Understanding Languages and Making Dictionaries.

Misha Gromov

January 25, 2015

Contents

Abstract

We discuss possible designs of algorithms for learning to *understand* "texts" in a given *library* L where these "texts" may be rather general arrays of "signals". We want to represent such an *understanding* by a certain combinatorial structure, called *(ergo)-dictionary* D, that, in particular, encodes algorithms for building this very D out of L as well as for using D for facilitating/speeding-up *understanding* of "texts" in other *libraries*.

1 Idea of Learning.

We want to understand human *"understanding"* up to a point where we shall be able, for instance, to design a simple computer program for evaluation of *plausibility* of phrases, e.g. such as the following two.

1. *Most cats like to eat grass.*
2. *Some cats like to eat mice.*

Understandably, 1 is more plausible than 2;[1] yet, it is not apparent how to formally derive this "*understand*" from the *observed distribution* of the constituents of *these and related phrases* in the body of the language.

What should a program ask Google to figure out *Who Talks*.

For instance dogs "talk" on thousands Google pages but *"do dogs ever tell"* is not found there; yet, *"do boys ever tell"* does appear on 10 pages.[2]

1.1 Signals and Learning.

Our study of linguistic structures will follow the guidelines of what we call *ergologic* that is based on the following premises.[3]

• Flows of signals coming from the external world carry certain structures "diluted" in them.

Learning is a process of extracting these structures and incorporating them into learner's own *internal structure*.

• The essential learning algorithms are *universal* and they indiscriminately apply to all kind of signals.[4]

• Universality is incompatible with any a priori idea of "reality" – there is no mental picture of what we call "real world" in the"mind" of the learner.

The only *meaning* the learner assigns to "messages" coming from outside is what can be expressed in terms of (essentially combinatorial) *structures* that are recognised and/or constructed by the learner in the process of incorporating these "messages" in learner's internal structure.

• Universality also implies that the actions of the learner – building internal structures and generating signals, both within itself and/or released outside,[5] *are not governed by goals* expressible in terms of the external world.

The learning is driven by learner's *"curiosity" and "interest"* in structural patterns the learner recognises in the incoming flows of signals and in the learner's delight in the logical/combinatorial beauty of the structures the learner extracts from these flows and the structures the learner builds.

Essential ingredients of the learning process are as follows.

• The learner discriminates between familiar signals and novelties and tries to match new signals with those recorded in its memory.

• The learner tries to *structurally extrapolate* the signals already recorded in its memory in order *to predict* the signals that are expected to come.

• Besides the signals coming from the external world, the learner perceives, records and treats some *signals internally generates* by the learner itself.

• The learner tends to *repeatedly imitate signals* being received, including some signals that come from within itself.

[1] It is 7:0 on Google

[2] It took me a while to find this example and it is not very convincing.

[3] The "ergo-prnciples" you see below summarize what is written in our articles *Structures, Learning and Ergosystems* and *Ergostructures, Ergologic and the Universal Learning Problem* on www.ihes.fr/~gromov.

[4] The learner's behaviour, that is learner's interaction/conversation with incoming signals, also depends on the learner's internal structure that has been already built at a given point in time. In particular, a prolonged exposure to a particular class of signals makes learner's behaviour more specialised (more efficient?) while learner's ability to absorb and digest different kinds if signals declines.

[5] These are the "actions" the human brain is engaged in.

(The repetitiveness of their basic operations allows a description of learning processes as *orbits under some transformation in the space of internal structures of a learner*. The learning program that implements this transformation must be quite simple and the learning process must be robust. Eventually, "orbits of learning" stabilise as they approach approximately fixed points.)

• The learner tends *to simplify signals* it tries to imitate.

• The learner systematically makes guesses and "jumps to conclusions" by *making general rules* on the basis of regularities it sees in signals.

• When the learner finds out that a rule is sometimes violated, the learner does not reject the rule but rather adds *an exception*.

• The learner tends to use *statistically significant* signals for building its internal structure as well as for making predictions. But sometimes, the learner assigns significance to certain exceptionally rare signals and use them as essential structural units within itself.[6]

• The learner probabilistic reasoning in uncertain environment is *yes-maybe-no* logic.

We impose the following restrictions on the abilities of our intended learner programs that are similar to those the human brain has.

• The learner *does not accept unstructured sets* with more than four-five items in them; upon encountering such a set the learner invariably assigns a certain structure to it.[7]

• The learner has *no built-in ability of sequential counting* beyond 4 (maybe 3); we postulate that $5 = \infty$ for the learner.

