The Complexity and Entropy of Literary Styles*

I. Kontoyiannis†

NSF Technical Report No. 97,
Department of Statistics, Stanford University
June 1996/October 1997

Abstract – Since Shannon's original experiment in 1951, several methods have been applied to the problem of determining the entropy of English text. These methods were based either on prediction by human subjects, or on computer-implemented parametric models for the data, of a certain Markov order. We ask why computer-based experiments almost always yield much higher entropy estimates than the ones produced by humans. We argue that there are two main reasons for this discrepancy. First, the long-range correlations of English text are not captured by Markovian models and, second, computer-based models only take advantage of the text statistics without being able to "understand" the contextual structure and the semantics of the given text.

The second question we address is what does the "entropy" of a text say about the author's literary style. In particular, is there an intuitive notion of "complexity of style" that is captured by the entropy?

We present preliminary results based on a *non*-parametric entropy estimation algorithm that offer partial answers to these questions. These results indicate that taking long-range correlations into account significantly improves the entropy estimates. We get an estimate of 1.77 bits-per-character for a one-million-character sample taken from Jane Austen's works. Also comparing the estimates obtained from several different texts provides some insight into the interpretation of the notion of "entropy" when applied to English text rather than to random processes, and the relationship between the entropy and the "literary complexity" of an author's style.

Advantages of this entropy estimation method are that it does not require prior training, it is uniformly good over different styles and languages, and it seems to converge reasonably fast.

---

*This paper was submitted as a term-project for the "Special Topics in Information Theory" class EE478 taught by Prof. Tom Cover (EE Dept., Stanford Univ.) during Spring 1996. This work was supported in part by grants NSF #NCR-9205663, JSEP #DAAH04-94-G-0058, ARPA #J-FBI-94-218-2.

†I. Kontoyiannis is with the Information Systems Laboratory Durand Bldg 141A, Stanford University, Stanford CA 94305. Email: `yiannis@isl.stanford.edu`

1 Introduction

The purpose of this note is to address the following two questions.

1. What is it that makes humans so much more efficient at estimating entropy than machines?

Shannon in 1951 [25] devised an experimental method for determining the entropy of a piece of text that was based on human subjects predicting the next character after having seen the preceding text. Using this method he estimated the "entropy of English" to be between 0.6 and 1.3 bits-per-character (bpc). This method was modified by Cover and King in 1978 [7] who asked their subjects to gamble on the next symbol outcome. Their method produced sharper estimates between 1.25 and 1.35 bpc.

But why use human subjects? An obvious method for entropy estimation is the following: Run an efficient compression algorithm on the data and calculate the compression ratio. With the great development of compression algorithms over the past 20 years and the tremendous advances in computer technology, one would expect that we should be able to get more efficient machine-based entropy estimates than the ones produced by humans. Experience shows it is not so. In 1992, Brown et al. [2] used a word trigram language model and a corpus of more than 500 million words to get an estimate of 1.75 bpc. More recently, Teahan and Cleary [26] in 1996 used a modification of a PPM-based arithmetic coding scheme to obtain estimates between 1.46 and 1.48 bpc.

Cover and Kings's result has remained, for the past 18 years, the benchmark for evaluating the performance of new compression algorithms.

2. What does entropy say about the "complexity" of language?

In other words, which facets of the notion of literary complexity does the entropy capture? We first need to be a little more specific about the meaning of the phrase "entropy of language" or "entropy of English." In information theory, the notion of entropy has a clear operational interpretation. It is the optimum compression ratio one can hope to achieve, on the average, over long messages emitted by a stationary source. It is the smallest number of bits per character required to describe the message in a uniquely decodable way. So entropy characterizes the redundancy of a source. Therefore different texts should be regarded as different sources and assigned different entropies in the same way that different authors have different styles, some of higher and some of lower

complexity than others.

Which brings us to the interpretation of entropy as a measure of complexity. The smaller the redundancy of the text the harder it is to predict it, and the more complex it seems. This connection between entropy and complexity was rigorously formulated by Kolmogorov in 1965 [12] and is discussed in some detail in [6] and [8]. But does entropy capture, in an intuitive way, any aspects of the complexity of an author's style in the literary sense? The entropy-complexity analogy is a very appealing one, and one that is easy to occasionally carry a little too far.

