Electronic Notes in Theoretical Computer Science 53 (2002)
URL: http://www.elsevier.nl/locate/entcs/volume53.html 17 pages

A Revision of Coding Theory for Learning from Language

Łukasz Dębowski

Institute of Computer Science, Polish Academy of Sciences,
ul. Ordona 21, 01-237 Warszawa, Poland
ldebowsk@ipipan.waw.pl

Abstract

Elements of a quantitative-symbolic theory of human language communication based on power-law entropic sublinearity are induced from independent results in quantitative linguistics, statistical NLP, information theory, plectics, neurosciences and statistical physics.

1 Introduction

Being a former physicist and studying natural language processing (NLP), the author has wondered why there is so little communication between statistical NLP, an applied science, and quantitative linguistics, its theoretical partner. Should there be no useful contributions possible? This eagered him to look into Zipf's law. [1] By hazard, the author found articles [16,17]. The articles report that quite trickily estimated Shannon entropy $H(N)$ of N-tuple of consecutive symbols in meaningful strings can be fitted by formula

$$H(N) = \epsilon_0 + \epsilon_\mu N^\mu + \epsilon_1 N, \tag{1}$$

with $\mu = \frac{1}{2}$ for natural language texts and $\mu = \frac{1}{4}$ for music transcripts. For English and German texts $H(N)$ could be safely estimated up to $N \approx 30$ characters with $\epsilon_0 \approx 0$, $\epsilon_{1/2} \approx 3.1$ bits and $\epsilon_1 \approx 0.4$ bits.

Eq. (1) suggests that natural language texts, or *parole* [13], and music transcripts are best describable in terms of neither any fixed deterministic models nor any fixed finitary probabilistic models. The reason is that for $N \to \infty$, one observes $\epsilon_0 \neq 0$, $\epsilon_i = 0$, $i = \mu, 1$ for deterministic processes

[1] [1] has shown that Zipf's law is met at least by strings of independently tossed letters and spaces. [19] reports on change in the law's exponent from -1 to -3 for ranks $\approx 10^4$, which cannot be explained by character independence. [20] proposes an alternative model.

and $\epsilon_i \neq 0$, $i = 0, 1$, $\epsilon_\mu = 0$ for finitary stochastic ones while any other sublinear component is decreasing exponentially w.r.t. N or faster [11]. A differentiation between "model" and "process" is adopted: Process is a distribution of infinite data. Model is a set of formulas that specifies a process completely.

Finitary stochastic processes form a special class of stochastic processes which can be generated by models with finite or short-range memory in the finite number of distinctly distributed observed or hidden variables. Some examples are Markov models, hidden Markov models (HMMs), short-range hamiltonian systems in physics. Probabilistic context free grammars (PCFGs) probably are also elements of the class (see appendix A). For finitary processes, the structure of random variable dependencies in the best model inferable from data almost stabilizes. The size of the best model grows at most logarithmically w.r.t. the dataset size [4].

Finitary stochastic processes are the only possible for classical frequentist statistics. [2] In frequentist statistics, firstly, one guesses an absolutely true but finitary model to generate the infinite data, and then one estimates a finite number of its parameters from the data. The strategy of frequentist statistics, however, has three potential fail points:

(i) One needs the "infinite" data: The reasonable training dataset size grows exponentially w.r.t. to the length of correlations captured by the model.

(ii) One needs to know the "true" model: If the model is wrong, no amount of data can influence its predefined structure and it gets estimated unreliably.

(iii) There may be no "true" model: The structure of variable dependence in the best model for data may strongly depend on the data size. Important elements of the structure may be edited infinitely when the dataset expands.

Case (iii) is the case of $\epsilon_\mu \neq 0$, $0 < \mu < 1$. For natural language one might not use the apparatus of frequentist statistics safely and no linguist could ever say *in principle* what is the complete model for a language.

Inferring any strong conclusion about the correct strategy for language modeling from several data points in [16,17] is not so sound as deductive reasoning. Nevertheless, the case of $\epsilon_\mu \neq 0$ can account for many independent intuitions what human language is while the case of $\epsilon_\mu = 0$ cannot. Eq. (1) with the fitted parameters could be called the fundamental law of

[2] Frequentist statistics has alternatives such as Bayesian approach introduced in [26]. [2] introduces into another "non-classical" statistics of time series with long memory. Long memory time series are defined as such that autocorrelation function between two positions in the series separated by d ones decays like $1/d$ or slower. [30] claims, however, that linguistic counterpart of correlation function being square root of mutual information between two characters separated by d ones is proportional only to $1/d^{1.5}$.

natural language.

The article was planned to summarize manuscript [14], written rapidly just after the author's first contact with eq. (1). Preparing this article, the author found also sources [11,4,35,24], which helped to refine his initial ideas.

