*Информационные процессы, Том 6, № 3, 2006, стр. 229–236.*
© *2006 Manin.*

# Experiments on predictability of word in context and information rate in natural language

**D.Yu.Manin**

*e-mail: manin@pobox.com*

Received June 13, 2006

**Abstract**—Based on data from a large-scale experiment with human subjects, we conclude that the logarithm of probability to guess a word in context (unpredictability) depends linearly on the word length. This result holds both for poetry and prose, even though with prose, the subjects don't know the length of the omitted word. We hypothesize that this effect reflects a tendency of natural language to have an even information rate.

## 1. INTRODUCTION

In this paper we report a particular result of an experimental study on predictability of words in context. The experiment's primary motivation is the study of some aspects of poetry perception, but the result reported here is, in the author's view, of a general linguistic interest.

The first study of natural text predictability was performed by the founder of information theory, C. E. Shannon [1]. (We'll note that even in his groundbreaking work [2], Shannon briefly touched on the relationship between literary qualities and redundancy by contrasting highly redundant Basic English with Joyce's "Finnegan's Wake" which "enlarges the vocabulary and is alleged to achieve a compression of semantic content".) Shannon presented his subject with random passages from Jefferson's biography and had her guess the next letter until the correct guess was recorded. The number of guesses for each letter was then used to calculate upper and lower bounds for the entropy of English, which turned out to be between 0.6 and 1.3 bits per character (bpc), much lower than that of a random mix of the same letters. Shannon's results also indicated that conditional entropy decreases as more and more text history becomes known to the subject, up to at least 100 letters.

Several authors repeated Shannon's experiments with some modifications. Burton and Licklider [3] used 10 different texts of similar style, and fragment lengths of 1, 2, 4, ..., 128, 1000 characters. Their conclusion was that, contrary to Shannon, increasing history doesn't affect measured entropy when history length exceeds 32 characters.

Fónagy [4] compared predictability of the next letter for three types of text: poetry, newspaper, and "a conversation of two young girls". Apparently, his technique involved only one guess per letter, so entropy estimates could not be calculated (see below), and results are presented in terms of the rate of correct answers, poetry being much less predictable than both other types.

Kolmogorov reported the results of 0.9–1.4 bpc for Russian texts in his work [5] that laid the ground of algorithmic complexity theory. Unfortunately, the paper contains no experimental details.

Cover and King [6] modified Shannon's technique by having their subjects place bets on the next letter. They showed that the optimal betting policy would be to distribute available capital among the possible outcomes according to their probability and so if the subjects play in an optimal way (which is not self-evident though), the letter probabilities could be inferred from their bets. Their estimate of the entropy of English was calculated at 1.3 bpc. This work also contains an extensive bibliography.

Moradi *et al* [7] first used two different texts (a textbook on digital signal processing and a novel by Judith Krantz) to confirm Burton and Licklider's results on the critical history length (32 characters), then added two more texts ("101 Dalmatians" and a federal aviation manual) to study the dependence of entropy on text type and subject (with somewhat inconclusive results).

A number of works were devoted to estimating entropy of natural language by means of statistical analysis, without using human subjects. One of the first attempts is reported in [8], where 39 English translations of 9 classical Greek texts were used to study entropy dependency on subject matter, style, and period. A very crude entropy estimate by letter digram frequency was used. For some of the more recent developments, see [9], [10] and references therein. By the very nature of these methods they can't utilize meaning (and even syntax) of the text, but by the brute force of contemporary computers they begin achieving results that come reasonably close to those demonstrated by human language speakers.

Our experimental setup differs from the previous work in two important aspects. First, we have subjects guess whole words, and not individual characters. Second, the words to be guessed come (generally speaking) from the middle of a context, rather than at the end of a fragment. In addition to filling blanks, we present the subjects with two other task types where authenticity of a presented word is to be assessed. The reason for this is that while most of the previous studies were eventually aimed at efficient text compression, we are interested in literary (chiefly, poetic) texts as works of literature, and not as mere character strings subject to application of compression algorithms[1]. Our goal in designing the experiment was to provide researchers in the field of poetics with hard data to ground some hypotheses that otherwise are unavoidably speculative. Guessing the next word in sequence is not the best way to treat literary text, because even an ordinary sentence like this one is not essentially a linear sequence of words or characters, but a complex structure with word associations running all over the place, both forward and backward. A poem, even more so, is a structure with strongly coordinated parts, which is not read sequentially, much less written sequentially. Also, practice shows that even when guessing letter by letter, people almost always base their next character choice on a tentative word guess. This is why guessing whole words in context was more appropriate for our purpose.

