Learning Language

Misha Gromov

March 2, 2020

Abstract

This is an edited version of two final lectures from my course on Large Spaces at the Courant institute in the Fall 2019.

Contents

1 Problems, Concepts, Notations **1**
- 1.1 Library: a Shadow of a Mind . . . 2
- 1.2 Learning Transformations, Annotation of Texts, Efficiency Functions, Modularity and Quasi-Commutativity, Initial, Final and Intermediate Models of Languges . . . 5
- 1.3 Numbers, Probability, Statistics . . . 8
- 1.4 Operational Segments and Contextual Domains . . . 10
- 1.5 Linguistic Trees and Linguistic Cores. . . . 12
- 1.6 Links, Units, Equivalence Classes and Similarity Networks . . . 14
- 1.7 Bootstrapping of Similarities, Coclustering and Composition of Functions. . . . 16
- 1.8 Four Levels of Understanding . . . 18
- 1.9 Universality, Novelty, Redundancy, Entropy, Information, Prediction, Generalization . . . 21

2 Commentaries to the sections 1.1-1.9 **22**
- 2.1 Linguistics, Biology, Mathematics. . . . 22
- 2.2 Similarities, Analogies and Inductive Reasoning . . . 26
- 2.3 Computations, Iterations, and Learning Algorithms . . . 26
- 2.4 Philosophy, Logic, Numbers, Probabilities, Ambiguities, Trees . 26
- 2.5 Thinking in Parallel . . . 26
- 2.6 Composition of Maps, Word embeddings and Neural Networks . 27
- 2.7 In the Mind of the Speaker . . . 27
- 2.8 Predictions, Jumping to Conclusions, Curiosity Driven Learning and Ergologic . . . 27

3 Bibliography. **27**

1 Problems, Concepts, Notations

The real problem is that of developing a hypothesis about initial structure that is sufficiently rich to account for acquisition of language, yet not so rich as to be inconsistent. Noam Chomsky

Nothing in linguistics makes sense except in the light of language learning.
Theodosius Dobzhansky misquoted[1]

The study of *Language* and *Language acquisition* spreads over several domains of linguistics, such as *cognitive* and *developmental linguistics* and of computer science, e.g. the theory/practice of artificial neural networks and *machine learning* in general.

Surprisingly for a mathematician, there is a sharp disagreement between linguists on several basic issues.[2]

1. What is *Language*? What is its main function?
2. What makes natural languages *learnable*? Is it an *innate universal grammar* in our minds?
3. Can a *statistical analysis* of texts reveal the *structure of language?*

A mathematician's answer would be

a natural *class* $\mathscr{M}$ of *models* $\mathcal{M}$ *of languages* and of (learning) transformations of $\mathscr{M}$ [3] which, when applied to (a collection of texts from) a given library (corpus) L, somehow embedded into $\mathscr{M}$, would result in a model $\mathcal{M} \in \mathscr{M}$ of the language $\mathcal{L}$ behind L.

We don't claim we know exactly what kind of mathematics is adequate for this purpose. All we can say is that it should *not be bound* to *the formal language theory* and/or to *(traditional) probability and statistics*[4], but must be developed along the lines of *Grothendieck's ideas* of

functorial naturality of mathematical concepts and constructions

and compatibly with

- basic linguistic principles,
- realities of natural languages,
- data of experimental psychology,
- experience of the machine learning theory.

In the following sections, 1.1-1.9 we shall briefly describe what kind of mathematics we have in mind and what we intend to achieve with it.

1.1 Library: a Shadow of a Mind

We do not understand, and, for all we know, we may never come to understand what makes it possible for a normal human intelligence to use language as an instrument for the free expression of thought and feeling. Noam Chomsky

Thoroughly conscious ignorance is the prelude to
every real advance in science. James Clerk Maxwell

[1] Nothing in biology makes sense except in the light of evolution.

[2] See `https://en.wikipedia.org/wiki/List_of_unsolved_problems_in_linguistics`, [11], [?] , [13] [16] [26], [2], [20].

[3] We don't say *algorithms* to avoid an inappropriate association with *Turing machines* and alike.

[4] See [?] for a brief overview of the probability problem

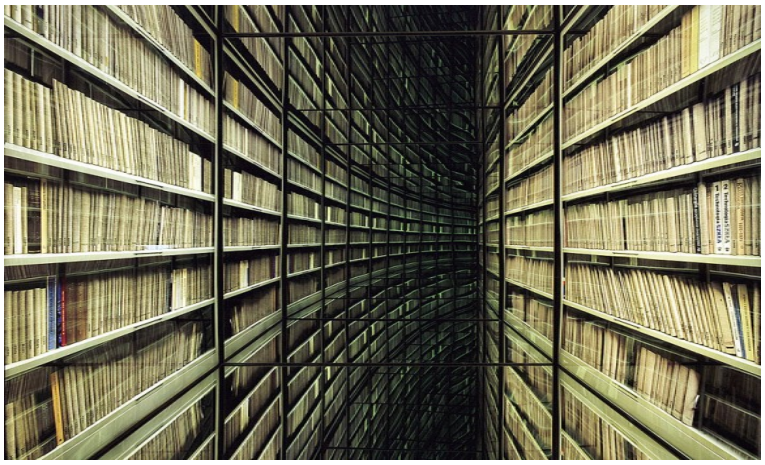


Figure 1: Library

Our starting point is *an imaginary library* L, a collection of billions strings in 25 symbols found on the bottom of the Indian Ocean.[5] We try, notwithstanding what Chomsky says, to build a dynamic mathematical structure/model that, in the case where L represents a *natural language* $\mathcal{L}$,[6] will be similar to what would be (conjecturally) happening in the (subliminal) mind of a child exposed to $\mathcal{L}$ in a natural way.

Compute, as physysts do, (relative) frequencies of different strings in L and evaluate correlations between them.

Then we discover

(1) *a systematic aperiodic recurrence*

of a few million, *seemingly random* substrings in 10-50 symbols.[7]

From a physicist's perspective this is *highly improbable*, virtually impossible.[8] Yet, this is how it is in Life: *multiple copies* – let them be only approximate ones – of *improbably complicated patterns* are seen everywhere – from DNA sequences and genetic code(s) to human speech and human artifacts.

The second essential, also Life-like, formal feature of languages is *discreteness*: texts are fragmented into

(2) *individual* (almost) non-ambiguously identifiable *units*:

words, phrases, sentences, paragraphs and other less pronounced ones. Because of this you can talk about what you read, similarly to how you make little stories of what you see in these pictures.

[5] We don't make any assumptions on where this library come from and/or on how it is organized. Yet, we speak of L as if it were an ordinary human library, say in English, where our use of such terms as "word", "phrase", etc. albeit metaphorical at the first stage prepares us to working out general mathematical definitions later on.

[6] Can the so mutilated idea of language be viable ? Depends on how you call this. Transform mutilation $\mapsto$ idealization and quote Chomsky again: *Opposition to idealization is simply objection to rationality.*

[7] A Google search for *"don't spend your time trying to figure out how"* returns 10 000 results, a single one for "don't spend your time trying how" and none for *"don't your time trying how"*.

[8] Yet, it may be instructive to design a simple dynamical system with a recurrence similar to the kind seen in natural languages.

Figure 2: Too long to go around

Figure 3: Sleeping beauty

But it is hard to make an informative story about a dead landscape, be it on Mars or on Earth.[9]

Figure 4: Mars

Figure 5: Earth

Despite their existence as independent entities, *textual units*, e.g. words, are

(3) *connected by a dense network of* functional linkages *in texts,*

neither mathematical definition of which nor their automatic recognition are obvious.

To appreciate the problem, try to justify in *general* terms, with no ad hoc grammar (or semantic) rule, why, in the following quote by Ferdinand de Saussure, far refers to idea but not to language.

Everyone, left to his own devices, forms an idea *about what goes on in language which is very* far *from the truth.*

(Read this sentence fast and you associate far with Everyone.)

[9]Natural languages are primarily adapted for narrations about Life, since 99% of what your senses receive is: *people, animals, plants, human behavior, human relationships, and everything made by human hands.* But once it comes to non-Life mathematics takes over. For instance, the concepts of *random field* and/or of *fractal* may describe textures of landscapes

(4) *Language* – as Chomsky says – *is a process of free creation*: learning to understand language, be it be written or spoken, is inseparable of learning to speak (or write); mathematically, this is a process of

generation of new fragments of a language by string modification,

where these "modifications" decompose into

several elementary steps

such as

string insertion.[10]

The fifth essential feature of our intended model $\mathcal{M}$ of a language $\mathcal{L}$, also motivated by how human's subliminal mind works, is

(5) a structural division of $\mathcal{M}$ into several *weakly interacting* parts, often called *moduli*, and of the building process of $\mathcal{M}$ from L, into several parallel subprocesses. [11]

This parallel is, in fact, an interconnected network of elementary processes of analysis of incoming language flows[12] and of synthesis of outgoing flows.

The structure of this network will be an essential part of our model of language and of language learning algorithms.

Classification and Similarities

The key to *extraction of the structure* of a language $\mathcal{L}$ from (samples taken from) L, lies in developing a *dynamic hierarchical*[13] *network* of *similarities* between (recurrent) patterns in L and of *classification/clusterization* schemes associated with (some of) these similarities.

Conceivably the rules for construction and use of this network may, at least functionally, stand for the Chomskyan *initial* [linguistic] *structure* hidden in the depth of the human mind.