2 Analytical theory of complexity measure

What general formalisms and strategies for their acquisition can be used for describing processes with $\epsilon_\mu \neq 0$ and various $0 < \mu < 1$? It is a technical question, which cannot be answered strictly using the present knowledge. Anyway, the best description of *all* empirical data is widely *believed* to be always the shortest one (principle of minimum description length).

Following [14], let assume that any admissible description for any data consists of two parts: (i) The codebook, being the definition of some decoding procedure C. (ii) The encoded data, being some argument A for procedure C. $C(A)$ is the original, unencoded data. In a simplified symbolic approach discussed in section 4, C is a set of some codeword definitions and A is a string of the codewords. No such assumption is made now. Let N be the length of unencoded data and let $D(N,C)$ be the length of their description using procedure C.

$$D(N,C) = \Delta(N,C) + \Theta(C), \tag{2}$$

where $\Delta(N,C)$ is the length of A and $\Theta(C)$ is the length of C. It is assumed that the length of A does not depend on anything else but N and C.

Let $C(N)$ stand for C which is applied to A in the shortest description. The best decoding procedure for data may depend on their size. Let assume that N and C can be approximated as continuous, and $\Delta(N,C)$, $\Theta(C)$ are differentiable functions of their parameters. Minimality of $D(N,C)$ for $C = C(N)$ can be stated as

$$\left.\frac{\partial D(N,C)}{\partial C}\right|_{N=\text{const},C=C(N)} = 0. \tag{3}$$

The notation means that for C parameterized by some parameters, all derivatives of $D(N,C)$ w.r.t. these parameters equal to 0 for $C = C(N)$. Let assume that for any constant C, A grows proportionally to the original data size,

$$\left.\frac{\partial D(N,C)}{\partial N}\right|_{C=\text{const}} = \frac{\Delta(N,C)}{N}. \tag{4}$$

One should expect eq. (4) for sufficiently large N if any fixed decoding procedure acts almost locally on sufficiently large encoded data.

$D(N) := D(N,C(N))$ is the minimum description length (MDL), $\Theta(N) := \Theta(C(N))$ is the length of the optimal codebook, $\Delta(N) := \Delta(N,C(N))$ is the optimal length of encoded data. Combining eqs. (2), (3), (4) yields

(5) $\Theta(N) = D(N) - ND'(N)$,

(6) $D(N) = \Theta(N) + \Delta(N)$.

Eq. (5) is arrived at with no more assumptions about the nature of process producing the data. It may remain valid also in extreme cases when the process cannot be described purely in terms of classical conditional probabilities or deterministic (rewriting) rules. For discrete N, eq. (5) may apply approximately for sufficiently large N.

Looking for the average length of the best code, Shannon has consciously omitted the length of the codebook itself and the formula for Shannon entropy reflects this omission. For finitary stochastic processes, it is $\Delta(N)$ rather than $D(N)$ that can be identified with N-tuple Shannon entropy $H(N)$. Let assume for a while that what Shannon entropy estimators designed to work well for finitary stochastic processes estimate estimates also some kind of $\Delta(N)$ for other data. [3] Solving eq. (5) for given $D(N)$, $\Delta(N)$, or $\Theta(N)$ is easy because these are intriguingly linear equations. For instance,

$$\Theta(N_2) = \Theta(N_1) + \int_{N_1}^{N_2} \frac{\Delta(N)}{N} dN - \Delta(N_2) + \Delta(N_1). \tag{7}$$

If one puts $\Delta(N) = H(N)$ then for eq. (1) and $\epsilon_\mu \neq 0$, the optimal codebook grows infinitely w.r.t. the amount of data collected,

$$\Theta(N_2) = \Theta(N_1) + \frac{1-\mu}{\mu} \epsilon_\mu \left[N_2^\mu - N_1^\mu \right]. \tag{8}$$

The constant term in $\Theta(N)$ may be unlearnable for purely deterministic systems and finite data, remember the problems with CFG induction from a finite number of positive examples only. The presence of terms $\propto N^\mu$, $\propto N$ in real world problems might imply a possibility of estimating the constant term in $\Theta(N)$ for them.

Quantities similar to $\Theta(N)$ have been being defined independently for the very recent years. [4,24] introduce predictive information I_{pred} as mutual information between the whole past and the whole future of any time point in the stationary stochastic process $..., S_1, S_2, S_3, ...$,

$$I_{pred} = \lim_{N \to \infty} I(S_1, ..., S_N; S_{N+1}, ..., S_{2N}). \tag{9}$$

[3] [14] contains a mistake of identifying $H(N) = D(N)$. Furthermore, there is an unsolved problem of estimating $\Delta(N)$ as the length of really optimally encoded data. Estimating $\Delta(N)$ as Shannon entropy $H(N)$ of the sample is dangerous. For $\epsilon_\mu \neq 0$, $\Delta(N)$ being average code-length for N-tuple used apart is significantly bigger than $H(N)$being average code-length for N-tuple of data immersed in the infinite ensemble which is all coded. It is also why recursive definitions introduced in section 4 do profit. At last, there may be no naive infinite ensemble as the usual law of large numbers can be hardly seen experimentally for some objects (problem of absolute probabilities for words rather than ranks).