However, the results we present here, as already mentioned, are not relevant to poetics proper, so we will not dwell on this further, and refer the interested reader to [11].

## 2. EXPERIMENTAL SETUP

In their Introduction to the special issue on computational linguistics using large corpora, Church and Mercer [12] note that "The 1990s have witnessed a resurgence of interest in 1950s-style empirical and statistical methods of language analysis". They attribute this empirical renaissance primarily to the availability of processing power and of massive quantities of data. Of course, these factors favor statistical analysis of texts as character strings. However, wide availability of computer networks and interactive Web technologies also made it possible to set up large-scale experiments with human subjects.

The experiment has the form of an online literary game in Russian[2]. However, the players are also fully aware of the research side, have free access to theoretical background and current experimental results, and can participate in online discussions. The players are presented with text fragments in which one of the words is replaced with blanks or with a different word. Any sequence of 5 or

[1] It should be noted though that efficient compression is important not only *per se*, but also for cryptographic applications as pointed out in [10]. In addition, language models developed for the purpose of compression are successfully used in applications like speech recognition and OCR, allowing to disambiguate difficult cases and correct errors.

[2] http://ygrec.msk.ru

more Cyrillic letters surrounded by non-letters was considered a “word”. Words are selected from fragments randomly. There are three different trial types:

type 1: a word is omitted, and is to be guessed.
type 2: a word is highlighted, and the task is to determine whether it is original or replaced.
type 3: two words are displayed, and the subject has to determine which one is the original word.

Incorrect guesses from trials of type 1 are used as replacements in trials of types 2 and 3.

Texts are randomly drawn from a corpus of 3439 fragments of mostly poetic works in a wide range of styles and periods: from Avantgarde to mass culture and from 18th century to contemporary. Three prosaic texts are also included (two classic novels, and a contemporary political essay).

As of this writing, the experiment has been running almost continuously for three years. Over 8000 people took part in it and collectively made almost 900,000 guesses, about a third of which is of type 1. The traditional laboratory experiment could have never achieved this scale. Of course, the technique has its own drawbacks, which are discussed in detail in [11]. But they are a small price to pay for statistical relevance, especially if it can’t be achieved in any other way.

## 3. RESULTS

The specific goal of the experiment is to discover and analyze systematic differences between different categories of texts from the viewpoint of how easy it is to a) reconstruct an omitted word, and b) distinguish the original word from a replacement. However here we’ll consider a particular property of the texts that turns out to be independent of the text type and so probably characterizes the language itself rather than specific texts. This property is the dependency of word unpredictability on its length.

We define *unpredictability* $U$ as the negative binary logarithm of the probability to guess a word, $U = -\log_2 p_1$, where $p_1$ is the average rate of correct answers to trials of type 1. For a single word, this is formally equivalent to Shannon’s definition of entropy, $H$. However, when multiple words are taken into account, entropy should be calculated as the average logarithm of probability, and not as the logarithm of average probability,

$$H = -\frac{1}{N}\sum_{i=1}^{N}\log_2 p_1^i$$

$$U = -\log_2 \frac{1}{N}\sum_{i=1}^{N} p_1^i$$

Indeed, the logarithm of probability to guess a word equals the amount of information in bits required to determine the word choice. Thus, it is this quantity that is subject to averaging. When dealing with experimental data, it is customary to use frequencies as estimates of unobservable probabilities. However, there are always words that were never guessed correctly and have $p_1 = 0$ for which logarithm is undefined (this is why Shannon’s techinque involves repeated guessing of the same letter until the correct answer is obtained). Formally, if there is one element in the sequence with zero (very small, in fact) probability of being guessed, then the amount of information of the whole sequence may be determined solely by this one element.