In particular, the learner is not able to produce or perceive five consecutive iteration of the same process, unless this becomes a *routine* delegated from *cerebral cortex* to *spinal cord*.[8]

Our main conjecture is that *universal learning algorithms exist* and, moreover, their formalised descriptions are quite simple.

The time complexity of such an algorithms must be at most log-linear (with no large constant attached) and the performance of an "educated/competent program" must be no worse than logarithmic.

In fact, the essential features of (ergo)learning as we know it, make sense only on a roughly "human" time/space scale: such a learning may apply to flows of signals that carry $10^9 - 10^{15}$ bits of information all-together and one hardly can go much beyond this.[9]

Ergo-logic, Universality and Doublethink. If one expects an analysis of a *flow of signals*, e.g. of a collection of texts in some language $\mathcal{L}$ to be anywhere close to the TRUTH, and if one wants to design an algorithm for learning $\mathcal{L}$, one must,

[6] It is the rare words in texts that are significant, not the most frequent ones.

[7] Partition of stars in the sky into constellations is an instance of this.

[8] Never mind *the kid that fought his dad that bought the car that struck the bike that hit the truck that brought the horse that kicked the dog that chased the cat that caught the rat that ate the bread.*

[9] The universal learning systems themselves, e.g. those residing behind our skulls, have no built-in ideas of *meaning*, of *time*, of *space*, of *numbers*. But any speculation on natural or artificially designed "intelligent" systems strikes one as *meaningless*, if *spacial* and *temporal* parameters of possible implementations of such systems are *not specified* and set within realistic *numerical* bounds.

following ergo-logic, *disregard* all one a priori knows about this $\mathcal{L}$, *forget* this is a language, *reject* the idea of meaning associated to it

But the only way to evaluate the soundness of your design prior to a computer simulation of it, is to compare its performance to that of the corresponding algorithms in a human head.

1.2 Discretization, Classification, Interaction.

The process of learning mainly consists in *structuralizing* the incoming flows of signals by identifying redundancies in these flows and representing "compressed flows" of these signals in a structurally efficient way.

It is a fundamentally unresolved problem in psychology to identify *mathematical classes of structures* that would *model mental structures* built by human brains that assimilate incoming "flows of signals".

We do not know what, specifically, these structures are but their three ingredients are visible.

1. *Discretization and Formation of Units.* The structures built in the course of learning are *assumed to be discrete*, i.e. composed of *distinct units*.

There are several mechanisms, call them *discretizers*, responsible for formation of these units.

Initially, such units are obtained by "meaningful segmentation" of incoming flows of signals, such, for instance, as division of texts into *words* and *phrases*.

Eventually, everything deserving a name becomes a unit.

2. *Classification and Reduction of Units.* This is called *categorisation* by linguists, it is often implemented by *clusterization* algorithms and depicted by arrows

$$u \mapsto v = class(u)$$

called *reductions*. There are several distinct mechanisms/algorithms of classifications, call them *classifiers*, that run in parallel.

Some classifiers find and/or establish *similarity relations* between different units. For instance, words are divided according to their grammatical functions, such as the traditional (and contraversial) division into eight "parts of speech".

Classes, as they are being formed, are incorporated as *units* in the learner's structure.

3. *Connections between Units.* Some units, be they incoming or internal, have non-trivial *connections* between them, also regarded as *relations* and/or *interactions*. These are found, identified and enregistered by several algorithms, called *connectors*.

Predominant number of connections link *pairs* of units – these are depicted by coloured edges between units where the colors represent (the names of) the corresponding connectors.

A significant part of connectors search for *similarities* between different units.

A quite different groups of connectors is occupied with finding pairs (also triples and possibly, quadruples) of units that *perform together* certain functions. This "togetherness" is manifested by systematic co-appearance of the corresponding units.

Another instance of a connection/relation is that between a unit and the class assigned to it by some classifier.

After singling out these three different, yet mutually interdependent, processes, we must design algorithms implementing them where these algorithm must be universal as well as simple.

Then, granted these algorithms perform as they should, we shall be able(?) to decide if there is *some unknown "else"* within human mind crucially involved in the "learning to understand" process that is fundamentally different from formation of units, their classification and their combinatorial organisation according to their connections and interactions.

The fundamental difficulty we face here appears when we attempt to structuralize not only incoming flows of signals, but also those *created and circulating within* learning system itself, where these "internal flows" are not, at least not apparently, grounded on any structure similar to what underlies "true flows": the linear (temporal or spacial) order between signals.