[11] provides many formal proofs on properties of entropic series $H(N)$. It proves also $I_{pred} = E$, where excess entropy E is defined as

$$E = \lim_{N\to\infty} E(N), \quad E(N) = H(N) - N \lim_{N'\to\infty} [H(N') - H(N'-1)], \tag{10}$$

Both $E(N)$ and $\Theta(N)$ meet the desirable conditions for so called complexity measures discussed in plectics (studies of complexity) [22]. They remain finite for easily formalizable deterministic or finitary systems (homogeneous substances) while possibly growing infinitely w.r.t. the system size for those that are extremely laborious to describe (life, language, society).

3 What is the human language for?

$\Theta(N)$ is the size of the theory which is most appropriate for describing the N-tuple of data. $\Delta(N)$ measures only the remaining randomness. For *parole*, unbound growth of $\Theta(N)$ is reasonable. The receiver of *parole* can profit from inferring new portions of generalizable knowledge from the incoming signal infinitely. Otherwise, it would be very difficult to explain what for the humans keep on communicating to each other in the long-time perspective.

Some readers might have a good objection: Why cannot humans send their messages to each other directly rather than as complex redundancy pattern of another signal? Probably, humans use language for two different purposes:

(i) For their own intellectual development (global communication by finding the best codebook of the input, i.e. exceeding any current formal system).

(ii) For commanding each other, in which they behave as computer-like machines (local communication by interpreting the input sequentially and deterministically, i.e. staying within the current formal system).

From the adults' short-term perspective, primary is purpose (ii) which dominates in practical dialogues while purpose (i) appears mainly in currently thoughtful statements. For a child starting at almost *tabula rasa* and learning from listening to the adults' *parole*, purpose (i) must be primary if it is ever to achieve interpreting purpose (ii) properly.

The unique feature of human language is that it forms a means optimized for both purposes (i) and (ii). Provided the rate of mechanical effort put into communication is not negligible but practically constant, *parole* would be optimized w.r.t. purpose (i) for $\epsilon_i = 0$, $i = 0, 1$, and w.r.t. purpose (ii) for $\epsilon_i = 0$, $i = \mu, 1$. Because both interests of children and adults must be addressed in the same *parole* for the species survival, quotient ϵ_μ/ϵ_1 gets stabilized on a nonzero finite value. The more a child of a species had

to learn the bigger the quotient were and the longer the childhood.

Does $\mu = \frac{1}{2}$ hold for $N \gg 30$ characters still? Do adults have so many brand new things to say all life at a regular rate? Casual conversations are not such a learning task. Having done them, one can largely forget them. $\mu = \frac{1}{2}$ for small N may be only a trace of a brain device for speeding up general language learning by children listening to casual adult speech. In order to be nonexistent in adults, the device should use extensively some hardware resources of brain. Might a possible physical mechanism of the device be the wave of neural plasticity and myelination propagating through neocortex [18,28]? For large N, greater than $\approx 10^3$ (length of simple question-answer block or causal statement), μ may decrease to smaller values depending on the text and the amount of "items" still learnable by an adult. If the author's hypothesis were verified then with the change of Zipf's law exponent [20], it might provide measurable evidence that general language processing is functionally different from more abstract conceptual processing by humans.

4 Computing codebooks by reversible compression

Using a really good compression algorithm for data with $\epsilon_\mu \neq 0$, maybe one could find the optimal structure for the data description, predict the future and generate the process. At the moment, optimal algorithms for reversible compression are known rigorously only for two cases [10]:

(i) Coding finite bag of atomic symbols restricted to context-free rewriting of the atomic symbols and assuming that the codebook takes up no place (Huffman code).

(ii) Coding *infinite* string with$\epsilon_1 \neq 0$ (Shannon entropy justification).

Case (ii) is derived as a limit of case (i) taken with: (a) allowing any number of codewords for all strings over the alphabet of atomic symbols to be defined in the codebook, (b) assuming that all relative frequencies of defined codewords are asymptotically proportional to some constants (probabilities).