1.2 Learning Transformations, Annotation of Texts, Efficiency Functions, Modularity and Quasi-Commutativity, Initial, Final and Intermediate Models of Languges

... we are interested in the operating principles of language because we hope that this will give us some clues about the operating principles of the human brain. Eric H. Lenneberg

Although library[14] L is presented as a subset of the set A^* of finite strings in *letters/symbols* a from an *alphabet* A,

$$L \subset A^* = \bigcup_{i=1,2,3,...} A^i = A \cup (A \times A) \cup (A \times A \times A) \cup ...,$$

[10] This is reminiscent of insertions and deletions of segments of DNA by mobile genetic elements.

[11] Brain's bottleneck is the low speed of sequential processing. This is compensated by multichannel parallelism and a large volume of the long term memory.

Parallel recognition, which relies on different independent cues, goes much faster than sequentially following one letter after another, where the time is (at least) proportional to the length of a string.

[12] Some of such "flows" in the human brain/mind may come from within.

[13] "Dynamic" means that the descriptions of (some of) similarities include the (algorithmic) processes they are obtained with and "hierarchical" signifies that this network contains *second order* similarity links between certain (first order) similarities.

[14] A linguist would call it "corpus".

it would be misleading *to define* the corresponding language as such a subset, where the obvious, albeit non-decisive issue, is non-existence of a *natural set* A: alphabets of *letters* are pronouncedly artificial, while *words* don't come in packages of well defined *sets*.[15]

In any case, we are not after *definition of Language* — what we want to eventually obtain is a *mathematical model* of representation of Language in the *human mind*.

We envisage such a model $\mathcal{M}$ of a language $\mathcal{L}$, or of *competence in* $\mathcal{L}$ derived from a library L,[16] as a *hierarchical network of units*, kind of

Dictionary & Extended Grammar of $\mathcal{L}$,

and of

algorithms for manipulating with these *dictionary units* and describing/prescribing their relations/interactions with *textual units*, such as *annotation*, e.g *parsing*, of texts from L, [17] where the decoration of (texts from) L with these annotations is denoted

$$\mathcal{A} = \mathcal{A}(L) = \mathcal{A}_{\mathcal{M}}(L).$$

For instance one of the *read algorithms* compares words/strings w from L with these represented by units in the dictionary, thus recording all appearances of each w in L.

Unlike what a linguist would do, we don't attempt to straightforwardly describe $\mathcal{M}$, but will characterise it as

a member of a certain *simple general explicitly mathematically defined, class/space* $\mathscr{M}$ of conceivable models $\mathcal{M}$.

The location of $\mathcal{M} = \mathcal{M}(L)$ in $\mathscr{M}$ can be specified in (at least) two ways.[18]

(1) *Implicit Localization of* $\mathcal{M} = \mathcal{M}(L)$ *in* $\mathscr{M}$. This is done by imposing some mathematical criteria on $\mathcal{M}$, e.g.

optimality of the speed of the read algorithm in balance with *descriptive complexity of* $\mathcal{M}$.

(2) *Constructive/Algorithmic Localisation.* This is obtained by following a *path of learning*, that, in the first approximation, is an *orbit of a transformation*

$$\Lambda : \mathscr{M} \to \mathscr{M},$$

where Λ consequently applies to *an input* $\mathcal{M}_0 = \mathcal{M}_0(L) \in \mathscr{M}$. This latter is defined on the basis of texts from the library L, and our path is the Λ-orbit of $\mathcal{M}_0$,

$$\mathcal{M}_0 \overset{\Lambda}{\mapsto} \mathcal{M}_1 \overset{\Lambda}{\mapsto} ... \overset{\Lambda}{\mapsto} \mathcal{M}_{st},$$

which is followed until you arrive at a *stationary* or *approximately stationary* point $\mathcal{M}_{\mathsf{st}} \in \mathscr{M}$, i.e. such that

$$\Lambda(\mathcal{M}_{\mathsf{st}}) = \mathcal{M}_{\mathsf{st}} \text{ or, at least, } \Lambda(\mathcal{M}_{\mathsf{st}}) \text{ is sufficiently close to } \mathcal{M}_{\mathsf{st}}.$$[19]

[15] Language is no more a subset in A^*, than a human being is a subset of atoms in the Milky Way galaxy. And speaking of *probability measures* on A^* instead of subsets doesn't help.

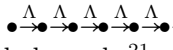
[16] Our concept of *linguistic competence* is not limited to grammar, but includes much of *semantics* and some aspects of *pragmatics* as well.

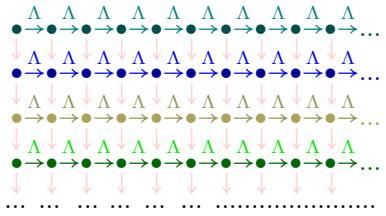
[17] These units may come from another library or generated by the learning algorithm itself.

[18] $\mathscr{M}$ stands for *Chomskyan universal grammar*: *... system of principles and structures that are the prerequisites for acquisition of language, and to which every language necessarily conforms;* specification of the position of $\mathcal{M}$ in $\mathscr{M}$ corresponds to the choice of *parameters*.

[19] To get an idea, think of $\mathscr{M}$ as the class of all finite graphs and of Λ as an instruction

Maximising Efficiency. Ideally, we would like to have a short description for an *efficiency function* $\mathcal{E} = \mathcal{E}_L(\mathcal{M})$, $\mathcal{M} \in \mathscr{M}$, and then define (the scheme of) learning as an optimization process for this $\mathcal{E}$ on $\mathscr{M}$.[20]

Parallelism and Commutativity. Learning, as all subliminal mental processes, is not linearly organized, it is not  , but runs in several (probably, a few hundred) parallel channels,[21] which correspond to the moduli, i.e. the weakly interacting parts into which $\mathscr{M}$, and, accordingly, $\mathcal{M}$ decompose:



What is essential, and what greatly *enhance efficiency* of learning is that (most of) these parallel processes are (locally approximately) *commute*, where this *quasi-commutativity* is due to the structural and dynamic *quasi-independence* of the corresponding moduli.[22]

Keeping this in mind, the second approximation of what we call *the scheme of learning* is pictured as a *partially ordered and approximately commutative semigroup* $\mathbf{\Lambda}$ generated by Λ, Λ, Λ, Λ ..., that operates on $\mathscr{M}$. This allows many different paths of learning corresponding to different compositions, e.g. $\Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda$ ... and $\Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda \circ \Lambda$..., which, when applied to $\mathcal{M}_0$, terminate close each to another in $\mathscr{M}$. (Later on we clarify and refine this picture.)

Decomposing $\mathcal{E}$. The beneficial role of modularity&commutativity in learning is most clearly seen in terms of $\mathcal{E}$.

In fact, since $\mathscr{M}$ and $\mathbf{\Lambda}$ decompose into weakly interacting parts, the full efficiency function $\mathcal{E}$, if it were *something like energy*, would (approximately) decompose into the sum

$$\mathcal{E} \approx \mathcal{E} + \mathcal{E} + \mathcal{E} + \mathcal{E}...,$$

and optimisation of $\mathcal{E}$ would (approximately) reduce to optimisation of all summands independently. (Some of these $\mathcal{E}$ are described in the following section.)

Beside "naked sums", the *non-cancellation property* – that is what makes *one-at-a-time optimization* possible – is satisfied by linear combinations with *positive* coefficients as well as by non-linear functions in $\mathcal{E}, \mathcal{E}, \mathcal{E}, \mathcal{E}...$ that are *monotone increasing*, e.g. $\inf(\mathcal{E}, \mathcal{E}, \mathcal{E}, \mathcal{E}...)$.

---

for modifications (building and rebuilding) graphs, where such a Λ is qualified as a learning algorithm only if it is *robust and simple*.

[20] The existence of such an $\mathcal{E}$ is postulated in most versions of the mathematical/machine learning theory, but we don't take this for granted.

[21] When it comes to programming, this needs to be represented by a *sequential algorithm*, which is, however, *by no means equivalent* to the original parallel one.

[22] Besides grammars of natural languages, where it makes little difference, for instance, in which order one learns conjugation of irregular verbs in English or conjugations of verbs and inclinations of nouns in Russian, modularity&quasi-commutativity is present in many domains of human intellectual activity, e.g. in mathematics, which is organized in a networks of quasi-independent "theorems".

But in any case, there is no a priori reason to think of $\mathcal{E}$ as a *number valued* function and to be concerned with linearity. What is essential, at least at the present stage, is *the partial order* induced by $\mathcal{E}$ on $\mathscr{M}$.

Summing up, we are faced by the following issues.

1. Description in most general terms of mathematical structures we expect to see in mental models $\mathcal{M}$ of languages and furnishing the totality $\mathscr{M}$ of all such $\mathcal{M}$ with a (preferably mathematically natural) structure, adapted for effectuating 2, 3, 4 below.

2. Representation of L by an $\mathcal{M}_0$ in $\mathscr{M}$.

3.Construction of parallel (algorithmic learning) families of transformations

$$\mathbf{\Lambda} : \mathscr{M} \rightrightarrows \mathscr{M},$$

possibly based on efficiency criteria for a positioning $\mathcal{M} = \mathcal{M}(L)$ in $\mathscr{M}$

4. Finding criteria for the quality of performance of a given $\mathcal{M}$ and/or of $\mathbf{\Lambda}$.

Between $\mathcal{M}_0$ and $\mathcal{M}$. Even if we roughly know what the immediate input from L to the mind of a learner is and and what, in general terms, the structure of languges could be, an essential conceptual difficulty remains, since we have no direct data on the structure of

imperfect models corresponding to *intermediate stages* of learning,

which are present, for instance, in the mind of a child who learns a language.

All we can do at this point is to quote Darwin:

Every new body of discovery is mathematical in form, because there is no other guidance we can have.