1.1 A nonparametric method for entropy estimation

We offer the following partial answer to the first question. The main problem of machine-implemented algorithms has been the fact that they are almost always based on parametric Markov models of the English language. It seems to be a well-understood fact that, as already argued by Chomsky 40 years ago [4], Markovian models are not adequate linguistic descriptions for natural languages. From our point of view (that of entropy estimation), one obvious deficiency of Markov models is that they have bounded context-depths and thus cannot capture the strong long-range dependencies encountered in written English. This motivates us to look for efficient nonparametric entropy estimation algorithms.

In this report we use nonparametric entropy estimators based on string matching [15] [13], that are closely related to the celebrated Lempel-Ziv data compression algorithms [31] [32], and are similar to methods that have been used in DNA sequence analysis [3], [9]. We apply two of these estimators that are described in detail in Section 4 below. Here we briefly describe how one of the two estimators works.

We model text as a string produced by a stationary process $\boldsymbol{X} = \{\ldots, X_{-1}, X_0, X_1, X_2, \ldots\}$, with each X_i taking values in a finite alphabet A (this assumption will remain in effect throughout). Suppose we are given a long realization of the process starting at time zero: $x_0 x_1 \ldots x_M$. For each position $i \geq 1$ of the "text" $x_0 x_1 \ldots x_M$ we calculate the length of the shortest prefix starting at x_i, that does not appear starting anywhere in the previous i symbols $x_0 x_1 \ldots x_{i-1}$, and denote this length by l_i. (We allow the possibility that there is overlap between the prefix starting at x_i and the matching string starting somewhere in $x_0 x_1 \ldots x_{i-1}$.) For example, if we are given the binary

string

$$01100101011001001$$

then for $i = 5$ we get $l_5 = 3$:

$$01100\underbrace{101}_{l_5=3}011001001.$$

The first entropy estimator [15] is given by the formula

$$\hat{H}_N = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{l_i}{\log(i+1)}\right]^{-1}, \tag{1}$$

for some $N < M$ (logarithms here and throughout this report are taken to base two). Motivation for this formula will be given in Section 4 below. The prefix-length l_i can be thought of as the length of the next phrase to be encoded by the Lempel-Ziv algorithm, after the past up to time $(i-1)$ has been encoded. Since, as i grows, there is no restriction on how far into the past we can look for a long match, the estimator (1) can take into account very long range correlations in the text, and this, in view of the discussion following our first question above, partly explains our motivation for introducing this method.

With respect to our second question now, it is interesting to recall the notion of Lempel-Ziv complexity which was originally introduced in conjunction with studying the complexity of finite strings [16]. There, a measure of complexity for finite strings was introduced based on Lempel-Ziv parsing, and its properties were discussed and compared with those of other complexity measures. This encourages us to interpret the estimate (1), for finite values of N, as a complexity measure of the author's style for a given text, although some remarks are in order here about what exactly we mean by that.

1.2 The complexity of style

We applied estimator (1) to several different English texts including the King James Bible, a concatenation of four novels by Jane Austen (*Sense and Sensibility*, *Northanger Abbey*, *Persuasion* and *Mansfield Park*) and two novels by James Joyce (*Ulysses* and *Portrait of the Artist as a Young Man*). We chose these texts because they are stylistically very different, and each one is representative of a different category. It is a well-known linguistic fact that, excluding proper

names, there are only about 500 roots in the Bible. This means that there is a lot of repetition and therefore considerably high redundancy, i.e. low descriptional complexity. But there is also a definite sense in which the bible is a very complex piece of writing. The thought process that is described is found (by many people) quite deep, and although the language is simple and easy to read, the meaning of the text is rather complex and occasionally hard to follow.

James Joyce, on the other, hand seems hard to read. The style does not seem to strictly follow any definite syntactic rules, it is not restricted to the standard English vocabulary, and is not easy to follow. In this sense James Joyce's writing is more complex that the Bible, but it may be that the Bible is more complex in what it means for a reader, given the context within which it is interpreted. Finally Jane Austen was chosen as a representative author of the standard 19th century heavy literary style.

We run estimator (1) on a 500,000-character piece of each one of the above three texts. The results are shown in the table below.

text	Entropy Estimate	number of chrs.
Bible	0.92 bpc	500,000
J. Austen	1.78 bpc	500,000
J. Joyce	2.12 bpc	500,000

These results agree well with our intuition that entropy captures statistical structure and descriptional complexity, but not the complexity that comes from the actual contextual and semantic meaning of the text.