With this gigantic increase of computational complexity, however, one gets tiny profit if one *does know* that atomic symbols are generated independently at random (0th order Markov model). Then, one saves only < 1 bit per atomic symbol while getting much less readable codebook! Even this saving is not approached quickly since the average number of additional entries grows only logarithmically w.r.t. to the dataset size [4]. Probabilistic independence assumption combined with whatever other statistical apparatus practically does not improve over simple symbolic processing given the counts of symbols in the data and given no other information about their linear order (syntax). Neither paradigm can compress N symbols as $\propto N^\mu$, $0 < \mu < 1$. The arguments can be iterated over

coding done by rewriting any *finite* number of previously chosen n-tuples of atomic symbols. One should just define these n-tuples as new atomic symbols. For finitary processes, n-tuples with sufficiently large n can be treated as probabilistically independent.

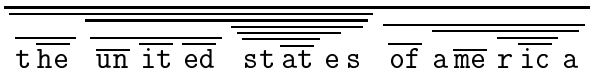
Thus, one concludes that if $H(N) \propto N^{\mu}$, the optimal number of entries must be asymptotically infinite. It provides another evidence for eq. (8). Nevertheless, there appears a problem of finding the optimal coded entities. Using more information about the linear order of symbols in the data can only improve the encoding. On the other hand, testing *all* possibilities of coding by n-tuples appearing in the data cannot be done in real time. Top-down frequentist statistics being unhelpful, bottom-up quantitative-symbolic processing purely driven by data will be adopted instead. The latter approach manages the syntactic information simpler and in each step, it generates only a finite number of new codebook hypotheses with partial present justification.

Each codeword standing for a composition of atomic symbols can be defined either in terms of codewords for atomic symbols only or also in terms of codewords for more complex compositions which together yield the same composition. The first kind of definitions will be called simple, the second one will be called recursive. For $\epsilon_1 \neq 0$ and fully reversible compression, the choice between using simple or recursive definitions does not change significantly the total description length for *infinite* data. For $\epsilon_\mu \neq 0$ and $\epsilon_1 = 0$, eqs. (1), (8) imply that the size of the very best codebook never can be neglected w.r.t. to the size of the encoded data. Recursive definitions shorten both the codebook and the encoded data simultaneously and significantly. Some induced codewords may be used only in the codebook for defining other codewords. Even for $\epsilon_\mu \neq 0$ and $\epsilon_1 \neq 0$, the use of recursive definitions enlarges the codebook significantly and improves its predictive quality although it does not change significantly the total description length for infinite data.

No long-time global optimization of search for the shortest description is feasible. If no genetic algorithms or simulated annealing are used, a feasible compression process reduces to the local search through the graph of all symbolic descriptions for given data. The nodes of the graph are symbolic descriptions themselves, i.e. they represent full fixed states of the codebook and the encoded data. Each description S in the graph is annotated with length $l(S)$ of its Huffman (or Shannon-Fano) coding. $l(S)$ is computed for fixed codebook, counting the codeword occurrences both in the encoded data and in the codebook. The set of atomic symbols is finite so the length of atomic symbol definitions is negligible. A link in the graph from node S to node S' exists if and only if description S' is yielded by applying to description S some symbolic transformation from the predefined finite set of transformations X. Reversibility of the compression is assumed so for each link from S to S', there must be a path from S' to

S. The path back need not be a single link because there may be no such *single* transformation in the set X. The search starts in the node with an empty codebook and encoded data equal to raw data. Then, the search moves from one node S to the next node S' having the minimal $l(S')$, always along existing links, until $l(S) > l(S')$.

An implementation of the presented principles is de Marcken algorithm [12]. In this algorithm, set X consists of two kinds of symbolic transformations: (a) defining a concatenation of two codewords with undefining 0, 1, or 2 elements of the pair, (b) undefining a codeword. Defining an object means introducing a new codeword, replacing all occurrences of the object with the codeword (both in the data and in the codebook) and enrolling the definition of the codeword as the object into the codebook. Undefining the codeword is the reverse. De Marcken algorithm produces a codebook consisting of recursive and mostly meaningful definitions of often syllables, morphemes, words and fixed phrases. Since more distant (and some internal) correlations are not pure concatenations, then it stops enlarging the codebook with compression rate about 2 bits/character for English texts. An example from [12] is:



The bars overbrace the concatenations that got defined as codewords. The text was deprived of spaces, capitalization, and punctuation.

Une langue est un système où tout se tient [13]. All linguistic entities can arise only if their frequencies roughly exceed the products of frequencies of their parts. The frequencies must be read not only from the encoded data but also from the optimal codebook, which allows acquiring language from raw *parole* but only when incrementally classifying its important correlations. There may be hardly any entities of language processing in the human brains before their frequencies are really observed and internally transformed. The exception would be for the meaningless "atomic symbols" of perception and basic "symbolic operations". Everything else would result from "macroeconomic control" of the trade between feasible definitions and available lengths of names (pointers). The same system of discrete entities can be inferred even for highly contextually dependent raw frequency distributions. That property enables quite free communication by adults while not disabling the bottom-up strategy of a child carefully listening to them. Formal grammars seem to be an abstract branch of statistics, abstracting even from the frequencies of the entities described. This abstraction, however, is not total. The continuous space of all frequency distributions of strings is cut into discrete basins of optimal grammars for them. For continuous changes of raw frequency distributions, an abrupt change in the optimally inferred formal system can occur.