An instance of such a "form" that is suggested by the mathematical category theory is a model of learning represented by a *function* (*functor*?)

$$L \rightsquigarrow \mathcal{M} = \mathcal{M}(L)$$

defined on a certain class of libraries L, e.g. *sub-libraries* of some $\mathbf{L}$ or of *mental sub-quotients*[23] of $\mathbf{L}$.

Although this, as much as any ad hock mathematical idea, can't be *directly* applied to the real life problem of learning, it does set you mind on a new course of thinking which we shall pursue in the following sections.???

1.3 Numbers, Probability, Statistics

The grand aim of all science is to cover the greatest number of empirical facts by logical deduction from the smallest number of hypotheses or axioms.[24]

ALBERT EINSTEIN

SMALL AND LARGE NUMBERS. Since our model $\mathcal{M}$ of the language $\mathcal{L}$ is supposed to imitate the pristine mind of a child, it will explicitly include only the numbers 1,2,3, rarely, 4, while everything from 5 on will be in the same basket as infinity. But to simplify, we shall allow manipulation with larger numbers, e.g.

[23]Such a sub-quotient will be formally defined later on as a kind of a mental perception of a sub-library of $\mathbf{L}$ by a learner.

[24]A physicist's ambition is a theory based on few, say $\kappa < 10$, basic rules with exponentially many, $\kappa \rightsquigarrow \kappa^n$, $n = 1, 2, 3, 4, 5, ...$ logically deducible empirical facts. In biology, one would be happy with $K \rightsquigarrow K^{n=2}$ for large (about million?) K with admissible 80% error rate. And in linguistic the realistic exponent, I guess, is $n = 3$.

for evaluation of relative frequencies of words in L, without explaining every time how it could've been performed in a number free combinatorial environment.

On the other hand, we use numbers in our (external) description of properties of $\mathcal{M}$, such as the size of $\mathcal{M}$ where – this is crucial in design of $\mathcal{M}$–

no model of a language, or of anything else associated with Life, makes sense outside a narrow range of numerical possibilities.[25]

Unlike the exponential size ($> 10^{100}$) of the number all conceivable strings/sentences in a language $\mathcal{L}$, our L, $\mathcal{M}$, $\mathscr{M}$, Λ must be moderate.

Thus, the realistic size of a library L needed for learning $\mathcal{L}$,

although above 10^6-10^8 words, must be below 10^{12}- 10^{14} words.[26]

The description of the expected model $\mathcal{M}$ (a representative list of vocabulary of $\mathcal{L}$, the grammar and the read and write algorithms) must be

in the range 10^6-10^8 words,

while the "universe" $\mathscr{M}$ of all possible $\mathcal{M}$ and the learning instructions that define the learning (transformation) $\mathbf{\Lambda}$ of $\mathscr{M}$ must admit fairly *short descriptions*:

a few pages for $\mathscr{M}$
and at most
a couple of hundred pages for $\mathbf{\Lambda}$,[27],

where, an *implicit* description of $\mathbf{\Lambda}$ via a suitable efficiency function $\mathcal{E}$ derived from *universal* principles of learning, may be describable just on a dozen pages.)

****

The true logic of this world is in the calculus of probabilities.
James Clerk Maxwell

... probabilistic models give no particular insight into some of the basic problems of syntactic structure. Noam Chomsky

Probability beautifully works in symmetric environment, e.g in *homogeneous* or *nearly homogeneous* spaces, such as the spaces of (states of) identical molecules of a gas in a box. But in order to be *meaningfully* applicable to *heterogeneous structures*identical one encounters in "non-physical worlds", e.g. in the world of languages, the traditional probabilistic formalism and its use must be *limited and modified* in several ways. For instance, one can't unrestrictedly iterate product formulas, e.g. chain rule $P(A\&B) = P(A|B) \cdot P(A)$, since accumulation of errors renders results meaningless, even, where, which is rare in languages, these $P(A|B)$ and $P(A)$ are meaningfully defined.[28]

In any case, since we don't admit large numbers into out model $\mathcal{M}$, it will be, at least overtly, *predominantly combinatorial* rather than stochastic.

[25] "Infinite", *"arbitrarily small"*, *"arbitrarily large"*, etc, are bona fide *mathematical* concepts. Their application in modeling physical systems (e.g. via differential equations) is (non-trivially) justifiable and (often inexplicably) successful. But their unrestrained use in biology, psychology, linguistics and philosophy of AI breeds nothing but pointless speculations.

[26] 10^9 words corresponds to a ten thousand average books, and 10^{14} to a hundred billion Google pages, where the number of all books in the world is of order 10^8 and the number of Google pages is about $3 \cdot 10^{13}$.

[27] This is, probably, how much the *inborn language learning programs/moduli*, most of them operating in parallel, occupy in the brains/subliminal minds of humans.

[28] If each consecutive word in the sentence *... probabilistic models...* is assigned probability $\approx \frac{1}{5}$ – this, albeit inaccurate, is meaningful – then the probability of the whole sentence will come up as *meaningless* $5^{-15} \approx \frac{1}{3 \cdot 10^{10}}$, where a minor perturbation, of $\frac{1}{5}$ to $\frac{1}{4}$ increases the result by huge (>28) factor. See ?? section for more about it.

Yet, the construction of $\mathcal{M}$ and evaluation of its performance will rely on statistics, such as (properly interpreted) distribution of (small) textual units, e g. of words, in our library L, where, *if it seems admissible*, we use basic probabilistic concepts, such as *frequency, correlation, entropy.*

1.4 Operational Segments and Contextual Domains

The full picture of the universe of models $\mathscr{M}$ may still remain vague but an essential part of the first step of the learning process $\mathbf{\Lambda} : \mathscr{M} \rightrightarrows \mathscr{M}$ is quite apparent. It is

SEGMENTATION OF L: identification of *operational segments* in the library, [29] where the basic such segments S, called *words* are, typically, 2 - 20 letters long. (Recall, there are 25 letters in the alphabet of L.)

These may contain significant subsegments called *morphemes*, while certain consecutive strings of words make larger segments called *phrases* and *sentences*, say, up to 50 words in them.[30]

We stop at *paragraphs* (100-200 words),[31] but we identify *large contextual domains*, called LCD in the language $\mathcal{L}$, where the corresponding, slightly ambiguous, fragments of L may be of various, often unspecified, size, say of at least million words in them (about 10 books) and where certain LCD may be decomposed into subdomains, subsubdomains, etc, e.g.

biology $\supset$ molecular biology...
$\cap$
science $\supset$ chemistry $\supset$ organic chemistry $\supset$ chemistry of hydrocarbons...
$\cup$ $\cup$
physics... chemistry of metals...

The number of such domains depends on diversity of cultures of contributors to L. In English, I guess, the number of *linguistically significant* LCD may be counted in hundreds (maybe thousands?), but languages spoken by small groups of people (who make no libraries) may have only 2-3 contextual domains[32]

These (large) domains D are identified *in parallel* with (short) significant/operational segments S by the following (seemingly circular) properties.

↻ A segment S is *operationally significant* if the string contained in it, denoted, $\overset{...}{S}$, *unreasonably frequently* appears in some (or several) D.

↻ A domain D is *semantically significant* if a particular significant segment (or a group of segments) appears in D *abnormally high frequency.*

FROM NUMBERS TO STRUCTURES.

Turning ↻ and ↻ into actual definitions will be done later on along the following lines.

Unreasonable will be defined by requiring the probability $P(S)$ of a significant S to be *significantly greater* than the product of probabilities of subsegments,

[29] We assume that the learner understands the geometry of the (imaginary) line which supports the symbols/letters from L.

[30] These numbers, may depend on peculiarities of L, but all languages generated by minds comparable to ours, are likely to be divided into segments with the information contents similar to how it is in English.

[31] Division into pages, chapters, volumes, and topical shelves, albeit helpful to the learner (if recognizable) is rather arbitrary.

[32] Properly linguistically defined, these domains may be multiple in all human languges.

$S_1, S_2 \subset S$ for all decompositions $S = S_1 \cup S_2$,

$$\frac{P(\dddot{S})}{P(\dddot{S}_1) \cdot P(\dddot{S}_2)} >> 1^{33}$$

(One may assume that $S_1, S_2 \subset S$ don't overlap.)

Besides, S must be *maximal* in the sense that

$$\frac{P(\dddot{S})}{P(\ddot{S}_+)} >> 1$$

for all segments $S_+ \supsetneqq S$

(This will be generalized to include rare words and phrases e.g. by introducing some equivalence between certain rare strings.)

Then, abnormally high will be defined by giving a precise meaning to S appearing in D *much more frequently* that outside D, where, additionally, this frequency must be, statistically speaking, uniform that is (roughly) the same in different parts of D.[34]

The above "definitions" illustrate the idea that

combinatorial structures of a (quasi)deterministic model of a language $\mathcal{L}$ can be derived from statistical analysis of texts from a sufficiently representative library L.

Segments instead of Strings and the Relation $\dddot{\sim}$. Each string of letters in L defines a segment on the imaginary real line, where different segments may contain equal strings. When speaking above about frequencies or probabilities of segments S, we always mean the *string contents of* S, denoted $\dddot{S}$, and observe that equality of the strings, $\ddot{S}_1 = \ddot{S}_2$, defines an *equivalence relation* on the segments, denoted

$$S_1 \,\ddot{\sim}\, S_2.$$

In fact, this equivalence relation on the segments in L uniquely defines the linguistic content of L with no reference to any alphabet.

