206.

4.1.2 Procedure

To compute per-word entropy, we trained n-gram models on the training set using the CMU-Cambridge language modeling toolkit, with the same parameters as in Experiment 1. Again, n was varied from 1 to 5. We determined the correlation between per-word entropy and sentence position for both the development set (derived from the BNC) and for the test set (the Embra corpus).

Then, we investigated the predictions of G&C's entropy rate principle by correlating the position and entropy of a sentence with its reading time in the Embra corpus.

The reading measure used was total reading time, i.e., the total time it takes a subject to read a sentence; this includes second fixations and re-fixations of words. We also experimented with other reading measures such as gaze duration, first fixation time, second fixation time, regression duration, and skipping probability. However, the results obtained with these measures were similar to the ones obtained with total reading time, and will not be reported here.

Total reading time is trivially correlated with sentence length (longer sentences taker longer to read). Hence we normalized total reading time by sentence length, i.e., by multiplying with the factor $\frac{1}{|X|}$, also used in the computation of per-word entropy. It is also well-known that reading time is correlated with two other factors: word length and word frequency; shorter and more frequent words take

	$c = 25$		$c = 76$		$c = \infty$	
Binned data	r	p	r	p	r	p
Entropy 1-gram	−0.5495	0.0044	−0.2510	0.0287	0.0232	0.7419
Entropy 2-gram	0.0602	0.7751	0.4249	0.0001	0.0392	0.5773
Entropy 3-gram	0.4523	0.0232	0.5395	0.0000	0.1238	0.0776
Entropy 4-gram	0.4828	0.0145	0.5676	0.0000	0.1229	0.0800
Entropy 5-gram	0.4834	0.0144	0.5723	0.0000	0.1223	0.0813
Sentence length	−0.8584	0.0000	−0.2947	0.0098	−0.2161	0.0019
Raw data	r	p	r	p	r	p
Entropy 1-gram	−0.0636	0.0000	−0.0543	0.0000	0.0364	0.0000
Entropy 2-gram	−0.0069	0.2783	0.0435	0.0000	0.0477	0.0000
Entropy 3-gram	0.0162	0.0103	0.0659	0.0000	0.0687	0.0000
Entropy 4-gram	0.0193	0.0022	0.0691	0.0000	0.0711	0.0000
Entropy 5-gram	0.0192	0.0024	0.0685	0.0000	0.0707	0.0000
Sentence length	−0.0747	0.0000	−0.1027	0.0000	−0.0913	0.0000

Table 3: Results of Experiment 2: correlation of sentence entropy and sentence position on the BNC

less time to read (Just and Carpenter, 1980). We removed these confounding factors by conducting multiple regression analyses involving word length, word frequency, and the predictor variable (entropy or sentence position). The aim was to establish if there is a significant effect of entropy or sentence length, even when the other factors are controlled for. Word frequency was estimated using the unigram model trained on the training corpus.

In the eye-tracking literature, it is generally recommended to run regression analyses on the reading times collected from individual subjects. In other words, it is not good practice to compute regressions on average reading times, as this fails take between-subject variation in reading behavior into account, and leads to inflated correlation coefficients. We therefore followed the recommendations of Lorch and Myers (1990) for computing regressions without averaging over subjects (see also McDonald and Shillcock (2003) for details on this procedure).

4.2 Results

Table 3 shows the results of the correlation analyses on the development set. These results were obtained after excluding all sentences at positions 1 and 2. In the newspaper texts in the BNC, these positions have a special function: position 1 contains the title, and position 2 contains the name of the author. The first sentence of the text is therefore on position 3 (unlike in the Penn Treebank, in which no title or author information is included and texts start at position 1).