The data obtained in this regard by neurophysiologists and psychologists do not tell us, at least not directly, how to proceed – we take our cues from what mathematics has to offer.

But when selecting mathematically natural algorithms, we keep in mind possibilities and limitations of their (potential) realisation by the brain: such an algorithms can not have many (say, more than 5) *consecutive operations* on each round (unit) of computation (that, roughly, corresponds to what we routinely do on 1 second time scale); yet, allowing several thousand operations running in parallel.[10]

1.3 Learning a Language.

We want to implement the process of *learning a language* $\mathcal{L}$ by an *orbit* of the *universal learning program* PRO that acts on the *linguistic space of* $\mathcal{L}$ and where this orbit must eventually converge to "*I understand* $\mathcal{L}$" state/program.

The principle existence of such a program PRO is demonstrated by the linguistic performance of the brain of (almost) every child born on Earth that receives flows of *electro-chemical signals* some of which come from linguistic sources and the "meaning" of which the child's brain learns to "understand".[11]

Another, closer to our experience scenario is that of a visitor from another Universe[12] who attempts to "understand" what is written in some human "library", e.g. on the English pages of internet.

In either case, the process of what we call "understanding" is interpreted as assembling *an (ergo)dictionary* D – a kind of "concentrated extract" of the combinatorial structure(s) that are present (but not immediately visible) in flows/arrays of linguistic signals.

The grammar of a language makes a part of D where the structural position of this grammar in D is supposed to imitate how it is (conjecturally) organised in the human mind.

[10] This parallelism is the "technical reason" why our basic mental (ergo)processes are inaccessible to our sequentially structured conscious minds.

[11] Bridging linguistic signals to non non-linguistic ones is an essential but not indispensable ingredient of "understanding Language" as it is witnessed by the linguistic proficiency of deafblind people.

[12] No imaginable Universe appears as dissimilar to ours as what the brain "sees" in the electrochemical world where the brain lives.

A particular dictionary $D = D(L) = D_{PRO}(L)$ is obtained from a collection of texts in some language, called a *library* L, according to some universal (functorial?) process/program PRO that drastically reduces the size of L and, at the same time, endows what remains with a *combinatorial structure* – a kind of "network of ideas", that is similar to but more elaborated than the structure of a *(partially directed) graph.*

This D can be thought of as (a record of) *understanding* of the underlying language by the learner behind PRO. This understanding, call it U_t, is time dependent, with D being an *approximate fixed point* of the learning process

$$U_{t_1} \underset{PRO}{\rightsquigarrow} U_{t_2},\ t_2 > t_1,$$

where, a priori, PRO can be applied to "understandings" U that were not necessarily built by this PRO.

The essential problem here is finding a *uniform/universal* representation that can be implemented as a *coordinatization* of "the space of understandings" U where a simple minded program PRO could act by consecutively adjusting "coordinates" $u_1, u_2, ...$ of $U$and where this space would accommodate incoming loosely structured flows of signals encoded by libraries as well as rigidly organised dictionary structures.[13]

Among relevant concepts and building blocks of an "understanding dictionary" and processes for assembling them we envisage the following.

Short range correlations,[14] *segmentation and identification/formation of units in flows of linguistic signals.*

Memory, information and prediction on different levels of structure.

Similarities, equalities, contextual classification, cofunctionality and coclustering.

Local and non-local, links and hyperlinks.

Tags, annotations, reduction, classification, coordinatization.

Structuralization and compression of redundancies.[15]

Ability and tendency for repetition and imitation.

Fast recognition of known, unknown, frequent, significant, improbable, nonsensical.

Evaluation of degree of "playfulness" or "metaphoricity" of words, phrases and sentences.[16]

Recognition of self-referentiality.[17]

Evaluation of parameters of ability/quality of predictions:
speed, precision, specificity, rate of success, the volume of the memory and the numbers of parallel and sequential "elementary operations" employed, etc.

[13]We know that such programs are fully operational in the brains of 2-4 year old children.

[14]Relative frequencies of "events" are essential for learning a language but such concepts as "probability", "corellation", "entropy", can not be applied to languages, without reservation.

[15]The essence of understanding is not so much extracting "useful information" but rather understanding the structure of redundancies in texts. Non redundant texts, such telephone directories, for instance, do not offer much of what is worth UNDERSTANDING.

[16]Playfulness is the first manifestation of what we call "ergo" in humans and certain animals.