There are parallels between the discussed reversible compression formalism and recent hypotheses in neurosciences. [31] argues thoroughly for a functional distinction between neocortex and hippocampus (parts of brain), which can be resumed that neocortex (larger one) stores mostly the codebook while hippocampus (smaller one) stores the recent encoded data. The recursive form of codebook by de Marcken algorithm resembles also the neurolinguistic stratificational model presented in [28].

5 Challenges for quantitative-symbolic processing

5.1 What are the proper symbolic transformations?

The substantial progress of machine language structure acquisition is not a question of improving the quantitative side of the learning scheme but rather considering more complex *reversible* symbolic transformations of the descriptions. What de Marcken algorithm does is rewriting its input as a nearly shortest context-free grammar (CFG). This CFG is restricted to one-to-one rewriting rules where each codeword on the left-hand-side must be precedent to all those on the right-hand-side accordingly to a partial order relation. To conform fully with CFG format, the encoded data should be treated as the rewriting for the initial symbol.

One might think that in general, all entries in the codebook might be the rules rewriting any strings of codewords as any strings of codewords. Then these rules needed to be also one-to-one, and because each rule had to contribute to the overall compression, their left-hand-sides had to be shorter than their right-hand-sides if counted in binary symbols. It means that the codebooks would have a restricted form of context-sensitive grammars (CSGs). Maintaining the condition of unique derivation of the initial symbol (compression reversibility) for a CSG is problematic. Besides that, CSG codebooks are too weak to generalize independent behavior of strings into paradigmatic definitions (for morphemes having phonetic alternations and harmonies, words having inflections, phrases instantiated as words).

In paradigmatic definitions, the entity is coded as an easily computable function of several arguments which are processed in parallel rather than in linear order. More regular domain of the function than simply attested argument combinations should be biased for as well. When dealing with a specific paradigmatic definition induction, such as word inflection, one can disable the recursivity of definitions [23,37]. The unlimited growth of *parole* codebooks inclines to think of infinite hierarchies of paradigmatic definitions and their coherent inference by compression. [4]

[4] Complex paradigmatic phenomena abound in vision. 2D images are coded as functions of 3D objects, their coordinates and rotations. The greater difficulty of vision than language processing can be also accounted by the nonexistence of context-free compres-

Necessity of using only reversible transformations is shared by reversible compression and quantum computing, introduced in [5]. The analogy may stimulate an exchange of algorithms in both domains even if human brains are not performing really microscopic quantum computing.

5.2 What is the proper searching method?

De Marcken algorithm has been introduced here as doing a completely local search. It is not true about the original one from [12], which defines and undefines codewords in large batches. Such scheme reduces the amount of symbolic processing but to optimize the use of pre-chosen transformations on the data, frequentist probabilistic optimization methods such as expectation maximization and Viterbi search are used.

For data with power-law entropic sublinearity, frequentist probabilistic optimization methods may be worse than simple but really the local search. In language, the law of large numbers is hardly met so one can learn not only the frequencies but also newly encountered structures of dependencies all the time. Probabilistic optimization methods assume that frequencies in the data meet the law of large numbers and they optimize w.r.t. small fluctuations from the "static" probabilities. When probabilistically optimizing a chain of actions, one is assuming that in the course of them, no structural learning is planned. One is very sure that the data are more wrong than one's model can be. In order to enable unexpected structural learning, one should reconsult the changed frequencies in data after each performed transformation on them and only then choose and perform the next best action. This is, however, the local search (compare [25]). Currently, at the lowest level of learning from *parole*, the local search finds better solutions than probabilistic optimizations techniques [7,29].

With being very low in the beginning and apparently growing later during the life, human ability to perform non-local search might be learnt during data processing. [8] points out that many cases of what has been considered global optimization and learning in biology, later appeared to be an effect of genetic algorithms (species evolution, immunological reaction). [8] proposes the same mechanism also for inconscious brain processing in the real time.

5.3 How to forget the past?

For $\epsilon_i \neq 0$, $i = \mu, 1$, the series learning is no more static recursive compression nor frequentist probabilistic inference. A time perspective appears and many mathematically ill-posed problems arise. $\epsilon_1 \neq 0$ means that the systematic codebook transmission using $\epsilon_\mu \neq 0$ is interfered by a portion of "random noise". The random noise cannot be compressed sublin-

sion algorithm of de Marcken's type for *discrete* multidimensional input data [14].

early and in its best encoded form, it gives no predictive information while wasting the memory resources [4]. The optimal strategy for learning from *parole* is to compress it irreversibly with forgetting anything that remains random but only if it is *really* random. There are some works on defining absolute tests of randomness for Turing machines [36] but the human tests for humans must be much simpler, easily computable and only a bit worse [34].