On the other hand, unpredictability as defined above is not sensitive to the exact probability to guess such words, but only on how many there are of them. While entropy characterizes the number

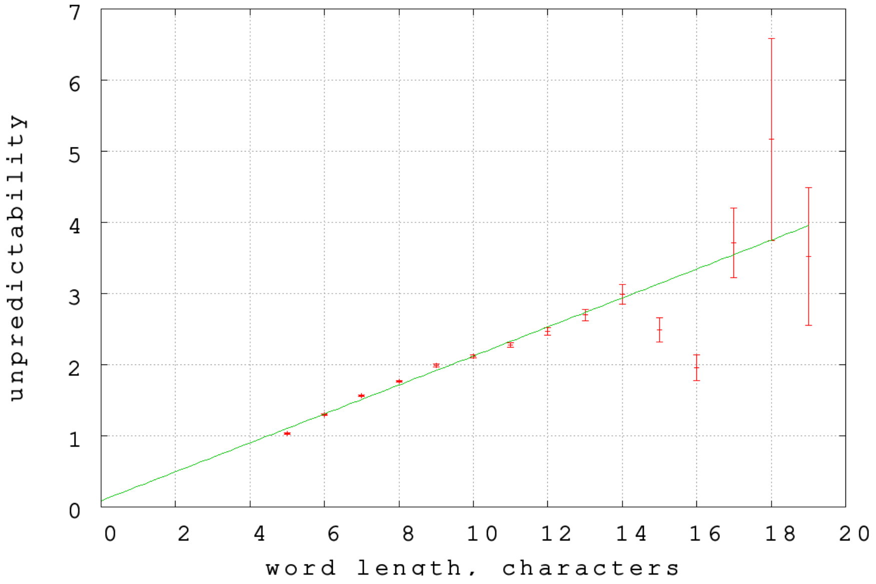

Fig. 1. Unpredictability as a function of word length in characters, all texts

of tries required to guess a randomly selected word, unpredictability characterizes the portion of words that would be guessed on the first try. They are equal, of course, if all words have the same entropy.

One way around the problem presented by never-guessed words would be to assign some arbitrary finite entropy to them. We compared unpredictability with entropy calculated under this approximation with two values of the constant: 10 bits (corresponding roughly to wild guessing using a frequency dictionary) and 3 bits (the low bound). In both cases, while $H$ is not equal numerically to $U$, they turned out to be in an almost monotonic, approximately linear correspondence. This probably means that the fraction of hard-to-guess words co-varies with unpredictability of the rest of the words. Because of this, we prefer to work in terms of unpredictability, rather than introducing arbitrary hypotheses to calculate an entropy value of dubious validity.

Unpredictability as a function of word length calculated over all words of the same length across all texts is plotted in Fig. 1 and Fig. 2 (where word length is measured in characters and syllables respectively). Confidence intervals on the graphs are calculated based on the standard deviation of the binomial distribution (since the data comes from a series of independent trials with two possible outcomes in each: a guess may be correct or incorrect).

In the range from 5 to 14 characters and from 1 to 5 syllables, an excellent linear dependence is observed. Longer words are rare, so the data for them is significantly less statistically reliable. We'll only discuss the linear dependence in the range where it is definitely valid.

## 4. DISCUSSION

It is very difficult, for the reasons mentioned above, to compare our results with previous studies. However, there are two points of comparison that can be made. First, we can roughly estimate the effect of word guessing in context as opposed to guessing the next word in sequence. Recall that Shannon [1] estimated zeroth-order word entropy for English based on Zipf's law to be 11.82 bits per word (bpw). Brown *et al* [9] used a word trigram model to achieve an entropy estimate of 1.72 bpc, which translates to 7.74 bpw for average word length of 4.5 characters in English. This means that trigram word probabilities contribute $11.82 - 7.74 = 4.08$ bpw for prediction of word in