We then conducted the same correlation analyses on the test set, i.e., on the Embra eye-tracking corpus. The results are tabulated in Table 4. Note we set no threshold for sentence position in the test set, as

	Binned data		Raw data	
	r	p	r	p
1-gram	−0.6505	0.0006	−0.2087	0.0195
2-gram	−0.3471	0.0965	−0.1498	0.0954
3-gram	−0.5512	0.0052	−0.1674	0.0620
4-gram	−0.5824	0.0028	−0.1932	0.0308
5-gram	−0.5750	0.0033	−0.1919	0.0320
Length	0.3902	0.0594	0.0885	0.3264

Table 4: Results of Experiment 2: correlation of sentence entropy and sentence position on the Embra corpus

	r	p
Entropy 2-gram	0.1551	0.007
Entropy 3-gram	0.1646	0.000
Entropy 4-gram	0.1650	0.000
Entropy 5-gram	0.1648	0.000
Sentence position	−0.0266	0.564

Table 5: Results of Experiment 2: correlation of reading times with sentence entropy and sentence position

the maximum article length in this corpus was only 24 sentences.

Finally, we investigated if the total reading times in the Embra corpus are correlated with sentence position and entropy. We computed regression analysis that partialled out word length, word frequency, and subject effects as recommended by Lorch and Myers (1990). All variables other than position were normalized by sentence length. Table 5 lists the resulting correlation coefficients. Note that no binning was carried out here. Figure 3 plots one of the correlations for illustration.
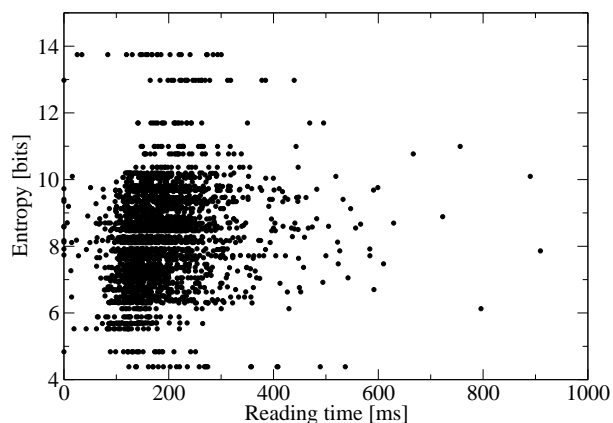
Figure 3: Experiment 2: correlation of reading times and sentence entropy (3-grams)

4.3 Discussion

The results in Table 3 confirm that the results obtained on the Penn Treebank also hold for the newspaper part of the BNC. The top half of the table lists the correlation coefficients for the binned data. We find a significant correlation between sentence position and entropy for the cut-off values 25 and 76. In both cases, there is also a significant correlation with sentence length; this correlation is particularly high (-0.8584) for $c = 25$. The entropy rate effect does not seem to hold if there is no cut-off; here, we fail to find a significant correlation (though the correlation with length is again significant). This is probably explained by the fact that the BNC test set contains sentences with a maximum position of 206, and data for these high sentence positions is very sparse.

The lower half of Table 3 confirms another result from Experiment 1: there is generally a low, but significant correlation between entropy and position, even if the correlation is computed for individual sentences rather than for bins of sentences with the same position. Furthermore, we find that sentence length is again a significant predictor of sentence position, even on the raw data. This is in line with the results of Experiment 1.

Table 4 lists the results obtained on the test set (i.e., the Embra corpus). Note that no cut-off was applied here, as the maximum sentence position in this set is only 24. Both on the binned data and on the raw data, we find significant correlations between sentence position and both entropy and sentence length. However, compared to the results on the BNC, the signs of the correlations are inverted: there is a significant negative correlation between position and entropy, and a significant positive correlation between position and length. It seems that

the Embra corpus is peculiar in that longer sentences appear later in the text, rather than earlier. This is at odds with what we found on the Penn Treebank and on the BNC. Note that the positive correlation of position and length explains the negative correlation of position and entropy: length enters into the entropy calculation as $\frac{1}{|X|}$, hence a high $|X|$ will lead to low entropy, and vice versa.

We have no immediate explanation for the inversion of the relationship between position and length in the Embra corpus; it might be an idiosyncrasy of this corpus (note that the texts were specifically picked for eye-tracking, and are unlikely to be a random sample; they are also shorter than usual newspaper texts). Note in particular that the Embra corpus is not a subset of the BNC (although it was sampled from UK broadsheet newspapers, and hence should be similar to our development and training corpora).