Deterministic interpretation may be applied as the single method only for fully learnt, purely finitary processes. Asymptotically there is no learning in them and the past can be forgotten with no risk. The child may be more eager to forget these parts of its current representation of *parole* which it can interpret deterministically, while remembering more often those which cause troubles for its deterministic interpretation. When the child is learning, its discrimination between these two classes evolves.

Two remarks are worth making for the future research: (i) Pure minimum-description-length formalism cannot contribute to clarifying the notion of deterministic non-interpretability. This formalism is just a construction of the shortest system of hierarchical rules and exceptions in one which exists for adding any new exceptional instance. (ii) If the counts of all codewords are memorized independently from their linear order in the description, forgetting the order of some codewords can be done with or without decreasing their memorized counts. Not decreasing the counts stabilizes the entries for the codewords the order of which is forgotten. On the other hand, it also can make new entry additions more difficult.

5.4 How to classify the future and generate the full process?

Short-term prediction capabilities are necessary for improving discrete linguistic classification of non-discrete acoustic percepts [27]. For finitary processes, the relative frequencies of coded entities in the shortest description of sufficiently large data are asymptotically equal to their (short-term predictive) probabilities in the optimal model. For a process with $\epsilon_\mu \neq 0$, $\epsilon_1 = 0$ linking compression and short-term prediction is a new and unanswered problem. Because the optimal codebook takes up a macroscopic space due to systematically recursive definitions, the codeword frequencies in the whole shortest description may differ significantly from the frequencies in the encoded data only. It is not clear yet how to use these two different frequency distributions for estimating the single predictive probability distribution in general.

Generating the full non-finitary process is a different task than short-term predicting it for the discrete classification only [33]. Also for simpler processes with long-range memory, short-term prediction is far easier than estimating the global parameters (probabilities) [2]. In the short-term prediction one can assume that the model for generating the data is

fixed. In order to generate a full nonfinitary process, one needs to simulate expanding the model continually with the data generation.

6 Is the language acquisition optimal?

[14] tries to explain why exactly $\mu = \frac{1}{2}$ holds for natural language if for all $0 < \mu < 1$, the optimal codebook grows infinitely. The first hypothesis was that the learning child memorizes only the optimal codebook. The natural optimality criterion would be $\Theta(N) = \max$, which yields the optimal μ dependent on N with $\mu \to 1$ for $N \to \infty$. The largest codebook could be extracted from sufficiently large amounts of almost complete noise!

A problem is that with growing μ, one *still* needs to store more and more previously encoded data because the newly extracted portion of the codebook may associate some objects in the previously encoded data and the fresh ones. Thus, the second hypothesis was that for $\epsilon_1 = 0$, the language learner would memorize both the codebook and the encoded data. $\mu = \frac{1}{2}$ for natural language could correspond to the solution of constraint

$$\frac{N}{D(N)}\Theta'(N) \approx \mu(1-\mu) = \max. \tag{11}$$

Such constraint might resume a tendency of the language learner to maximize its overall compression factor $N/D(N)$ and an effective tendency of *parole* to maximize the learner's rate of effective codebook acquisition $\Theta'(N)$.

An alternative explanation of $\mu \approx 0.5$ might use the low precision of estimates of μ. By an analogy to partial autocorrelation function [6], let introduce partial mutual information

$$\begin{aligned} I(A;C|B) &:= I(S_1,...,S_A;S_{A+B+1},...,S_{A+B+C}|S_{A+1},...,S_{A+B}) \\ &= H(A+B)+H(B+C)-H(A+B+C)-H(B). \end{aligned} \tag{12}$$

Condition $I(N;N|N)/H(N) = \max$ yields $\mu \approx 0.574$. *Parole* may be well optimized for maximizing delayed and independent partial explanations of the current speech acts.

Yet another explanation of the values of μ might be searched in the theory of formal languages. One of powerful methods of the theory is to map formal properties of the language L into analytic properties of its generating function $G(z) = \sum_{n\geq 0} g(n)z^n$, where $g(n)$ is the number of strings of length n belonging to L. E.g. $G(z)$ is an algebraic function of z if language L is generated by a context free grammar that is unambiguous [9]. If language L consists of all N-tuples appearing in some string S_1^∞ and all N-tuples appear equally often for given N, then $\log_2 g(N) = H(N)$. Intriguingly, [21] asked a question whether languages with $\log_2 g(N) \propto \sqrt{N}$ can be generated by context-free grammars at all.