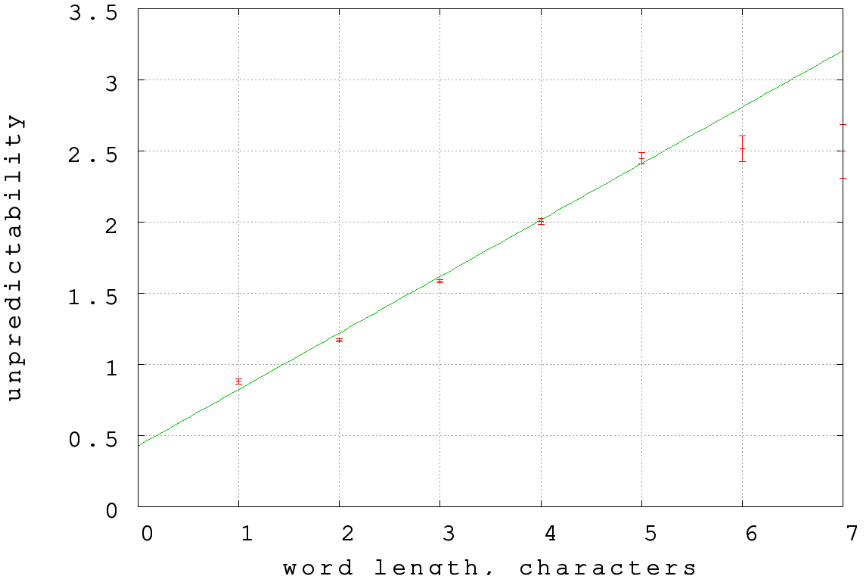

Fig. 2. Unpredictability as a function of word length in syllables, all texts

sequence. But word in context participates in three trigrams at once: as the last, the middle and the first word of a trigram. Only the first trigram is available when the model is predicting the next word, but all three trigrams could be used to fill in an omitted word (this is a hypothetical experiment which was not actually performed). Of course, they are not statistically independent, and as a rough estimate we can assume that the last trigram contributes somewhat less information than the first one, while the middle trigram contributes very little (since all of its words are already accounted for). In other words, we could expect this model to have about 4 bpw more information when guessing words in context, which is very significant.

The second point of comparison is provided by [13] (Fig. 13 there), where entropy is plotted for the $n$-th letter of each word, versus its position $n$. Entropy was estimated using a Ziv–Lempel type algorithm. It is well-known that guessing is least confident at the word boundaries for both human subjects and computer algorithms, and this chart quantifies the observation: the first letter has the entropy of 4 bpc, which drops quickly to about 0.6–0.7 bpc for the 5th letter and then stays surprizingly constant all the way through the 16th character. This chart is practically the same for the original text and a text with randomly permuted words, which gives a telling evidence of the current language models' strengths and weaknesses. For the purposes of this discussion, the data allows to reconstruct the dependency of word entropy on the word length as $h_n^{(w)} = \sum_{i=1}^n h_i^{(l)}$, where $h_n^{(w)}$ is the entropy of words of length $n$, and $h_i^{(l)}$ is the entropy of the $i$-th letter in a word. This dependency, valid for the language model in [13], has a steep increase from 1 through 5 characters, and then an approximately linear growth with a much shallower slope of 0.6–0.7 bpc. This is very different from our Fig. 1, and even though our data is on unpredictability, rather than entropy, the difference is probably significant.

In fact, our result may at first glance seem trivial. Indeed, according to a theorem due to Shannon (Theorem 3 in [2]), for a character sequence emitted by a stationary ergodic source, almost all subsequences of length $n$ have the same probability exponential in $n$: $P_n = 2^{-Hn}$ for large enough length ($H$ is the entropy of the source). However, this explanation is not valid here for several reasons. Even if we set aside the question of natural language ergodicity, from the formal point of view, the theorem requires that $n$ is large enough so that all possible letter digrams are likely to be encountered more than once (many times, in fact). Needless to say that the length of a single word is much less than that. Practically, if this explanation were to be adopted, we'd expect the probability to guess a word to be on the order $P_n$, which is much smaller than the observed probability. In fact, the only reason our subjects are able to guess words in context is that the words are connected to the context and make sense in it, while under the assumptions of Shannon's theorem, the equiprobable subsequences are asymptotically independent of the context.

Another tentative argument is to presume that the total number of words in the language (either in the vocabulary or in texts, which is not the same thing) of a given length increases with length, which makes longer words harder to guess due to sheer expansion of possibilities. If there had been exponential expansion of vocabulary with word length, we could argue that contextual restrictions on word choice cut the number of choices by a constant factor (on the average), so the number of words satisfying these restrictions still grows exponentially with word length. However, the data does not support this idea. Distribution of words by length, whether computed from the actual texts or from a dictionary (we used a Russian frequency dictionary containing 32000 words [14]), is not even monotonic, let alone exponentially growing. The number of different words grows up to about 8 characters of length, then decreases. This behavior is in no way reflected in Figs 1, 2, so we can conclude that the total number of dictionary words of a given length is not a factor in guessing success.