Thus the set of segments in L turns into a *vertex set of a graph*, call it $SG = SG(L)$, where some edges correspond to $\ddot{\sim}$-relations[35] and some encode the geometry of mutual positions of segments, e.g. indicating adjacency and/or inclusions between them. Then the resulting 3-*colored graph* on the set of *significant segments* provides a *combinatorial representation* of L which will be used later on for the definition/construction of the initial model $\mathcal{M}_0 = \mathcal{M}_0(L) \in \mathscr{M}$.

[33] Although the definition of probabilities of S, S_1 and S_2 is problematic without specifying the contextual domain D these belong to, the ratio $\frac{P(S)}{P(S_1)\cdot P(S_2)}$ is rather independent of D, while specific meaning of $>> 1$ is a parameter that must be eventually adjusted to the reality of L.

[34] Large contextual domains, similarly to those of landscapes (be them on Mars or on Earth as in pictures on p.4) are characterized by distributions of various patterns, which, besides keywords and key (short) phrases may include, for instance, specific grammatical constructions, such as particular *(longish) sentences with subordinate clauses.*

[35] In spoken language, the sound matching between utterings is only an approximate one while the (almost) perfect linguistic matching between *clouds of sounds* emerges at a later stage of speech analysis.

Composable Transformations. Some of the edges in the graph SG, e.g. those representing an inclusion of stings, such as $\mathsf{rtyui} \hookrightarrow \mathit{eaty}\mathsf{rtyui}\mathit{abcyc}$.

An essential feature of these and some other transformations of segments, is *composability*: given

$$f_{12} : S_1 \to S_2 \text{ and } f_{23} : S_2 \to S_3 \text{ there is a } \textit{composition } f_{13} = f_{12} \circ f_{23} : S_1 \to S_3,$$

which satisfies the associativity relation

$$(f_{12} \circ f_{23}) \circ f_{34} = f_{12} \circ (f_{23} \circ f_{34}),$$

that brings to one's mind the formalism of *the mathematical category theory*.

Warning. Most linguistic transformations even if *composable* but then only *approximately*, and also there are limits to the number, probably < 5, of consecutive compositions one can meaningfully perform.

1.5 Linguistic Trees and Linguistic Cores.

It seems, that the main mathematical actors in models $\mathcal{M}$ of languages must be *trees* rather than numbers.

Most common in linguistic are *parsing trees* that correspond to nested segmentations of texts, e.g.
([.....][(.......)(.....)(.......)][.....][.....])([.....]{[....][.....][...]}[.....][...]),
where the corresponding tree is as follow.

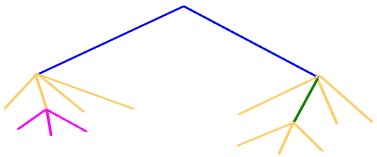


Figure 6: Tree of Segments

Definition of $\mathcal{T}(\dddot{S})$. Bulkier but logically simpler trees $\mathcal{T}$, which represent branchings of strings $\dddot{S}$ in L, are defined as follows.

Given a string $\dddot{S}$, assume, to be specific, it is a word, let $\dddot{S}_*$ denote the copies of $\dddot{S}$ in L and let $\overleftrightarrow{S} \supset \dddot{S}_* \subset L$ be the segments in L, which extend $\dddot{S}_*$ on the left and/or on the right by at most k words on both side, say, for $k = 10$. (Thus, each $\dddot{S}_*$ is contained in ten or less of such $\overleftrightarrow{S}$.)

Then, glue pairwise all pairs of these extended segments, say $\overleftrightarrow{S}_1, \overleftrightarrow{S}_2$ along subsegments

$$\overleftrightarrow{S}'_1 \subset \overleftrightarrow{S}_1 \text{ and } \overleftrightarrow{S}'_2 \subset \overleftrightarrow{S}_2,$$

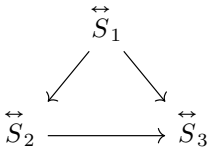
such that

- $\overleftrightarrow{S}'_1$ *contains* $\dddot{S}_1$ and $\overleftrightarrow{S}'_1$ *contains* $\dddot{S}_1$,

• the segments $\overset{\leftrightarrow}{S}_1'$ and $\overset{\leftrightarrow}{S}_2'$ have *equal string contents*, moreover – this is a necessary precaution – that there is an isomorphism between the strings contained in them *which sends* $\overset{\cdots}{S}_1$ *to* $\overset{\cdots}{S}_2$.

• the subsegments $\overset{\leftrightarrow}{S}_1'$ and $\overset{\leftrightarrow}{S}_2'$ are the *maximal ones* in $\overset{\leftrightarrow}{S}_1$ and $\overset{\leftrightarrow}{S}_2$ with the above two properties.

It takes a bit of thinking to realize that that these gluing mutually agree for triples of segments on the subsegments where all three gluings (arrows) are defined,



and that the graph obtained by gluing all these is a *tree*, which we denote

$$\mathcal{T}(\overset{\cdots}{S}) = \mathcal{T}_L(\overset{\cdots}{S}) = \mathcal{T}_{k,L}(\overset{\cdots}{S}).$$

Notice that the direction in L – we assume that the strings in L are directed– induces directions in the trees $\mathcal{T}(\overset{\cdots}{S})$ for all strings/words $\overset{\cdots}{S}$. Accordingly, $\mathcal{T}(\overset{\cdots}{S})$ decomposes into the union of the the *backward tree* and *forward tree*,

$$\mathcal{T}(\overset{\cdots}{S}) = \overset{\leftarrow}{\mathcal{T}}(\overset{\cdots}{S}) \cup \overset{\rightarrow}{\mathcal{T}}(\overset{\cdots}{S}),$$

which intersect over a single segments corresponding to $\overset{\cdots}{S}$.[36]

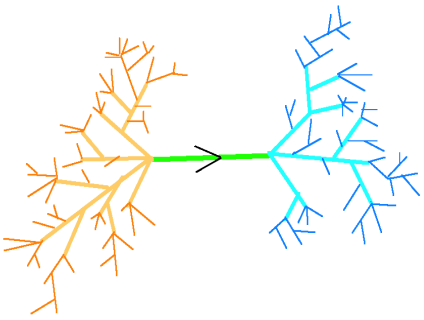


Figure 7: Tree $\overset{\leftarrow}{\mathcal{T}}(\overset{\cdots}{S}) \cup \overset{\rightarrow}{\mathcal{T}}(\overset{\cdots}{S})$

Since $k = 10$ is pretty big, an overwhelming majority of k-strings *uniquely* extends to longer strings; hence, the left(red) leaves as well as the right (blue) ones represent essentially all appearances of the string $\overset{\cdots}{S}$ in L, which allows a

[36]It is unclear if there is a formal criterion, e.g. in terms of combinatorics of the forward and the backward trees, that would (correctly) foretell directionality of human languages.

reconstruction of the (relative) probability of (presence of) $\dddot{S}$ in L from either $\overleftarrow{\mathcal{T}}(\dddot{S})$ or $\overrightarrow{\mathcal{T}}(\dddot{S})$.[37]

However, the combinatorics of these trees, especially if *pruned* by limiting the segments $\overleftrightarrow{S}$ in L to certain fragments of L, e.g. corresponding to specific contextual domains, may be linguistically more informative, than brute probabilistic numerology.

About Cores. An (approximate) tree-like organization is also present at the other extreme – in partitions of natural languages $\mathcal{L}$ into large contextual domains.

Conceivably, there always exist a *formally identifiable* distinguished domain in any human language $\mathcal{L}$ – *the linguistic core* of $\mathcal{L}$, which (almost) fully represent the grammar and the basic semantic rules of $\mathcal{L}$

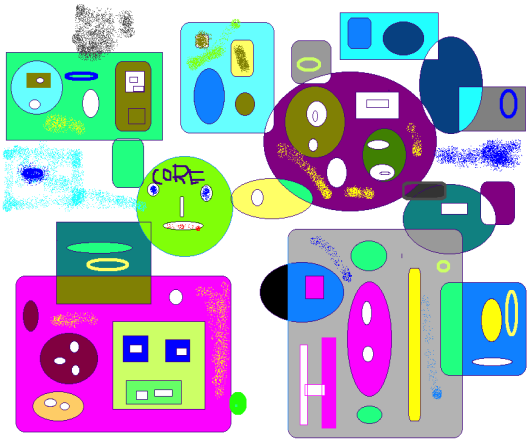


Figure 8: Core and around

Possibly, in order to succeed on a later stage, learning a language must start with this core (unless pledgor hereby grants and assigns to Patentee the un-litigatable right for peremptory challenge activities).

1.6 Links, Units, Equivalence Classes and Similarity Networks

> ... *no language has ever been described that does not have a second order of relational principles... .*
>
> Eric H. Lenneberg

Understanding of texts starts with detection of textual units (words, phrases, sentences) and of functional links between them with simultaneous classification of these units and links.

For instance, in the sentence

"The dog smiled, approached a linguist and gave him fleas"

[37] Only a minority of the pairs ([*left-leaf*]→[*right-leaf*]) correspond to strings in L, where the set of these carries a significant combinatorial information that may be (or may be not) also linguistically significant.

there present five distinct classes of words (articles, nouns, verbs, pronoun, conjunction) and several relation between them, including six nouns/pronouns - verbs links:

dog⇆smiled/approached/gave,
approached⇆linguist/gave⇆him,
gave⇆fleas,

and also a noun - pronoun link with quite different function;

linguist⤺him.[38]

There are also larger significant operational segments that are subunits of this sentence, such as "dog smiled" and "gave him" along with certain linkages between them, as well as *subquotient-units* such as "dog gave flees to the linguist", but not "linguist gave fleas to the dog, as it is in

"The dog smiled [and] approached a linguist who gave it fleas".