1.3 Organization

The rest of this report is organized as follows: In the next section we outline the history of the problem of entropy estimation. In Section 3 we discuss our first question, namely the differences between humans and computers estimating entropy. Section 4 contains a description of our entropy estimator, and Section 5 a discussion of our results.

2 Some History

In his landmark 1948 paper, [24], Shannon defined the entropy and the redundancy of a language, and he presented a series of approximations to written English based on finite-order Markov chains. He considered zeroth to third order letter approximations and first and second order word approximations, and based on these he computed an entropy estimate of 2.3 bpc for 27-character English (26 letters plus space). Three years later, Shannon [25] came up with a more sophisticated experimental procedure for determining the entropy of a given text. Instead of postulating a Markovian model for the data and then estimating the relevant model parameters, the new method utilized the knowledge of the language statistics possessed by those who speak it, and was based on human subjects predicting the next letter after having seen the preceding text. Experiments on an excerpt from Dumas Malone's *Jefferson the Virginian* [17] produced upper and lower bounds of 0.6 and 1.3 bpc, respectively. It is worth noting here that the transition from using a method based on a Markov model for the data to a method that uses humans for prediction can be thought of as a transition from a parametric to a nonparametric family of models.

Shannon's estimates [25] were calculated on the basis of how many guesses it took the subject to correctly identify the next letter after having seen the preceding text. This method was improved upon by Cover and King [7] in 1978. They asked the subjects to not only try to predict the following character, but at the same time identify the probability with which they thought their guess was correct. This was done by placing sequential bets on the next symbol occurrence. The subjects gambled on a sentence extracted from [17] and produced entropy estimates between 1.29 and 1.90 bpc. A plausible method for combining these estimates was employed by Cover and King which produced overall results between 1.25 and 1.35 bpc. These seem to be the most reliable entropy estimates to date.

Now for the computer-based methods. The obvious way to obtain an upper bound for the entropy of a given piece of text is to run a compression algorithm on it and then calculate the compression ratio. The most successful methods, that is, the methods producing the lowest estimates, have been the ones based on the "Prediction by Partial Matching" (PPM) variation of the arithmetic coding algorithm – see [1] for a comprehensive discussion of PPM-related methods. The first experiments [19] [1] that used initially empty contexts (and no training) reported results

between 2.19 and 2.40 bpc. In a recent paper, Teahan and Cleary [26] used statistical preprocessing of the text and a training method to obtain estimates as low as 1.46bpc.

Finally, Brown et al. [2] used a large corpus of more than 500 million words to create a word trigram language model which produced an estimate of 1.75 bpc. As pointed out in [26] part of the reason why Brown et al.'s estimate is not as high as one would hope for given the amount of data they used, is that instead of 27-letter English they did their experiments using all 96 printable ASCII characters.

Several other methods have appeared in the literature since Shannon's original papers, where they are applied to specific problems in many areas outside information theory such as psychology, education, music, linguistics. Newman and Waugh [21] in 1960 used a statistical method to study the differences in entropy across languages. They came up with estimates between 2.4 to 2.8 bpc for different English texts, but their method was not fully rigorously justified and has not been widely used. Jamison and Jamison in 1968 [10] used Shannon's method of guessing to get an estimate of 1.65 bpc and they pointed out a connection between entropy and partial knowledge of languages in linguistics. The use of information theoretic methods in the study of language has been studied extensively by Mandelbrot [18], Chomsky [4], Newman [20], Yaglom, Dobrushin and Yaglom [30] and Paisley [23], among many others. The texts [29], [27] [1], and the paper [7] contain extensive bibliographies on the subject.

3 Machines vs Humans

How come humans, using just a few characters, can estimate entropy so much more accurately than powerful computers using hundreds of millions of words? It seems that humans not only have a very good knowledge of the statistics of language, they can also extract information from the context and the semantics of the given passage. We can guess the answers to rhetorical questions, we can predict the characters' psychological reactions in novels. Humans are good at keeping track of long-range dependencies within long texts. Typically, machine-based language models keep track of contexts of at most a few characters long, whereas humans can easily follow book plots that are hundreds of pages long. If, for example, the first two words in a book are "The Bible" we are pretty certain that we will not come across the phrase "non-differential manifolds" somewhere

in the middle of the book, whereas if we were merely looking at statistics, this phrase would have a definite positive probability.