In fact, the word length distribution could have had a direct effect on unpredictability only if the word length were known to the subject. But this is generally not the case. Subjects in our experiment are not given any external clue as to the length of the omitted word. Since Russian verse is for the most part metric, the syllabic length of a line is typically known, and this allows to predict the syllabic length of the omitted word with a great deal of certainty. However unpredictability depends on word length in exactly the same way for poetry and prose (see Fig. 3), and in prose there are no external *or* internal clues for the word length. [3]

This leaves us with the only reasonable explanation for the observed dependency: in course of its evolution, the language tends to even out information rate, so that longer words carry proportionally more information. This would be a natural assumption, since an uneven information rate is inefficient: some portions will underutilize the bandwidth of the channel, and some will overutilize it and diminish error-correction capabilities. In other words, as language changes over time, some words and grammatical forms that are too long will be shortened, and those that are too short will be expanded and reinforced.

It is interesting to note that this hypothesis was also proposed in passing by Church and Mercer in a different context in [12]. Discussing applications of trigram word-prediction models to speech recognition, they write (page 12):

> In general, high-frequency function words like *to* and *the*, which are acoustically short, are more predictable than content words like *resolve* and *important*, which are longer. This is convenient for speech recognition because it means that the language model provides more powerful constraints just when the acoustic model is having the toughest time. One suspects that this is not an accident, but rather a natural result of the evolution of speech to fill the human needs for reliable communication in the presence of noise.

A feature that is "convenient for speech recognition" is, indeed, not to be unexpected in natural language, and from our results it appears that its extent is much broader than could be suggested

[3] It is also worth noting that average unpredictability of words in poetry and prose is surprisingly close. In poetry, it turns out, predictability due to meter and rhyme is counteracted by increased unpredictability of semantics and, possibly, grammar. Notably, these two tendencies almost balance each other. This phenomenon and its significance is discussed at length in [11].

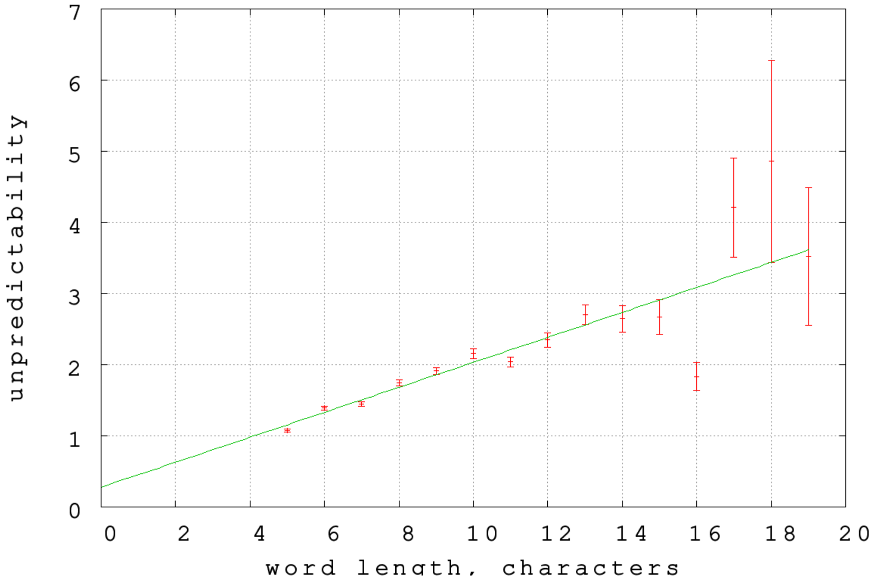

Fig. 3. Unpredictability as a function of word length in characters, prose only

by Church and Mercer's observation. Of course, this is only one of many mechanisms that drive language change, and it only acts statistically, so any given language state will have low-redundancy and high-redundancy pockets. Thus, any Russian speaker knows how difficult it is to distinguish between *mne nado* 'I need' and *ne nado* 'please don't'. Moreover, it is likely that this change typically proceeds by vacillations. As an example consider the evolution of negation in English according to [15] (p. 175–176):