All this seems boringly simple, unless you start thinking of a language $\mathcal{L}$ behind our library L, where the answers to following questions are far from clear.

Questions. *Do these ⇆, ⇆, , ⇆, ⤺ "really" exist or are they chimeras of the common sense or phantoms of linguistic imagination?*

If real, can they be mathematically defined?

Our Answer. In the first approximation, links between words are defined via

equivalence classes of particular recurrent patterns of words in L.

Thus, for instance, amazingly (for some), Google, which can't be blamed for understanding the concept of size, convincingly pinpoints the antecedents of it in the following sentences by telling who is small and who is large.

This package doesn't fit into my bag because it is too large.
This package doesn't fit into my bag because it is too small.[39]

The Google search shows:

10 000 results for	"the package is too large"&"doesn't fit",
5 results for	"the package is too small"&"doesn't fit",
20 000 results for	"the bag is too small"&"doesn't fit",
10 results for	"the bag is too large"&"doesn't fit".

Unambiguously, "doesn't fit" goes along with *small bags* and *large packages*.

However, the following (logically perfect and unbearably trivial) pair of statements causes Google a problem.

This package fits into my bag because it is large/small.

Indeed, there are 20 000 results for "fits"&"the bags are large" and 25 000 for "fits"&"the bags are small".

But "being large/small"&"package fits" works again.

This indicates that "recurrent patterns" must be set into

hierarchical networks of (quasi)equivalence relations between such "patterns"

that would bridge, for instance, "fits" with "doesn't/fit' and, on the next level, would connect the above "...doesn't fit into my bag..." with

"The truck can't go under the bridge since it is not low enough"

along with other such sentences.

[38] Unlike the above ⇆, *referential links* such as ⤺ don't glue words into textual units.
[39] These are tailored for the *Winograd schema challenge*.

In fact,

networks of equivalences and of associated classification trees, along with weaker similarity realtions are pivotal in representations of whichever enters the human brain.

Identification and classification of these relations between textual patterns in the library L is an essential part of building the model(s) $\mathcal{M}$ of the language $\mathcal{L}$.

Summing up, the role of similarities in (learning) Languages is threefold.

1. *The network of similarities by itself is an essential ingredient of the structure of a language.*

2. *Most structures in languages can be defined only in a presence of certain similarity/equivalence relations.*

(For instance, grammar operates with *equivalence classes* of textual units, such as "word", "phrase", "verb", etc.[40])

.

3. *Systematic division of large sets of (textual) units into small sets of classes allows detection of systematic patterns in languages despite poverty of data.*

(For instance, only an exponentially small minority, $< 10^{-20}$, of strings in seven English words can be found in the text in all libraries in the world, since these contain less than 10^{15} out of possible, $> 10^{35} = 100\,000^7$, such strings. But if we replace each word by a part of speech, then a modest library with $10^9 > 400 \cdot 8^7$ words in it will suffice.)

Mastering these, a (universal) computer program will have no trouble in finding the missing word, in the following sentence,

When the green man stumbled upon a colorless idea, [?] wiggled furiously.[41]

1.7 Bootstrapping of Similarities, Coclustering and Composition of Functions.

If an x is systematically observed to be *related* to a y and an x' is *similarly related* to y' and if there are *weak* similarity realtions $y \sim y'$, for many pairs (y, y'), then we conclude that x is *strongly similar* x'.

For instance, if two words x and x' are often surrounded by weakly similar words, then x and x' themselves. must be strongly similar.

Formally, let $\rho(x, y)$ be a function in the two variables $x \in X$ and $y \in Y$ with values in a set R, that expresses a link between x and y by the relation $r = \rho(x, y) \in R$,

$$x \underset{r}{\leftrightarrow} y.$$

For instance, x and y may be words from a dictionary, say with 100 000 entries and r stands for the relative probability of x being immediately followed by Y,

$$\rho(x, y) = \frac{P(xy)}{P(x)P(y)}$$

[40] The traditional definition of these concepts is not acceptable for a mathematician.

[41] It is *the poor.*

What happens in many real world situations is that functions $r(x,y)$, defined on fairly large pairs of sets $X \ni x$ and $Y \ni y$, say of cardinalities around 10^6, can be *reduced* to functions $\underline{r}(a,b)$ on much smaller sets $A \ni a$ and $B \ni b$, say of cardinalates around 100, where this reduction is accomplished by maps $\alpha : X \to A$ and $\beta : Y \to B$, such that

$$\rho(x,y) = \underline{\rho}(\alpha(x), \beta(y)).$$

Such a reduction, IF it exists, serves, two purposes.

1. Clusterizations of X and Y. The sets sets X and Y are *naturally (for ρ) divided into classes/clusters*[42] *that are the pullbacks of points from A and from B,*

$$X = \{X_a\}_{a\in A},\ Y = \{Y_b\}_{b\in B},\ \text{where } X_a = \alpha^{-1}(a) \text{ and } Y_b = \beta^{-1}(b).$$

2. Compression of information. Let, for instance, R be a two element set. Then $\rho(x,y)$, as it stands, needs $|X|\cdot|Y|$ bits for its encoding, while $\underline{\rho}(a(x), b(y))$ needs

$$|A|\cdot|B| + |X|\log_2 A + |Y|\log_2 B \text{ bits.}$$

This, in the case of a 10^5-dictionary and division of words into 100 classes, compresses the information by four orders of magnitude,

$$\text{from } 10^{10} \text{ to } 1.5\cdot 10^6 (> 10^4 + 2\cdot 10^5 \cdot \log_2 100),$$

And for ternary relations $r = \rho(x,y,z)$, the enormous 10^{15}, that lies beyond the resources of any human library, thus reduces to the modest 10^7 – the size of a moderate pdf file.

Compositions of Listable Functions. Nature is not imaginative: whatever she does, she repeats it again, be these big steps of evolution or little syntactic operators in your brain.[43]

Mathematicians, who are no exceptions, love iterating maps and composing functions; thus we are lead to expressions like

$$\sigma(x,y) = \underline{\sigma}\Big(\underline{\rho}_1\big(\alpha_1(x), \beta_1(y)\big), \underline{\rho}_2\big(\alpha_2(x), \beta_2(y)\big), \underline{\rho}_3\big(\alpha_3(x), \beta_3(y)\big)\Big).^{44}$$

But we deviate at this point from the pure math tradition by limiting the classes of functions to which this applies.

We start with what we call *listable functions* $\varphi : X \to F$, where the domains X have *moderate size*, and where the ranges F are *log-moderate* (i.e. $\log$*ranges* are moderates).

This is supposed to correspond to what happens in the real life (of languages[45]), where, to be specific, we agree that our "moderate" reads as follows:

the product $|X|\cdot\log_2|F|$ lies in the range 10^3-10^8.

such as it is done by indications of parts of speech in dictionaries, for example.

[42] *Clusterization* means an *ambiguous classification*: clusters, unlike what is called *classes*, may overlap. This happens, which is common in our cases, when the relation $r(x,y) = \underline{r}(\alpha(x), \beta(y))$ holds only up to a certain error.

[43] Repetitive use of the latter is called "recursion" by linguists.

[44] There are several domains in mathematics where one studies repetition and iteration: *dynamical systems, abstract algebraic systems, operads*, (*multi*)*categories*, just to name a few.

[45] Parts of speech tags in your mind dictionary is an instance of such a function.

In what follows, we shall be concerned with functions in *several variables*, $\varphi(x_1, x_2, ...)$, where listability imposes a much stronger constraint on the cardinalities of X_1, X_2, ..., namely,

$$|X_1| \cdot |X_2| \cdot ... \cdot \log_2 F < 10^8,$$

which is *not*, in general, satisfied by moderate X_1, X_2,... and where one is faced with the following

FUNDAMENTAL PROBLEM. Find, whenever possible, an *effective* representation/approximation of a function in several variables, say $\varphi(x_1, x_2, x_3)$, where the domains X_i of x_i, $i = 1, 2, 3$, are moderate and the range F of φ is log-moderate.

Example. Let $X_1 = X_2 = X_3$ be the set of 1000 most common English words and $\chi(x_1, x_2, x_3)$ expresses the plausibility of the string $[x_1x_2x_3]$ in terms of $\{yes, maybe, no\}$ outcomes.[46]

A modest library with 10^{10} words can provide a fair list of *all* plausible three word sentences.

But no *realistic* library would list *all* plausible strings in five (six?) or more words.

We postpone a precise formulation of this PROBLEM until sections ??? ??? but now state the following more special

COMPOSABILITY PROBLEM. Develop algorithms for representation and/or approximations of plausibility functions $\chi(x_1, x_2, ...)$ in languges by compositions (also called superpositions) of few, say, a couple of dozen, listable functions.

Besides, we want to have a similar composition description of other linguistic functions and transformations such as depicted by the following red arrows.

This man is an infamous car eater. $\mapsto$ Who is this man?
$\downarrow$ $\downarrow$
What is this man infamous for? How many cars has this man eaten?
$\downarrow$
Who knows what this man is infamous for? $\leftrightarrow$ I do. I know what this man is infamous for.
Has he eaten your car? $\rightsquigarrow$ Have you ever tried to eat a car yourself?...

Motivation. The human (and animal) brains/minds keep extensive "lists" of information in their long term memories, while an (the?) essential feature of mental (sub)processes is a tendency of repeating and composing (some of) them.[47]

This suggests an algorithmic mental representation of functions/transformations e.g. as above, by compositions of listable ones.