According to Tom Cover [5] "When a computer can compress text as well as the best human can, we should say that the machine *understands* the text." Computers, on the other hand, can calculate and store statistics extremely accurately. Given the current state of computer technology computers can have much larger memory and process the information stored much faster than a human could. Based on these considerations, Teahan and Cleary [26] remark that "There is no reason that machine models cannot do better than humans."

It was already argued in the Introduction and it seems to be a generally accepted fact that there exists a part of natural language, corresponding to the intuitive notion of understanding the content of a phrase, that is not captured by purely statistical models. In view of this there are two questions that remain unanswered:

1. How big is the entropy content of that part of the language? How much can *understanding* the message reduce the length of its most efficient description?

2. Humans are good at "understanding" but machines are extremely efficient at keeping and processing statistics. What is the nature of this tradeoff? Will computers eventually strictly dominate human performance as memory and processing time become faster and cheaper?

4 The Method

In this section we describe two recent entropy estimators, (2) and (3) below [15] [13] [9].

Consider a stationary process $\boldsymbol{X} = \{\ldots, X_{-1}, X_0, X_1, X_2, \ldots\}$ with values in the finite alphabet A. We denote an infinite realization of the process by $x = (\ldots, x_{-1}, x_0, x_1, x_2, \ldots)$ and for $i \leq j$, x_i^j denotes the string $(x_i, x_{i+1}, \ldots, x_j)$. In 1989 Wyner and Ziv discovered the following interesting connection between match lengths along a realization of the process and its entropy rate H. Given a realization x and an integer $N \geq 1$, we let L_N denote the length of the shortest string starting at x_0 that does *not* appear starting anywhere within the past N symbols x_{-N}^{-1}, where we allow for overlap between the two matching strings. Equivalently, L_N can be thought of as the longest

match length plus one:

$$L_N = L_N(x) = \max\{k \geq 0 \,:\, x_0^{k-1} = x_j^{j+k-1} \text{ for some } -N \leq j \leq -1\}.$$

Wyner and Ziv [28] conjectured that, as the size N of our "database" grows to infinity, L_N will grow logarithmically with slope equal to $1/H$:

$$\frac{L_N}{\log N} \to \frac{1}{H} \quad \text{a.s.}$$

Ornstein and Weiss [22] formally established this result. One would hope to use this in practice to estimate the entropy, but simulations show that the convergence is very slow. Also, in an intuitive sense, we do not seem to be making very efficient use of our data. So we suggest the following modification. We fix a large integer N as the size of our database. Given a realization x and some fixed time instant i, instead of looking for matches starting at time zero we look at the length of the shortest string that starts at time i and does not appear starting anywhere in the previous N symbols X_{i-N}^{i-1}. If we call that length $\Lambda_i^N = \Lambda_i^n(x)$ then it is clear that $\Lambda_i^N(x) = L_N(T^i x)$, where T is the usual shift operator. Therefore the stationarity of $\boldsymbol{X}$ immediately implies that, for any fixed index i, $\Lambda_i^N / \log N$ will converge to $1/H$ with probability one.

Let us remark in passing that Λ_i^N can be interpreted as the length of the next phrase to be encoded by the sliding-window Lempel-Ziv algorithm [31], when the window size is N. Similarly Λ_i^i can be thought of as length of the phrase that would be encoded next by the Lempel-Ziv algorithm [32] with knowledge of the past x_0^{i-1}. Observe that Λ_i^i is the quantity we called l_i in the introduction.

Following [9] [15] [13], in order to make more efficient use of the data we suggest that instead of just looking at the match length Λ_i^N at just one position i, we look at several positions $i = 1, 2, \ldots, M$ and we average these estimates out. This can be done by either sliding a window of size N behind the current position,

$$\frac{1}{M} \sum_{i=1}^{M} \frac{\Lambda_i^N}{\log N},$$

or by considering a window of growing size,

$$\frac{1}{M} \sum_{i=1}^{M} \frac{\Lambda_i^i}{\log(i+1)}.$$

The following is proved in [13]:

Theorem. *Let $\boldsymbol{X}$ be a stationary ergodic process with entropy rate H and let N grow linearly with M as $M \to \infty$. Then*

$$\frac{1}{M} \sum_{1 \le i \le M} \frac{\Lambda_i^N}{\log N} \stackrel{M\to\infty}{\longrightarrow} \frac{1}{H}, \quad \text{a.s. and in } L^1, \tag{2}$$

$$\frac{1}{M} \sum_{1 \le i \le M} \frac{\Lambda_i^i}{\log(i+1)} \stackrel{M\to\infty}{\longrightarrow} \frac{1}{H} \quad \text{a.s. and in } L^1, \tag{3}$$

provided the following mixing condition holds:

Doeblin Condition (DC): *There exists an integer $r \ge 1$ and a real number $\beta \in (0,1)$ such that with probability one,*

$$P\{X_0 = x_0 | X_{-\infty}^{-r}\} \ge \beta, \quad \text{for all } x_0 \in A.$$

We interpret (DC) as saying that "if we wait for r time-steps, anything can happen with non-zero probability." Or, to be a little more precise, "conditional on any realization of the infinite past from time $-\infty$ to $-r$, any symbol can occur at time zero with strictly positive probability." For the case of English text we believe that the validity of (DC) is a plausible assumption. In the context of string matching problems (DC) was first introduced by Kontoyiannis and Suhov [14], where its properties are discussed in greater detail.

5 Results

In this section we present preliminary results based on our experiments using the estimators described in the previous section.

5.1 Texts

Below we present the results of the sliding-window type estimator (2) applied to the three texts described in the introduction. The window length is equal to $(1/2)\times$(#of characters $-$ 500).

text	Entropy Estimate	number of chrs.
Bible	1.46 bpc	10,000
	1.31 bpc	20,000
	1.10 bpc	400,000
	0.92 bpc	**800,000**
J. Austen	2.00 bpc	10,000
	1.87 bpc	100,000
	1.78 bpc	400,000
	1.77 bpc	**1,000,000**
J. Joyce	2.21 bpc	100,000
	2.11 bpc	200,000
	2.12 bpc	400,000
	2.15 bpc	**1,000,000**

5.2 Sentences

In human-based experiments, the subjects are typically asked to predict or gamble on a piece of text no longer than two or three hundred characters long. So in a sense the quantity that is calculated is the entropy of the sentence examined, given the subjects' knowledge of the preceding text and of the language it is written in. Cover and King in their experiments [7] used a 75 character long sequence from [17] that was contained in a one sentence. This brings up the question about how typical was the sentence picked by Shannon or by Cover and King (this question is discussed in [11] and [26]).

We modified our estimator (2) to examine just one sentence at a time, using a large window size. In (2) we take position $i = 1$ to be the beginning of a sentence, let N large (window length) and $M = 75$ to be the length of the part of the sentence to be analyzed. Below are the results we obtained on sentences taken from Jane Austen's works:

sentence	Entropy Estimate	Window length
1	1.25 bpc	2310598
2	1.37 bpc	2361100
3	1.49 bpc	2319260
4	1.52 bpc	2362203
5	1.84 bpc	2360000

– **sentence 1:** "good heavens charles how can you think of such a thing take a box for tomor"

– **sentence 2:** " of sir william it cannot be doubted that sir walter and elizabeth were sho"

– **sentence 3:** " could imagine she read there the consciousness of having by some complicat"

– **sentence 4:** "she had but two friends in the world to add to his list lady russell and mr"

– **sentence 5:** "imself to raise even the unfounded hopes which sunk with him the news of hi"

6 Text sources

All the texts that were used for our experiments were downloaded from public domain sites on the Internet. The *King James Bible* and Jane Austen's *Mansfield Park*, *Northanger Abbey*, and *Persuasion* we got from Project Gutenberg at http://192.76.144.75/books/gutenberg; Jane Austen's *Sense and Sensibility* we got from the Educational Resources of the University of Maryland at College Park Web site at http://www.inform.umd.edu:8080/EdRes/Topic/WomenStudies/ReadingRoom; James Joyce's *Portrait of the Artist as a Young Man* and *Ulysses* we got from the Bibliomania Web site at http://www.bibliomania.com/Fiction.

References

[1] J.G. Bell, T.C. Cleary and I.H. Witten. *Text Compression*. Prentice Hall, New Jersey, 1990.

[2] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, J.C. Lai, and R.L. Mercer. An estimate of an upper bound for the entropy of English. *Computational Linguistics*, 18(1):31–40, 1992.