[17]Omnipresent self-referentiality, along with "playfulness", distinguishes languages from other flows of signals. The simplest instance of this is seen in *noun↔pronoun* linkages.

Using a Dictionary. The essential role of a dictionary D of a language $\mathcal{L}$, be it fully or partially assembled, is for reading and understanding texts in this language. From some moment on, the process of learning becomes an *elaborate interaction* of D with texts in $\mathcal{L}$ that is guided by *the simple* core program in PRO.

1.4 Teaching and Grading.

A universal language learning problem PRO is supposed to models a mind of child and it needs only a minimal help from a "teacher", such as ordering texts according to their complexity[18] and allowing PRO a flexible access to texts.

On the other hand, evaluation of the *quality of understanding* by PRO is harder (albeit much easier than designing a learning program itself), since no one has a clear idea what *understanding* is.

Our formal approach is guided, in part, by how it goes in physics, where an unimaginably *high level of understanding* is reflected in the *predictive power* of mathematically formulated *natural laws* that encapsulate enormously *compressed data*.

This lies in a category quite different from what we call *"knowledge"*.

For instance, ancient hunters *knew* more of how planets wander in skies than most modern people do. But understanding of this wandering depends on "compression" of this knowledge by setting it into the slender frame of mathematically formulated laws of motion.[19]

Similarly, understanding languages depends on compression of structural redundancies[20] in flows of linguistic signals. albeit this compression is not as substantial as in physics.

Besides "sheer knowledge", *understanding* should be separated from *adaptation*. For instance, an experienced rodent (or a human for this matter) competently navigates in its social environment. But only metaphorically, one may say that the rodent (or human) *"understands"* this environment.

With the above in mind, we indicate the following two mutually linked attributes of what we accept as "*understanding*".

[1] *Structural compression of "information".*

[2] *Power of prediction.*

These **[1]** and **[2]** can be quantified in a variety of ways. For example, one may speak of the degree of compression versus the "percentage" of structure lost in the course of compression, while the essential characteristics of a prediction is *specificity* versus *frequency of success* .

Later on, we shall use this kind of quantification for *partially ordering "levels of understandings"* and shall suggest tests for evaluating progress achieved by a learning program PRO in these terms.

[18] One may also equip PRO with an ability, similar to that possessed by children up to the age 2-3, to resist a "bad teacher" by rejecting environmental signals that are detrimental for the learning. (This ability deteriorates with age as one has to adapt to the environment in order to survive.)

[19] Ancients astronomers came to *understand* periodicity of planetary motions and were able to make rather accurate predictions.

[20] One can not much compress the "useful information" without loosing this "information" but if we can "decode" the structure of redundancy it can be encoded more efficiently.

Another attribute of *"understanding"* that is easy to test but hard to quantify is as follows.

[3] *Ability to acquire knowledge.*

For instance, a program *PRO* minimally proficient in English, would "know", upon browsing through Encyclopedia Britannica, that cows eat grass and cats eat mice.[21]

Also, the following can be seen as a hallmark of understanding.

[4] *Ability to ask questions.*

(Those whose business is UNDERSTANDING – scientists and young children – excel in asking questions.)

Besides the ability to understand, learning programs *PRO* may be graded according to their "internal characteristics", such as the *volume of the memory* a *PRO* has to use, the *number of elementary operations* and *the time* needed for it to make, for instance, a particular prediction.

At some point letter on , we shall comprise a list of specific criteria a learning/understanding program must satisfy.

2 Linguistic Flows and their Structures.

2.1 Features of Linguistic Signals.

The essential attributes of "verbal signals" be they transmitted and/or perceived auditory, visually (sign languages) or via tactile channels (in deaf-blind communication) are as follows.

(1) *Fast Language Specific Clustering.* Formally/physically *different* signals, e.g. sounds, are perceived/recognised as *identical* verbal units, e.g. phonemes, words, phrases, where this is achieved within *half-second* time intervals.

The clusterization of *phonemes* (and, probably, of other, including non-auditory, basic verbal units) depends on a particular language and the mechanism of learning these clusters by children (that deteriorates with age) is poorly (if at all) understood.

Yet, abstractly speaking, this is the easiest of our problems as it is witnessed by the efficiency of (non-contextual?) speech recognition algorithms.

(2) *Formalized Division into Units.* Flows of speech are *systematically* divided (albeit non-perfectly) into *(semi) autonomous units*, where the basic ones are what we call *"words"*.