7 Linguistics and thermodynamics

Eqs. (5), (11) are quite interesting if one sees them in a wider scientific context. For a physicist, eq. (5) states that optimal codebook length $\Theta(N)$ is a Legendre transform of minimum description length $D(N)$ w.r.t. system size N. Similar Legendre transforms appear in thermodynamics of inanimate nature. The difference is that where in the transforms in nature appear entropy and energy, in language appear entropy and complexity measure. Physical (mechanical) energy seems to be completely canceled out from the macroscopic language behavior! Is it that "nature" is the interplay between entropy and energy function while "culture" is the interplay between entropy and complexity measure? If quantities $\Theta(N)$, $D(N)$ of entropic dimensionality were interpreted as volume and N was interpreted as number of particles in the system, then eq. (11) would mean that given the initial $N = N_1$ and final $N = N_2$, the number of particles and volume grow in such way that the useful work done by the system is maximal.

Theoretical statistical physicists have been conscious of acute problems that are met by maximum Shannon entropy method [3] in inferring the macroscopic behavior of long-range hamiltonian systems. For short-range hamiltonian systems, the method predicts the macroscopic behavior excellently. To deal with long-range hamiltonian systems, non-extensive thermodynamic formalism [35] was proposed several years ago. Non-extensive thermodynamics modifies the definition of entropy for which the maximum entropy method is applied. As a result, the method produces power-law distributions for the same constraints which yield Gaussian distributions when applied to Shannon entropy. While Gaussian distributions abound in inanimate nature, power-law distributions abound in complex systems such as life, language and economy. A new wave of interest in physics is whether some extensions of non-extensive thermodynamics could not only reproduce the power-laws but also explain consistently all large-scale behavior of the complex systems. The question of plausible links between linguistics and thermodynamics is discussed a little more in [15].

8 Conclusions

In order to explain experimental measurements of power-law entropic sublinearity in *parole*, elements of a quantitative-symbolic theory of language communication have been inferred. The theory predicts that *parole* can be fully described neither by finite formal nor finite probabilistic models, so infinite learning of *parole*'s codebook is both possible and necessary. The way to the full and mathematically rigorous theory of human language communication is long still. Even now, there appear interesting

analogies and differences between this theory and nonextensive thermodynamics. The analogies and differences may stimulate further *bidirectional* transfer of ideas between the community of physics and plectics and the community of linguistics and computer science. Influential can be also neuroscientific inspirations.

Finding by hand the full *parole*'s codebook, the part of which is Saussurean *langue*, seems to be no longer feasible. A potentially faster way to understanding human language *as a system* and making practical use of this knowledge seems to be understanding the *natural intelligence* in its own mathematical terms of huge self-organization of computing. This new paradigm may combine and still differ from deterministic computing and frequentist probabilistic inference. Accordingly to [28], it is the context of language processing rather than other cognitive tasks, such as vision, which offers the largest amounts and of easiliest interpretable data for understanding the natural intelligence.

A Fixed PCFGs can fail to generate *parole*

[11] gives processes generated by hidden Markov models (HMMs) as instances of finitary processes and remarks that non-finitary processes may be generated by context-free grammars (CFGs). This remark may be false if the probabilistic CFGs (PCFGs) are meant. PCFGs consist of a finite set of context free rules with probabilities of derivation depending only on the currently expanded node. Given these probabilities, $A_1 D(N) > H(N) > B_1 D(N) > 0$ and $A_2 N > D(N) > B_2 N > 0$, where $D(N)$ is the number of derivation rules used in a derivation tree for N-terminal production. $H(N)$ for typical PCFGs may have a large linear component and possibly only very small sublinear ones. Analogical reasoning is valid also for HMMs, which are probabilistic regular grammars and form a subset of PCFGs. There is little evidence at the moment that HMMs and other PCFGs belong to different classes of entropic sublinearity generators.

The preceding argumentation yields a view onto entropic sublinearity by the process generator rather than its interpreter. Generalizing, $A_1 D(N) > H(N) > B_1 D(N) > 0$ where $D(N)$ is the mean number of independent arbitrary decisions that the generator must make in order to generate N consecutive signs of *parole*. The amount of new elements of the codebook inferable from these N signs does not depend on ϵ_1. Quotient ϵ_1/ϵ_μ measures the amount of redundancy for unsupervised learning rather than for sequential deterministic interpretation. In a message composed exceptionally for unsupervised learning, ϵ_1/ϵ_μ would equal to 0 and the number of intended arbitrary decisions $D(N)$ would be only $\propto N^\mu$. ([32] reports that scientific texts have the lowest ϵ_1.) The necessity of finding such specific decisions before publishing makes the composition of longer and still thoughtful texts so hard. Surely, no HMM or PCFG could produce $D(N)$

with $\propto N^{\mu}$ term but with no or text-dependent $\propto N$ term. Even humans use generate-and-test to compose longer narrations as if their language generators could not cope it in real time.