[3] S. Chen and J.H. Reif. Using difficulty of prediction to decrease computation: Fast sort, priority queue and convex hull on entropy bounded inputs. In *34th Sympsium on Foundations*

of Computer Science, pages 104–112, Los Alamitos, California, 1993. IEEE Computer Society Press.

[4] N. Chomsky. Three models for the description of language. *IRE Trans. Inform. Theory*, 2(3):113–124, 1956.

[5] T.M. Cover. private communication. April 1996.

[6] T.M. Cover, P. Gacs, and R.M. Gray. Kolmogorov's contributions to information theory and algorithmic complexity. *Ann. Probab.*, 17(3):840–865, 1989.

[7] T.M. Cover and R. King. A convergent gambling estimate of the entropy of English. *IEEE Trans. on Inform. Theory*, 24(4):413–421, 1978.

[8] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. J. Wiley: New York, 1991.

[9] M. Farach, et al. On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. In *Proceedings of the 1995 Sympos. on Discrete Algorithms*, 1995.

[10] D. Jamison and K. Jamison. A note on the entropy of partially-known languages. *Inform. Contr.*, 12:164–167, 1968.

[11] F. Jelinek. Self-organized language modelling for speech recognition. In *Readings in Speech Rcognition*, pages 450–506. Morgan Kaufmann Publishers Inc., 1990.

[12] A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11, 1965.

[13] I. Kontoyiannis, P.H. Algoet, and Yu.M. Suhov. Two consistent entropy etimates for stationary processes and random fields. *NSF Technical Report no. 91, Statistics Department, Stanford University*, April 1996.

[14] I. Kontoyiannis and Yu.M. Suhov. Prefixes and the entropy rate for long-range sources. Chapter in *Probability Statistics and Optimization* (F.P. Kelly, ed.). Wiley, New York, 1994.

[15] I. Kontoyiannis and Yu.M. Suhov. Stationary entropy estimation via string matching. In *Proceedings of the Data Compression Conference, DCC 96*, Snowbird, UT, 1996. IEEE Computer Society Press.

[16] A Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Trans. Inf. Theory*, 22(1):75–81, 1978.

[17] D. Malone. *Jefferson the Virginian*. Little Brown and Co., Boston, 1948.

[18] B. Mandelbrot. An informational theory of the statistical structure of language. In W. Jackson, editor, *Communication Theory*, pages 485–502. New York: Academic Press, 1953.

[19] A. Moffat. Implementing the PPM data compression scheme. *IEEE Trans. Comm.*, 38(11):1917–1921, 1990.

[20] E.B. Newman. Men and information: a psychologists view. *Nuovo Cimento Suppl.*, 13(2):539–559, 1959.

[21] E.B. Newman and N.C. Waugh. The redundancy of texts in three languages. *Inform. Contr.*, 3:141–153, 1960.

[22] D. Ornstein and B. Weiss. Entropy and data compression schemes. *IEEE Trans. Inf. Theory*, 39(1):78–83, 1993.

[23] W.J. Paisley. The effects of authorship, topic structure and time of composition on letter redundancy in English texts. *J. Verbal Learning and Verbal Behaviour*, 5(1):28–34, 1966.

[24] C.E. Shannon. A mathematical theory of communication. *Bell System Technical J.*, 27:379–423, 623–656, 1948.

[25] C.E. Shannon. Prediction and entropy of printed English. *Bell System Technical J.*, 30:50–64, 1951.

[26] W.J. Teahan and J.G. Cleary. The entropy of English using PPM-based models. In *Proceedings of the Data Compression Conference, DCC 96*, Snowbird, UT, 1996. IEEE Computer Society Press.

[27] K. Weltner. *The Measurement of Verbal Information in Psychology and Education.* Springer-Verlag, Berlin-Heidelberg-New York, 1973.

[28] A. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inf. Theory*, 35(6):1250–1258, 1989.

[29] A.M. Yaglom and I.M. Yaglom. *Probabiliity and Information.* Kluwer, Boston, 1983. Translation of: Veroiatnosti informatsiia, 3rd rev. and enl. ed., Izd-vo Nauka, Moscow.

[30] I.M. Yaglom, R.L. Dobrushin, and A.M. Yaglom. Information theory and linguistics. *Problems of Linguistics*, 1:100–110, 1960.

[31] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3):337–343, 1977.

[32] J. Ziv and A. Lempel. Compression of individual sequences by variable rate coding. *IEEE Trans. Inf. Theory*, 24(5):530–536, 1978.