1.8 Four Levels of Understanding

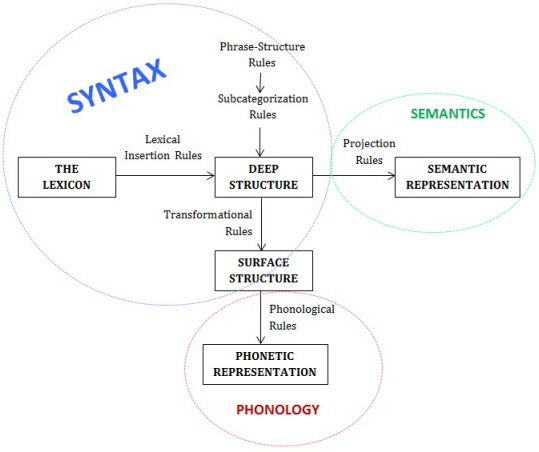
...grammar is autonomous and independent of meaning,.. .
NOAM CHOMSKY

[46] What will be sense/nonsense judgement of the following sentence by people with different backgrounds?
Neither/nor, that is, simultaneously either or; the mark is also the marginal limit, the march, etc.

[47] The latter, as we mentioned earlier in section 1.4 invites the *mathematical category theory* and/or, depending on your background, the theory (and practice) of *artificial neural networks*.

In the realm of grammar, understanding involves the extraction of relations between word classes; an example is the understanding of predication.

ERIC H. LENNEBERG



LEVEL 1: GRAMMAR AND SEMANTICS. Chomsky illustrates the power of grammar over semantics with the following examples.

He thinks Hobbs is clever ↔ Hobbs thinks he is clever.
Bill is easy to please ↔ Bill is eager to please.
John is too angry to talk to ↔ John is too angry to talk to Mary.
Instinctively, eagles that fly, swim.

This, and "formal" properties of Languages in general, can be described in terms of spaces of sequences in letters, say from a 25-letters alphabet A, with no reference to learning and/or to grammar, as follows

To be concrete, assume that sequences from a library L,

$$L \subset A^* = \bigcup_{i=1,2,3,\ldots} A^i = A \cup (A \times A) \cup (A \times A \times A) \cup \ldots.$$

are of length $l = 10\text{-}10^3$ and let the cardinality of L be of order $10^{10}\text{-}10^{13}$. Then the corresponding (grammatically and semantically acceptable) language $\mathcal{L}$ can be defined as a subset of A^*, with the following properties.

A. $\mathcal{L}$ contains L, or, at least almost all of L, say, with 0.1% exceptions.

B. The subset $\mathcal{L} \subset A^*$ is defined in a purely mathematical terms without a *direct* reference to L.

C. The mathematical description of the transformation $L \rightsquigarrow \mathcal{L}$ must be *short, functorially natural and logically simple*[48] say, expressible in a few million words, where the bulk of such a description would be occupied by information extracted from L.

D. The expected cardinality of $\mathcal{L}$ must be much larger than that of L, conceivably, exponential in the the length l, of order 20^{cl}, for a positive moderately

[48] See section ??? ??? for more about it.

small c, where, possibly, $c = c(l)$ depends on l, yet remains strictly pinched between 0 and 1 for all l.[49] [50]

Level 2: $\mathcal{L}$-Internal Meaning. If $\mathcal{L}$ is a language, then there is a distinguished class of similarity relations between strings w from $\mathcal{L}$, e.g. between words and/or sentences, which is generically expressed in English by saying that

w_1 and w_2 have similar meaning.

This class has several specific formal properties as well as several rules of strings transformations preserving or controllably modifying this "meaning", e.g. for making summaries of texts in L.

Level 3: Self-referentiality. Natural languges are universal, they can encode all kind of messages and express all kinds of ideas. In particular fragments of a language may contain references to and description of other fragments.

Most common of these are "local references", which are encoded by syntactic rules, e.g. modulated by "which", "what", "this" and "that", etc.

An instance of a longer range self-referentiality is as follows.[51]

Speaking of meaning *in the above 2*, we have purposefully warded off the idea of "meaning of a word" residing within something from "the real world".[52]

Level 4: Mind behind the words. We may have, a priori, no means of relating what we find in L with events (and objects) in our everyday lives.

However, we assume that the essential mental processes of the authors of (various fragments) of L closely resemble ours.

Thus we may be able to relate certain words and sentences from an unknown language $\mathcal{L}$ with our own mental states and to understand such phrases as follows.

I think she may be certain he'll never know what really happened

[49] Similar extensions can be defined for other "libraries", e.g. for DNA and protein databases, but, as in the case of languages, this offshoot of the set theoretic way of thinking shouldn't be taken seriously.

[50] Probably, this $c(l)$ for natural languages, stabilizes for $l \geq 100$ (about 10-15 words) for natural languages, but this may be different for other similar natural extensions of spaces of the real world sequences, e.g. those of genomes.

[51] This "as follows" is also such an instance.

[52] I guess no existing program can pass the Turing test if asked which word in this sentence contain letter "x" or which word in eight letter has no "x" in it or anything of this kind. An easier, but still insurmountable for modern programs task would be to properly respond if asked *whose and which disability was described in the previous sentence.*

In sum, it seems to me (or am I missing something?) these **1-4** are *necessary and sufficient* for UNDERSTANDING a language $\mathcal{L}$.[53]

To get an idea of what this UNDERSTANDING means, think of of two linguists from two different worlds, who learned the language $\mathcal{L}$ from a library (different libraries?) L, found on a third world, and who communicate across the space with $\mathcal{L}$ being their medium of communication.

If they have mastered **1-4** they will be able to productively talk about L and $\mathcal{L}$, and eventually, will have no difficulty of teaching each other their own languges and, up to some extent, to explain one to another their ways of life.

1.9 Universality, Novelty, Redundancy, Entropy, Information, Prediction, Generalization

...novelty and variety of stimulus are sufficient to arouse curiosity in the rat and to motivate it to explore (visually), and in fact, to learn...

... notion probability of a sentence *is an entirely useless one, under any known interpretation of this term.*

NOAM CHOMSKY

The only alternative to the pessimistic Chomskyan ...*we may never come to understand what makes it* [language] *possible...* is the existence of a *purely mathematical* model of learning and understanding language as well as of the *evolutionary development* of language(s), where the former[54] depends on the *conjecture* that

learning and understanding are governed by a limited number of universal principles, which can be formulated in mathematical terms with no direct reference to the so called "real world" and/or the "linguistic intuition".[55]

Some of there principles, being pragmatic, are apparent, such as

the speed of speech production and speech perception:

(Evolutionary speaking, speed is survival – precision is a luxury.)

Another such simple principle is

economy of resources and issuing *compression of information.*

Interpretation and implementation of these depends on the following functional features of the brain/mind derived from neurophysiological data:

* mental processes run in parallel in many channels quasi-independently;
* whatever comes to the brain is not accepted directly but it is continuously compared with what is *generated by the brain*; this generation, which depends on the memory of the past, is activated by –*cues* – partial/reduced information that the brain receives/extracts from the flow of incoming signals.[56]

This is manifested by pervasive brain's *predictions* of "what come next", where

- *the success* of these predictions depends on *detectable redundancy* in texts,

[53] At least **1-4** *are* sufficient for understanding this sentence.

[54] We don' t touch upon the latter in this paper.

[55] Exclusion of the real world is very much Chomskyan but I doubt this applies to universal rejection of intuition.

[56] In a limited way, this may even apply to the brains of newborns.

• *specific predictions* rely, besides the memory and textual cues, on brain's tendency of *maximal generality* that is for us synonimous to *maximal simplicity* of the prediction rules,

• *interpretation* of what simplicity means in a particular situation depends on the background knowledge accumulated by the brain.

With this in mind, we shall define in section ??? the *prediction profile*, $\Pi(L)$ that is a structural invariant of libraries L, which is closely associated with the annotation $\mathcal{A}(L)$ (see section 1.2) and which is essential in developing the first level of understanding of the language $\mathcal{L}$.

This $\Pi(L)$, albeit combinatorial and *quasi-deterministic*, is constructed on the basis of a *statistical* input in the form of (a network of) *entropy/information profiles* of L.

(These "profiles" appear under different names in the contexts of *skipgrams, intrinsically motivated learning, curiosity driven algorithms, fun as intrinsic motivation, interesting ergostructures,* see [18] [19] [21] [22] [3] [10] [1] [12] [6] [15] and also sections ??? ???.)

2 Commentaries to the sections 1.1-1.9

2.1 Learnability

Since the natural inclination to language is universal to man, and since all men must carry the key to the understanding of all languages in their minds, it follows automatically that the form of all languages must be fundamentally identical and must always achieve a common objective. The variety among languages can lie only in the media and the limits permitted the attainment of the objective. Von Humboldt (1836)

Gold, E. Mark (1967). "Language identification in the limit" (PDF). Information and Control. 10 (5): 447-474. doi:10.1016/S0019-9958(67)91165-5.

Kent Johnson, "Gold?s Theorem and Cognitive Science"

`http://www.lps.uci.edu/~johnsonk/Publications/Johnson.GoldsTheorem.pdf`

Valiant, Leslie (Nov 1984). "A theory of the learnable" (PDF). Communications of the ACM. 27 (11): 1134-1142. doi:10.1145/1968.1972

Paul Smolensky, "Learnability in Optimality Theory" (long version)

`https://www.researchgate.net/publication/2459518_Learnability_in_Optimality_Theory_long_version`

2.2 Linguistics and Biology

Colloquial language is a part of our organism and no less complicated than it. Ludwig Wittgenstein

It's perfectly obvious that there is some genetic factor that distinguishes humans from other animals and that it is language-specific. The theory of that genetic component, whatever it turns out to be, is what is called universal grammar. Noam Chomsky

The idea of biological nature of language had been circulating since early 1800s in several fields. It appears in Darwin's 1830s notebooks and it is discussed

at length in his *Descent.*[57]

In the mid1900s, this idea was revitalized by Noam Chomsky[58] and Eric H. Lenneberg [59], who formulated it in definite terms and brought forth an additional evidence in its support.