This division, that is sharper than that of signals coming from "natural sources", is based in a significant extent on *universal* principles of *segmentation* that are applicable to all kind of signals where the markers separating "segments" are associated with *pronounced minima* of the *stochastic prediction profiles* (see section ???) of signal flows, where determination of such a profile depends on structural patterns characteristic for a particular flow.

(3) *Medium and Long Range Structure Correlations.* There are more "levels of structure" in languages than in other flows of signals. This is seen, in part,

[21] Properly responding to *"Do black cats eat fresh mice?"* instead of plain *"Do cats eat mice?"* would need a study of a more representative corpus of English than Encyclopedia Britannica by *PRO*.

in a presence of non-local "correlations" between different fragments in texts.

For instance, if a sentence starts with "*There are more ... in ... "* one may rightly expect "*than in "* coming next with abnormally high probability.[22]

And if "*Jack"* appears on every second page in a book and "*his eyes sparkled again*" than, you bet, *"Jack's eyes sparkled*" on the previous page.

(4) *Verbal Reduction of non-Linguistic Signals.* Many different non-verbal signals, corresponding to objects, events, features or actions may be encoded by the same word.[23] For instance, hundreds small furry felines that have ever crossed your field of vision reduce to a single "cat".

Non-verbal signals are many while their word-names are few. The use of a language replaces the bulk of the "raw memory" in the brain by a network of *"understand*" links between individual items in this memory. This is why small children visibly enjoy the process of the verbal classification/unification of "natural signals" from the "external world" as they learn to identically name different objects.[24]

(5) *Imitation, Repetition and Generation of Lingustic Signals.* Humans, especially children, have the ability to reproduce linguistic signals σ they receive, including those emitted by themselves, where, to be exact, *not signals σ themselves* are generated but members σ' of the *same class/cluster* as σ and where the choice of a particular *classification rule* is a most essential matter. (We shall return to this later on.)

One can hardly analyse languages without being able to generate them[25], where the language generative mechanisms – called *generative grammars* – result from the repetitive nature of imprecise imitation.

(6) *Many Levels of Self-Referentiality.* No other flow of signals, and/or human medium of communications have the propensity of self-reference that is characteristic of Language. The ergo-structures of languages contain multiple reflections of their own "selves", their internals "egos", such as

noun-pronouns pairs, allusions to previously said/written items, summaries of texts, titles of books, tables of content, etc.

Understanding a language is unthinkable without ability of generation and interpretation of self-referential patterns in this language.

(7) *Pervasive Usage of Metaphors.* Metaphors you find in dictionaries are kind of frozen reflections of their precursors in multiple coloured mirrors of Language (where such a precursor may not exit anymore). But many metaphors are ephemeral; they appear once and never come again.

2.2 Grammars and Dictionaries.

Making a dictionary involves several *interlinked* tasks where a starting point is

[22]Try: *there are more * in* on Google.

[23]This may be contrasted with the existence of *synonymous* words, but the multiplicities and significance of the latter are incomparable to the power of the verbal reduction.

[24]Children of this age are close to being ideal ergo-learners – the strive to lean and to understand is the main drive of ergo-systems.

[25] Neuronal signal generation mechanisms play an essential role also in vision: much of what you "really see" is conjured by your own brain, but the details of this process are inaccessible to us.

Annotation &Parsing, that is identification and classification of ***textual units*** that are persistent and/or significant fragments in short strings s (say, up to 50-100 letter-signs) as well as attaching ***tags*** or *names* to some of these fragments.

Tagging may be visualised as colouring certain fragments in texts, where these fragments and the corresponding colours may overlap. Or, one may represent an annotation by several texts written in parallel with the original one, where the number of different color-words is supposed to be small, a few hundred (thousand?) at most, with a primitive "grammar" that is a combinatorial structure organising them.[26]

One may think of such annotations as being written in strings positioned on several *levels*[27] over the original strings s, where the new tug-strings on the level l are written in the tug-words specific to this level and where the number of such l-tugs (at least) exponentially fast decays with l.

An ergo-dictionary is obtained by several consecutive *reductions* or *factorizations* applied to a library of annotated texts where the resulting combinatorial structure of the dictionary is quite different from that of (annontated or not) texts.

We describe below the basic combinatorial/categorical structure of texts and libraries and briefly indicate how parsing is performed.

2.3 Categories and Diagrams of Presyntactic and Syntactic Insertions .

Deep linguistic structures display some *approximate category theoretic* features, e.g. *abridgements* may be seen as *semantic epimorphisms*, or as *functors* of a kind rather than mere "morphisms".