References

[1] V. Belevitch. Théorie de l'information et statistique linguistique. *Académie royale de Belgique, Bulletin de la classe des sciences*, page 419, 1956.

[2] J. Beran. *Statistics for Long-Memory Processes*. Chapmann & Hill, 1994.

[3] A. L. Berger, S. A. Della Pietra, V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39, 1996.

[4] W. Bialek, I. Nemenman, N. Tishby. Predictability, complexity and learning. http://xxx.lanl.gov/abs/physics/0007070v2, 2000.

[5] S. L. Braunstein. *Quantum computation: a tutorial*. http://www.sees.bangor.ac.uk/~schmuel/comp/comp.html, 1995.

[6] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, 1987.

[7] E. Brill. Transformation-based error-driven learning and natural language processing: A case study of part-of-speech tagging. *Computational Linguistics*, 21:543, 1995.

[8] W. H. Calvin. *The Cerebral Code. Thinking a Thought in the Mosaics of the Mind*. The MIT Press, 1996.

[9] N. Chomsky and M. P. Schützenberger. The algebraic theory of context-free languages. In P. Bradford and D. Hirschberg, editors, *Computer Programming and Formal Systems*. North-Holland, 1963.

[10] T. M. Cover, J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[11] J. P. Crutchfield, D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. Santa Fe Institute Working Paper 01-02-012, 2001.

[12] C. G. de Marcken. *Unsupervised Language Acquisition*. PhD thesis, Massachussetts Institute of Technology, 1996.

[13] F. de Saussure. *Cours de linguistique générale*. Payot, 1916.

[14] Ł. Dębowski. Quantitative considerations on finding the shortest descriptions for meaningful symbolic sequences. ICS PAS Reports, Nr 924, Instytut Podstaw Informatyki PAN, 2001. [5]

[5] Author's papers are available also from http://www.ipipan.waw.pl/~ldebowsk.

[15] Ł. Dębowski. Entropic Subextensivity in Language and Learning. In *Interdisciplinary Applications of Ideas from Nonextensive Statistical Mechanics and Thermodynamics*. Santa Fe Institute, 2001.

[16] W. Ebeling, T. Pöschel. Entropy and long-range correlations in literary English. *Europhysics Letters*, 26:241, 1994.

[17] W. Ebeling. Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO. *Physica D*, 109:42, 1997.

[18] J. L. Elman. Origins of language: A conspiracy theory. In B. MacWhinney, editor, *The emergence of language*. Lawrence Earlbaum Associates, 2000.

[19] R. Ferrer, R. V. Solé. Two regimes in the frequency of words and the origins of complex lexicons. Santa Fe Institute Working Paper 00-12-068, 2000.

[20] R. Ferrer, R. V. Solé. The small-world of human language. Santa Fe Institute Working Paper 01-03-016, 2001.

[21] P. Flajolet. Analytic models and ambiguity of context-free languages. *Theoretical Computer Science*, 49:283, 1987.

[22] M. Gell-Mann. What is complexity? *Complexity*, 1, 1995.

[23] J. Goldsmith. Unsupervised learning of the morphology of a natural language. University of Chicago.

[24] P. Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25:907, 1986.

[25] E. T. Jaynes. Entropy and search-theory. In C. R. Smith and Jr. W. T. Grandy, editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*. D. Reidel, 1985.

[26] E. T. Jaynes. *Probability Theory—The Logic of Science*. http://bayes.wustl.edu/etj/node2.html.

[27] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, 1997.

[28] S. M. Lamb. *Pathways of the Brain. The Neurocognitive Basis of Language*. John Benjamin's Publishing Co., 1998.

[29] S. Lawrence, C. L. Giles, S. Fong. Natural language grammatical inference with recurrent neural networks. *IEEE Transactions on Knowledge and Data Engineering*, 1998.

[30] W. Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60:823, 1990.

[31] J. L. McClelland, B. L. McNaughton, R. C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102:419, 1995.

[32] N. W. Petrova. Code — Merkmale des schriftlichen Textes. In *Sprachstatistik*. Berlin, 1973.

[33] K. Rateitschak, W. Ebeling, J. Freund. Nonlinear dynamical model for texts. *Europhysics Letters*, 35:401, 1996.

[34] P. M. Todd, G. Gigerenzer, ABC Research Group. *Simple Heuristics That Make Us Smart*. Oxford University Press, 1999.

[35] C. Tsallis. Entropic nonextensivity: A possible measure of complexity. Santa Fe Institute Working Paper 00-02-043, 2000.

[36] P. Vitányi and M. Li. Minimum description length induction, Bayesianism and Kolmogorov complexity. *IEEE Transactions on Information Theory*, 46:446, 2000.

[37] D. Yarowsky and R. Wicentowski. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL-2000, Hong Kong*. 2000.