Let us now turn to Table 5, which lists the results of the analyses correlating the total reading time for a sentence with its position and its entropy (derived from n-grams with $n = 2, \ldots, 5$). Note that these correlation analyses were conducted by partialling out word length and word frequency, which are well-know to correlate with reading times. We find that even once these factors have been controlled, there is still a significant positive correlation between entropy and reading time: sentences with higher entropy are harder to process and hence have higher reading times. This is illustrated in Figure 3 for one of the correlations. As we argued in Section 2, this relationship between entropy and processing effort is a crucial prerequisite of the entropy rate principle. The increase of entropy with sentence position observed by G&C (and in our Experiment 1) only makes sense if increased entropy corresponds to increased processing difficulty (e.g., to increased reading time). Note that this result is compatible with previous research by McDonald and Shillcock (2003), who demonstrate a correlation between reading time measures and bigram probability (though their analysis is on the word level, not on the sentence level).

The second main finding in Table 5 is that there is no significant correlation between sentence position and reading time. As we argued in Section 2, this is predicted by the entropy rate principle: the optimal way to send information is at a constant rate. In other words, speakers should produce sentences with constant informativeness, which means that if context is taken into account, all sentences should be equally difficult to process, no matter which position they are at. This manifests itself in the absence

of a correlation between position and reading time in the eye-tracking corpus.

5 Conclusions

This paper made a contribution to the understanding of the entropy rate principle, first proposed by Genzel and Charniak (2002). This principle predicts that the position of a sentence in a text should correlate with its entropy, defined as the sentence probability normalized by sentence length. In Experiment 1, we replicated the entropy rate effect reported by Genzel and Charniak (2002, 2003) and showed that it generalizes to a larger range of sentence positions and also holds for individual sentences, not just averaged over all sentences with the same position. However, we also found that a simple baseline model based on sentence length achieves a correlation with sentence position. In many cases, there was no significant difference between the entropy rate model and the baseline. This raises the possibility that the entropy rate effect is simply an artifact of the way entropy rate is computed, which involves sentence length as a normalizing factor. However, using partial correlation analysis, we were able to show that entropy is a significant predictor of sentence position, even when sentence length is controlled.

In Experiment 2, we tested a number of important predictions of the entropy rate principle for human sentence processing. First, we replicated the entropy rate effect on a different corpus, a subset of the BNC restricted to newspaper text. We found essentially the same pattern as in Experiment 1. Using a corpus of eye-tracking data, we showed that entropy is correlated with processing difficulty, as measured by reading times in the eye-movement record. This confirms an important assumption that underlies the entropy rate principle. As the eye-tracking corpus we used was a corpus of connected sentences, it enabled us to also test another prediction of the entropy rate principle: in context, all sentences should be equally difficult to process, as speakers generate sentences with constant informativeness. This means that no correlation between sentence position and reading times was expected, which is what we found.

Another important prediction of the entropy rate principle remains to be evaluated in future work: for out-of-context sentences, there should be a correlation between sentence position and processing effort. This prediction can be tested by obtaining reading times for sentences sampled from a corpus and read by experimental subjects in isolation.

References

Aylett, Matthew. 1999. Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration. In *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco.

Burnard, Lou. 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

Clarkson, Philip R. and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of Eurospeech*. ESCA, Rhodes, Greece, pages 2707–2710.

Genzel, Dmitriy and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, pages 199–206.

Genzel, Dmitriy and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In Michael Collins and Mark Steedman, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Sapporo, pages 65–72.

Just, Marcel A. and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review* 87:329–354.

Kuhn, Roland and Renato de Mori. 1990. A cache-based natural language model for speech reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(6):570–583.

Lorch, Robert F. and Jerome L. Myers. 1990. Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(1):149–157.

McDonald, Scott A. and Richard C. Shillcock. 2003. Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research* 43:1735–1751.