Then translations from one language to another come as functors between categories (*2-categories* if abridgements are regarded as functors) of languages, where the category theoretic formalism should be relaxed to accommodate imprecision and ambiguity of linguistic transformations.

But we shall be concerned at this point with the following more apparent combinatorial category-like structure that is universally seen in all kind of "flows of signals".

Let a *library* L in, say English, language $\mathcal{L}$ be represented by a collection of tapes with strings s of symbols, e.g. letters or words, written on them, where many different tapes may carry "identical" or better to say *isomorphic* strings with the notation $s_1 \simeq s_2$, with the equality notation $s_1 = s_2$ reserved to same strings in the same location on the same tape.

Let arrows $s_1 \hookrightarrow s_2$ correspond to *presyntactic insertions* between strings, i.e. where such an arrow associates a substring $s_1' \subset s_2$ to s_1, where $s_1' \simeq s_1$.

We assume our strings are relatively short, no more than 10-20 words of length: this is sufficient for describing any "library" since every 10 words long

[26] An annotation may include references to *non-linguistic* signals but this would contribute to one's *knowledge* rather than to one's understanding.

[27] These levels l can be regarded as numbers $0, 1, 2, 3, ...$ with $l = 0$ corresponding to strings in the original text, where the number of level is small, something between 3 and 5. But, as we shall see letter on, these levels are organised according to a structure that is not quite linear order.

string *uniquely* (with negligibly rare exceptions) extends (if at all) to longer strings, since the total number of strings in any language is well below $100^{10} << n^{10}$ for n being the number of symbol-words in a language.[28] As for L one might thing of something with the number N of words in it in the range 10^6-10^{12}.

The resulting ***category*** $\mathcal{C}_{\hookrightarrow} = \mathcal{C}_{\hookrightarrow}(L)$ *carries the full information about* L.

DISCUSSION.

[+] Invariance. $\mathcal{C}_{\hookrightarrow}$ is invariant under the changes of "alphabets" – names of the symbols.

[++] Universality and Robustness The categorical description of languages satisfies the most essential ergo-requirement that is UNIVERSALITY.

For instance, *spoken languages* can be similarly described in categorical terms, whire, unlike written languages the arrows must correspond to *approximate* insertion relations between auditory or visual patterns.

In fact, allowing *approximate* presyntactic insertions with *sequence alignments* (with a margin of error 5-10%) in place of syntactic isomorphisms between strings would enhance the *robustness* of categorical descriptions of *written* languages as well.

[*] Non-locality. The $\mathcal{C}_{\hookrightarrow}$-description of libraries depends on comparison between strings that may be positioned mutually far away from each other in texts.

[**] Long Term Memory. This comparison between strings, depends on the presence of a structurally organised, albeit in a simple way, *memory* within the learning program.[29]

REDUNDANCY AND EXCESSIVE LOCAL COMPLEXITY OF $\mathcal{C}_{\hookrightarrow}$.

[–] The full category $\mathcal{C}_{\hookrightarrow}(L)$ contains many "insignificant" arrows, e.g. insertions of *single letters* into ten word sentences and arrows between "non-linguistic" strings, such as "tic stri".

This can be corrected by

allowing only **TEXTUAL UNITS** *for objects in* $\mathcal{C}_{\hookrightarrow}$.

and by

selecting a ***representative subdiagram*** $\mathcal{D}_{\hookrightarrow} \subset \mathcal{C}_{\hookrightarrow}$.

Such a diagram $\mathcal{D}_{\hookrightarrow}$ (that is a *network* of directed *arrow-edges* between *strings for vertices*) *must generate* (most of?) $\mathcal{C}_{\hookrightarrow}$ as a monoid, and also it must be "small", e.g. being a *minimal* subdiagram generating $\mathcal{C}_{\hookrightarrow}$.

(There is no apparent natural or canonical choice of $\mathcal{D}_{\hookrightarrow} \subset \mathcal{C}_{\hookrightarrow}$, but it may depend on the order in which the learner encounters texts in the library.)

[–+] Pruning and Structuralizing $\mathcal{D}_{\hookrightarrow}$. No matter how you choose $\mathcal{D}_{\hookrightarrow}$ it has *too many arrows* issuing from certain (relatively short) strings s, where the number of such arrows grows with the size of a library. Thus, in order to comply with the principles of ergo-logic, our learning algorithms must automatically reorganise $\mathcal{D}_{\hookrightarrow}$ in order to correct for this excessive branching. This may be

[28] Never mind the saying: "*there are infinitely many possible sentences in a natural language*".