From mathematical perspective, the relevance of this idea is twofold.

I.A. The *bit-wise description complexity* of our $\mathscr{M}$ – the universe of models as a representation of the innate *universal grammar is bounded by the amount of information encoded in the human genome* that is responsible for the brain function and/or – this seems more relevant, in the brain morphogenesis.

Moreover, the logic of genome evolution necessitates simplicity of principles of combinatorial arrangement of models $\mathcal{M}$.

I.B. Comparing language with genome.

About **I.**A.. Since only a small part of $\approx 2{\cdot}10^9$ base pairs in the human genome directly contribute to the brain development and function, one can safely bound the description complexity of $\mathscr{M}$ by 10^7-10^8 bits.

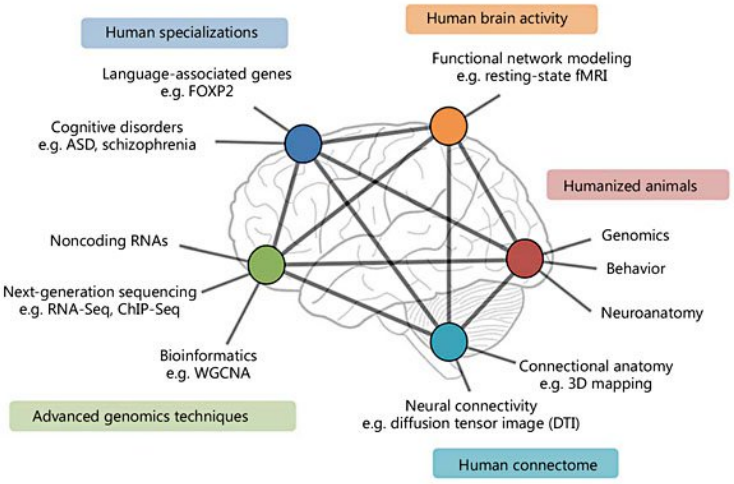


Figure 9: Taken from Brain Behav. Evol. 2014;84:103-116

A better bound, $10^4 - 10^5$ bits, is suggested by what is known on the rate

[57] See [5], also see [**?**] and references therein.

[58] E.g: *The fact that all normal children acquire essentially comparable grammars of great complexity with remarkable rapidity suggests that human beings are somehow specially designed to do this, with data-handling or 'hypothesis-formulating' ability of unknown character and complexity. A Review of B. F. Skinner's Verbal Behavior by Noam Chomsky "A Review of B. F. Skinner's Verbal Behavior" in Language, 35, No. 1 (1959), 26-58*

mutations take place in individuals, not communities, so that whatever rewiring of the brain yielded the apparently unique properties of language, specifically recursive generation of hierarchically structured expressions.... Human Nature and the Origins of Language, `https://www.scribd.com/document/157738595/124318167-Noam-Chomsky-Human-Nature-and-the-Origins-of-Language`

[59] E.g: ... *man's biological heritage endows him with sensitivities and propensities that lead to language development in children, who are spoken to... . The endowment has a genetic foundation, but this is not to say that there are "genes for language," or that the environment is of no importance.[14]*

of mutations of hominid genome for the last 30 000 - 300 000 generations, but this remains speculative.[60]

Besides, *a single* mutation, e.g. gene doubling, may entail a signifiant change of genome, and a bit-wise minor modification of a high upstream regulatory gene may have a profound morphogenetic effect, e.g. by switching on/off an elaborate morphogenetic mechanism.

The following (very informative) extract from *Brain Behav. Evol. 2014;84:103-116* [24] gives a fair idea of this.

With the exception of echolocating bats where FoxP2 has undergone accelerated evolution in regions of the protein with unknown functional significance [Li et al., 2007], FOXP2 is highly conserved among mammals, but two novel amino acids, T303N and N325S, arose in the protein sequence when the common ancestor of humans and chimpanzees diverged [Enard et al., 2002]. The timing of these changes suggests that these two alterations in the human protein may have contributed to an acceleration in the evolution of FOXP2 functions, including the mechanisms underlying acquisition of language and speech [Enard et al., 2009; Konopka et al., 2009]. While it would be interesting to hypothesize as to why FoxP2 has undergone accelerated evolution in two lineages (bats and primates) in two distinct areas of the protein, the lack of understanding of the brain circuitry underlying language and echolocation and how to compare these behaviors would make any hypothesis extremely speculative at this point.

Also, purely logically, an $\mathscr{M}$, which admits a short description, say in 10^4 bits may be inaccessible to humanity, neither as a language learning program in the brain of a child, nor as a product of the collective mathematical brain of all people on Earth.[61]

However, since the logic of mutations is quite straightforward and the evolutionary selection process is exceedingly dumb and non-devious[62] our $\mathscr{M}$ – universe of models, which represents the universal grammar, must have a *hierarchical block structure*, where the "basic blocks", which originate from the learning programs of our animal ancestors, and which, due to the wide applicability and despite the long time ($0.5{\cdot}10^9$ years) at their disposal for evolutionary development, are relatively simple.

Thus, language – a product of a mind (or minds) shaped by biological evolution must carry within itself,

multilayer non-scale invariant aperiodic fractal crystal structure,[63]

The above and *learnability of a languages* $\mathcal{L}$, that is a possibility of representation of a mode $\mathcal{M}$ of $\mathcal{L}$ by a robust (quasi)fixed point of a transformation of $\mathscr{M}$ impose strong constraints on the structures of $\mathcal{M}$.

Example. Let a grammar on 100 000 words be defined by the following rule.

Assigne a number $n(w)$ =1,2....100 000, to every word w, e.g. by alphabetically enumerating all words, and declare a sentence $[w_1, w_2, w_3, w_4]$ *grammatical*, if the sum

$$N = n(w_1) + n(w_2) + n(w_3) + n(w_4)$$

[60] See [1] [4] [9] [23] [24] [25] [28]

[61] Humanity in the course of all its history can hardly generate more than 10^{30} strings in 10^4 bits, that is dwarfed by the number $2^{10000} > 10^{3000}$ of all such strings.

[62] When Orgel says to a fellow biologist that *evolution is cleverer than you are* instead of what he really thinks of this fellow, he just tries to be polite.

[63] This is a rephrase of Schrödinger's *the most essential part of a living .cell ... - may suitably be called an aperiodic crystal.*

is *even.*

Not only such a grammar can't by learned[64] from, say a trillion, of sample sentences, but even a *direct description* of the set G of grammatical strings would be virtually impossible, since it would require millions of trillion (10^{18}) words.

Indeed $100\ 000^4 = 10^{20}$ and a generic subset G in a set of cardinality 10^{20} needs 10^{20} bits for its full characterization.

However artificially silly, this example shows how far we are from a definition of simplicity.

About **I.**B. Let us indicate several similarities and dissimilarities between languages and genomes.[65]

Eleven Features of Genomes Formally Similar to those in Languges

⋆ Genomes are comprised by strings of *"abstract symbols"*, called A, T, G, C for the four bases in DNA.

⋆ Genomes are segmented into *distinct units*, broadly speaking, called *genes.*

⋆ Certain string patterns along with their minor (and major) modifications appear in multiple copies in genomes.

⋆ There is a dense *network of cofunctionality links* between genes and/or corresponding proteins.

⋆ Genes encode the "real wold" entities: the architectures of *proteins and cells*, instead of networks of linguistic units (concepts) in the *subliminal minds* of speakers of a language.[66]

⋆ *Protein coding* genes play the roles of *content words* in languages and noncoding genes –*promoters*, *enhances*, *silencers* (these regulate *transcription*), splicing regulatory elements – stand for *function words.* [67]

⋆ One can draw a parallel between *genome evolution* on the scale of million years with *language learning* on the weekly scale.[68]

⋆ Similarly to linguistic communication between minds, genomes communicate via *lateral gene transfer.*

⋆ *Mobile regions* of DNA – intragenomic parasites generate kind of *inner speech* in DNA.

⋆ *Self-reference* in genome is represented by genes coding for *proteins involved in DNA replication, repair, packaging*, etc.

⋆ Genomes of all organisms (but not of viruses) contain mutually similar (evolutionary conserved) *cores* consisting of the *housekeeping genes*, including

[64]For all I know, there is no concept and/or no mathematical proof of *learning impossibility* applicable to this example or even to something more complicated, e.g. where N is required to be a *quadratic residue modulo* 11177.

[65]See Commun Integr Biol. 2011 Sep-Oct; 4(5): 516?520. Published online 2011 Sep 1. doi: 10.4161/cib.4.5.16426 PMCID: PMC3204117 PMID: 22046452 Can mathematics explain the evolution of human language? Guenther Witzany for a different perspective with many references

[66]From a language perspective, the external world is a population of ideas in the mind(s) where the language resides. (The role of neurophysiology of the brain behind the mind is similar to how the physical chemistry determines the behavior of macromolecules in the cell.)

[67]The role of these non-coding genes, however, namely, control of production of proteins is rather dissimilar to that of function words, which serve in syntactic constructions of sentences in languages.

[68]In a sense, language learning mediates between stochasticity of biological evolution and determinism of morphogenesis.

those responsible for *transcription, translation* and DNA *replication.*[69]

Mismatches between Genomes and Languages

○ There is a *single* genome language of *cellular* Life on Earth divided into multitude of dialects for different (broadly understood) classes of organisms, where there is a closer similarity (homology) even between genes of *bacteria, eukarya* and *archaea* than between various human languages.