> the original Old English word of negation was *ne*, as in *ic ne wāt*, 'I don't know'. This ordinary mode of negation could be reinforced by the hyperbolic use of either *wiht* 'something, anything' or *nāwiht* 'nothing, not anything' [...]. As time progressed, the hyperbolic force of *(nā)wiht* began to fade [...] and the form *nāwiht* came to be interpreted as part of a two-part, "discontinuous" marker of negation *ne ... nāwiht* [...]. But once ordinary negation was expressed by two words, *ne* and *nāwiht*, the stage was set for ellipsis to come in and to eliminate the seeming redundancy. The result was that *ne*, the word that originally had been the marker of negation, was deleted, and *not*, the reflex of originally hyperbolic *nāwiht* became the only marker of negation. [...] (Modern English has introduced further changes through the introduction of the "helping word" *do*.)

This looks very much like oscillations resulting from an iterative search for the optimum length of a particular grammatical form. It's all the more amazing then, how this tendency, despite its statistical and non-stationary character, beautifully manifests itself in the data.

*Acknowledgements.* I am grateful to D. Flitman for hosting the website, to Yu. Manin, M. Verbitsky, Yu. Fridman, R. Leibov, G. Mints and many others for fruitful discussions. This work could not have happened without the generous support of over 8000 people who generated experimental data by playing the game.

## REFERENCES

1. Shannon C.E. Prediction and entropy of printed English. *Bell System Technical Journal*, 1951, vol. 30, pp. 50–64.
2. Shannon C.E. A mathematical theory of communication. *Bell System Technical Journal*, 1948, vol. 27, pp. 379–423.
3. Burton N.G., Licklider J.C.R. Long-range constraints in the statistical structure of printed English. *American Journal of Psychology*, 1955, vol. 68, no. 4, pp. 650–653
4. Fónagy I. Informationsgehalt von wort und laut in der dichtung. In: *Poetics. Poetyka. Поэтика.* Warszawa: Państwo Wydawnictwo Naukowe, 1961, pp. 591–605.
5. Kolmogorov A. Three approaches to the quantitative definition of information. *Problems Inform. Transmission*, 1965, vol. 1, pp. 1–7.
6. Cover T.M., King R.C. A convergent gambling estimate of the entropy of English. *Information Theory, IEEE Transactions on*, 1978, vol. 24, no. 4, pp. 413–421.
7. Moradi H., Roberts J.A., Grzymala-Busse J.W. Entropy of English text: Experiments with humans and a machine learning system based on rough sets. *Inf. Sci.*, 1998, vol. 104, no. 1–2, pp. 31–47.
8. Paisley W.J. The effects of authorship, topic structure, and time of composition on letter redundancy in English text. *J. Verbal. Behav.*, 1966, vol. 5, pp. 28–34.
9. Brown P.F., Della Pietra V.J., Mercer R.L., Della Pietra S.A., Lai J.C. An estimate of an upper bound for the entropy of English. *Comput. Linguist.*, 1992, vol. 18, no. 1, pp. 31–40.
10. Teahan W.J., Cleary J.G. The entropy of English using PPM-based models. In: *DCC '96: Proceedings of the Conference on Data Compression*, Washington: IEEE Computer Society, 1996, pp. 53–62.
11. Leibov R.G., Manin D.Yu. An attempt at experimental poetics [tentative title]. To be published in: *Proc. Tartu Univ.* [in Russian], Tartu: Tartu University Press, 2006
12. Church K.W., Mercer R.L. Introduction to the special issue on computational linguistics using large corpora. *Comput. Linguist.*, 1993, vol. 19, no. 1, pp. 1–24.
13. T.Schürmann and P.Grassberger. Entropy estimation of symbol sequences. *Chaos*, 1996, vol. 6, no. 3, pp. 414–427.
14. Sharoff S., The frequency dictionary for Russian. *http://www.artint.ru/projects/frqlist/frqlist-en.asp*
15. Hock H.H., Joseph B.D. Language History, Language Change, and Language Relationship. Berlin–New York: Mouton de Gruyter, 1996.

*This paper was recommended for publication by J.D.Apresjan, a member of the Editorial Board*