[29] Conceivably, this organisation corresponds to how languages are perceived by their principal learners – 1- 4 year old children, where the $\mathcal{C}_{\hookrightarrow}$-categorical organisation of memory is the "ground level" of what we call "understanding" of $\mathcal{L}$.

achieved, as we shall see later on, by operations of *reduction*[30] applied to (the sets of) strings and arrows.

Categories $\mathcal{C}^{\hookrightarrow\downarrow}$ and Diagrams $\mathcal{D}^{\hookrightarrow\downarrow}$ of Annotated Texts.

If the texts in a library are annotated with tug-strings s' that are written on several level over original strings s, then the category with "horizontal" arrows $s'_1 \to s'_2$ is augmented by the "vertical" *position arrows* $s'' \downarrow s'$ saying that s'' lies over s', where such "mixed categories" and their representative subdiagrams are denoted by $\mathcal{C}^{\hookrightarrow\downarrow}$ and $\mathcal{D}^{\hookrightarrow\downarrow}$.

The presence of vertical arrows serves two purposes.

[1] Vertical arrows significantly *increase the connectivity* of diagrams since *a bound on the number* of tug-words on the *high levels* of annotations yields the existence of *many* horizontal (syntactic insertion) arrows between strings on these levels that were not present on the lower levels.

[2] And

the notion of a representative diagram $\mathcal{D}^{\hookrightarrow\downarrow}$
is modified in the presence of vertical arrows

by replacing many horizontal arrows issuing from lower levels strings in an annotated text by the corresponding arrows on the higher levels where the "low level information" is encoded by (inverted) vertical arrows. Thus one (partly) compensates for the excessive branching of $\mathcal{D}_{\rightarrow}$.

Remarks. (a) *Reconstructing* Arrow of Time. Linguistic strings are directed by *the Arrow of Time*. The category $\mathcal{C}_{\rightarrow}$ is unaware of this arrow but, probably, the time direction in strings can be reconstructed by some rule *universally* applicable to *all* human languages. Possibly, a predominantly *backward orientations* of self-references in texts may serve for such a rule.

(b) *Structures in Symbols.* In our categorical description (the alphabet of) the basic symbols, say letters, carry no internal structure of their own. But in reality letters in alphabets are non-trivially structured in agreement with one of the ergo-logic principles that allows no unstructured set of objects with more than three-four members in it. I am not certain what one should do about it.

(c) *Dimension of Vision.* Visual signals[31] For instance the visual are customary recorded on 2-*dimensional* backgrounds such as on photographs and/or eye retina, where the extra dimensions of depth and of time (in moving pictures) carry only auxiliary information. The morphisms $s_1 \hookrightarrow s_2$ here correspond to similarities between visual patterns s_1 and subpatterns in s_2.

But, probably, a significant part of visual perception is 1-*dimensional* being implemented/encoded by the neurobiology of *saccadic eye movements*. This suggests unified algorithms for learning to see and for learning to speak.

---

[30] This is also may be called *clusterization, classification,categorisation, factorization.*

[31] There is a demarcation line separating visual structures of *Life* – plants, animals, humans, human artefacts, from those of *non-Life* – stretches of water, rocks, mountains. These two classes of images are, possibly, treated differently by the visual system.

2.4 Fragmentation, Segmentation and Formation of Units.

Certain fragments of incoming signals[32] e.g. particular strings of letters such as *"words"*,[33] or some distinguished regions in visual fields, such as "*perceived objects* or "things"[34] qualify as *textual units*.

One can hardly give a comprehensive definition of such a unit, or a *signal-unit* in general, but one may indicate the following essential feature common to most units.

Probability of encountering a unit u among a multitude of other signals in the same category as u (here "category" means *class*) is *significantly greater than the product of probabilities of "disjoint parts of u"*.

For instance the word "probability" that has 11 letters in it may, a priori, appear only once or twice in a library with billion books ($<< 26^{11}$ letters) in it,[35]; but in reality it appears millon-fold more often than that.

This does not work quite so nicely for short words: scrabble dictionaries offer ≈1000 three-letter English words and ≈4000 four-letter words where many of them, e.g. *qat* (an African plant) or (to) *scry* (to practice crystal gazing) come rarely, but the improbable frequency of such a word may be seen in appearance of several copies of it in a single volume, or even on the same page.