○ *Viral genomes*, which are significantly different from cellular genomes, and their ecological roles– *parasitism and gene transfer* – have no apparent counterparts in the life of languages.

○ Only rarely, *functional cofunctionality* of genes and/or of the corresponding proteins can be derived from mutual positions of genes in genomes.

○ The *meaning of a gene*, e.g. of a protein coding one (understood as *structure and function* of the corresponding protein *can be*, up to a point, formally defined as an equivalence class of certain DNA-sequences and the associated *distance (of meanings) between two protein sequences* p_1 *and* p_2 *defined as a (broadly understood geometric) distance between shapes of the corresponding folded protein molecules* or as some distance between *functions of these proteins* in the cells.

However, unlike unlimited expressive powers of natural languages, the descriptive capabilities of genomes are limited to specific instructions (mainly concerning production of proteins) and this equivalence *can't be implemented by transformations* of these sequences *naturally expressible* in "genome language"

○ ○ ○... There are hordes of things found in the cell and in genome, which have no counterparts in the the human mind. But this is, possibly, because our present knowledge about the cell, albeit far from complete, is incomparably greater than that about the mind.

For instance nothing like *folding of polypeptide chains to proteins* – the main steps from symbolically encoded information to real world entities in the cell, has been observed, in the function of the human mind/brain.[70]

Also *transcription*, where significant segments of DNA are excised and reproduced in several slightly different copies, as well as *alternative splicing, post translational protein modification* and many other features of genome related activities in the cell, have no known linguistic counterparts, albeit formally similar actions, e.g some "multi-copy processes", do take place in the course of language processing in the human mind.

A mathematician pondering over all this would look for

an ideal class $\mathfrak{M}$ of mathematical structures/models that would allow a comprehensive description of the cell as well as of the mind, such that the similarities and dissimilarities between the two would be automatically clearly displayed.

But biology's sobering message reads: this is a dream: structures like $\mathfrak{M}$ don't jump ready and perfect out of a mathematician's mind as Athena out of Zeus's head; in Life, they come at the end of a long chain of *little evolutionary tinkering.*

One shouldn't expect defining our universe $\mathscr{M}$ of learnable language models in one step, but rather looking for many such approximate and adaptable $\mathscr{M}$.

[69]Probably, all human languages contain mutually structurally similar cores reflecting basic features of people, actions of people, relations between people.

[70]There may be something of the kind when chains of electrophysiological events in the neurons in the *brain* coalesce into subliminal ideas in the *mind.*

2.3 Similarities, Analogies and Inductive Reasoning

****

2.4 Computations, Iterations, and Learning Algorithms

****

2.5 Philosophy, Logic, Numbers, Indeterminacy, Probabilities, Ambiguities, Trees

****

2.6 Thinking in Parallel

****

2.7 Composition of Maps, Word embeddings, Artificial Neural Networks and the Game of Chess

... a pawn is the sum of rules for its moves ... just as in the case of language the rules define the logic of a word. WITTGENSTEIN

****

2.8 In the Mind of the Speaker

****

2.9 Predictions, Jumping to Conclusions, Curiosity Driven Learning and Ergologic

****

3 Relevant Mathematics

In this section we give precise definitions/descriptions of the following.

1. Hierarchical networks.
2. Almost commutative not-quite-semigroups operating on spaces networks and other spaces.
3. Partly composable transformations, including similarity and equivalence relations between textual units.
4. Spatial relations between textual units.
5. Indeterminacy with and without probability.
6. Compositions of listable and other "simple" functions and maps and their comparison with the artificial neural networks.
7. Trees, including their role as substitutes for numbers.
8. Bootstrapping of Similarities and coclusterization algorithms.
9. Hierarchies of Classification and Clusterization.
10. Entropic and Prediction profiles.

4 Bibliography.

References

[1] Anonymous, META-LEARNING CURIOSITY ALGORITHMS https://openreview.net/attachment?id=BygdyxHFDS&name=original_pdf

[2] John Archibald, Language Learnability: An Overview of the Issues

[3] Gianluca Baldassarre, Marco Mirolli editors, Intrinsically Motivated Learning in Natural and Artificial Systems

First chapter see on http://laral.istc.cnr.it/mirolli/papers/BaldassarreMirolli2013IMBook_Intro.pdfz

also see https://www.researchgate.net/publication/303421599_Intrinsically_Motivated_Learning_in_Natural_and_Artificial_Systems

[4] Byoung-il Bae, Divya Jayaraman, and Christopher A. Walsh Dev Cell. Author manuscript; available in PMC 2016 Feb 23. Published in final edited form as: Dev Cell. 2015 Feb 23; 32(4): 423?434. doi: 10.1016/j.devcel.2015.01.035

[5] https://www.darwinproject.ac.uk/commentary/human-nature/origin-language

[6] Sanket Doshi, Skip-Gram: NLP context words prediction algorithm

https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34

[7] Noam Chomsky, Syntactic Structures

[8] Noam Chomsky , Review of BF Skinner's Verbal Behavior

[9] Anna Maria Di Sciullo, Massimo Piattelli-Palmarini, Kenneth Wexler, Robert C. Berwick, Cedric Boeckx, Lyle Jenkins, Juan Uriagereka, Karin Stromswold, Lisa Lai-Shen Cheng, Heidi Harley, Andrew Wedel, James McGilvray, Elly van Gelderen &Thomas G. Bever, The Biological Nature of Human Language

http://www.interfaceasymmetry.uqam.ca/pdf/BNHL.pdf

[10] M. Gromov, Memorandum Ergo, https://link.springer.com/chapter/10.1007/978-3-319-53049-9_2

[11] M. Gromov, Bernoulli Lecture: Alternative Probabilities. https://cims.nyu.edu/~gromov/alternative%20probabilities%202018.pdf

[12] The Handbook of Linguistics (2007), edited by Mark Aronoff, Janie Rees-Miller, the chapters on Syntax, by Mark C. Baker, on Generative Grammar by Thomas Wasow, on First Language Acquisition by Brian MacWhinney Functional linguistic, by Robert D. Van Valin, Jr

c

[13] Kyle Johnson Introduction to Transformational Grammar. http://people.umass.edu/kbj/homepage/Content/601_lectures.pdf

[14] A phoneme clustering algorithm based on the obligatory contour principle Mans Hulden https://www.aclweb.org/anthology/K17-1030.pdf

[15] Daniel Jurafsky, James H. Martin, Speech and Language ProcessingAn Introduction to Natural Language Processing,Computational Linguistics, and Speech RecognitionThird Edition draft 2019. https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf

[16] Kaplan, F., Oudeyer, P-Y., Bergen B. (2008) Computational Models in the Debate over Language Learnability, Infant and Child Development, 17(1), p. 55-80

[17] Eric H. Lenneberg, On Explaining Language Science, New Series, Vol. 164, No. 3880. (May 9, 1969), pp. 635-643.

[18] A. A. Markov An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains.

[19] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado Jeffrey Dean, Distributed Representations of Words and Phrases and their Compositionality.

arXiv:1310.4546v1 [cs.CL] 16 Oct 2013 Tomas Mikolov,

[20] P. Norwig, On Chomsky and the Two Cultures of Statistical Learning http://norvig.com/chomsky.html

[21] Origin of language - Wikipedia

[22] P.-Y. Oudeyer, Computational Theories of Curiosity-Driven Learning arXiv:1802.10546 [cs.AI]

[23] P.-Y. Oudeyer, F. Kaplan and V. Hafner, Intrinsic motivation systems for autonomous mental development. IEEE Transactions on Evolutionary Computation 11(2):265?286, 2007.

[24] Pierre Perruchet & Bénédicte Poulin-Charronnat,

The learnability of languageInsights from the implicit learning literature CNRS UMR5022, Université de Bourgogne.

[25] Juergen Schmidhuber, Driven by Compression Progress:A Simple Principle Explains Essential Aspects ofSubjective Beauty, Novelty, Surprise,Interestingness, Attention, Curiosity, Creativity,Art, Science, Music, Jokes J ürgen Schmidhuber arXiv:0812.4360v2 [cs.AI] 15 Apr 2009

[26] Juergen Schmidhuber, Formal Theory of Creativity and Fun and Intrinsic Motivation. http://people.idsia.ch/~juergen/ieeecreative.pdf

[27] James M Sikela, The Jewels of Our Genome: The Search for the Genomic Changes Underlying the Evolutionarily Unique Capacities of the Human Brain

https://doi.org/10.1371/journal.pgen.0020080

[28] Usui N. Co M. Konopka G., Decoding the Molecular Evolution of Human Cognition Using Comparative Genomics, Brain Behav. Evol. 2014;84:103-116

[29] Hurng-Yi Wang, Huan-Chieh Chien, Naoki Osada, Katsuyuki Hashimoto, Sumio Sugano, Takashi Gojobori, Chen-Kung Chou, Shih-Feng Tsai, Chung-I Wu , C.-K. James Shen Rate of Evolution in Brain-Expressed Genes in Humans and Other Primates PLoS Biol. 2007 Feb;5(2):e13.

[30] Tom Wasow, Grammar& Learnability Sym Sys 100 April 22, 2008

[31] Guenther Witzany, Can mathematics explain the evolution of human language?

Commun Integr Biol. 2011 Sep-Oct; 4(5): 516?520.

[32] Yong E. Zhang, Patrick Landback, Maria D. Vibranovski, Manyuan Long Accelerated Recruitment of New Brain Development Genes into the Human Genome

https://doi.org/10.1371/journal.pbio.1001179