The abnormal frequency alone, however, does not define units: the string "obabili" appears at least as often as the full "probability"; thus, one has to augment the "definition" of a unit by the following

completeness/maximality condition: If a string s is a unit, then larger strings $s' \supsetneqq s$ are *significantly* less probable than s.

Segments and Boundaries. Fragmenting texts into units is naturally *coupled* with the process of *segmentation* that is introduction of *division points* that make boundaries of string-units in texts.[36]

Determination when the position d in a string S between two letters may be taken for a division point depends on the strings s "to the left" and "to the right" from d in S, where such a string, say s_{left}, being a unit is an essential indication for d being a division point.

But it may also happen that there is no such clear cut units next to d in S but there is a 20 letter string S' somewhere else in the library that contains

[32] We temporarily ignore overlaps between fragments such as *"hard to see"* and *"to see it"* in the unit-phrase *"hard to see it"*. (*"Hard to"* makes a perfect "unitary uttering"; yet this is a weaker unit than *"hard to see"*.)

[33] A textual unit may be "disconnected", e.g. it may consist of two (more?) strings separated by other strings in a text. This happens, for instance, to separable prefixes in German that are moved to the end of the sentences. Also this **is** *not exceptionally rarely* **seen** in English.

[34] The rigid concept of *object-unit* will be later combined with classification/reduction and applied to "things" that come in many shapes such as words with flexible morphological forms, the human body, or to something inherently random such as an image of a tree with multiple small branches. When out eye looks at such a tree, our mind, conjecturally, sees (something like) a *branch/shape distribution law* rather than the sample of such a distribution implemented by an individual tree.

[35] The number of different books in the world is estimated at about 100 million. This seems to imply 5-50 million authors, half of whom must be among 7 billion people who live today on Earth – lots of writers around us!

[36] Boundaries of the so called "words" are marked in most written languages by white spaces while phrases and sentences are pinched between division punctuation signs. But we pretend being oblivious to this for the moment.

isomorphic copies of five letter strings to the left and to the right from d and such that the corresponding d' is recognisable as a division point in S'. Then we may accept d as a division point in S and to use this for identifying previously unseen units in S.

The coupled Fragmentation + Segmentation is a multistage process each step of which is a part of the learning transformation PRO on a certain space of pairs $(Frag, Seg)$ that will be incorporated into the full "understanding space" later on.

This process must comply with *"please, no numbers" principle*: the program PRO we want to implement must function similarly to an infant's brain that, unlike an extraterrestrial scientist, has very limited ability of counting and of manipulating large numbers (e.g. frequencies) as well as small ones such (e.g. probabilities).

This is achieved, as we shall explain later on, by consecutive "internal fragmentation" of the process PRO itself into a network of simple processors/directories where, they all, individually, perform (almost identical) "baby operations" with the global result emerging via communication between these processors.

Syntatactic from Presyntactic. Eventually, we isolate strings (sometimes pairs of strings) that are serve as *textual units* and also we identify *significant insertions* between them that we call *syntactic insertions.*

Linguistic 2-Spaces $\mathcal{P}_{\hookrightarrow} = \mathcal{P}_{\hookrightarrow}(L)$ and $\mathcal{P}^{\hookrightarrow\downarrow}$.

Let us represent strings from a given library L by line segments of lengths equal the numbers of letters in them. Attach rectangular 2-cells to the disjoint union of all these strings, where these "rectangles" are Cartesian products $s \times [0,1]$, with s being some strings/segments of length ≥ 5 letters each and where the attachment maps are syntactic insertions from the segments $s \times 0$ and $s \times 1$ to some string segments S_0 and S_1 such that the images are *maximal mutually isomorphic* (i.e. composed of the same letters) substrings in S_0 and S_1.[37]

In fact, it is more instructive to use the maps corresponding *not to all* syntactic insertions in the category $\mathcal{C}_{\hookrightarrow} = \mathcal{C}_{\hookrightarrow}(L)$ but only to those from a *minimal* diagram $\mathcal{D}_{\hookrightarrow} \subset \mathcal{C}_{\hookrightarrow}$, that *generates* all morphisms from $\mathcal{C}_{\hookrightarrow}$ on strings of length≥ 5.

Then the resulting 2-dimensional cubical (rectangular) polyhedron $\mathcal{P}_{\hookrightarrow} = \mathcal{P}_{\hookrightarrow}(L)$ adequately encodes the library L and if L is sufficiently large, this $\mathcal{P}_{\hookrightarrow}$ carries all structure knowledge of the corresponding language $\mathcal{L}$ with segmentation into basic units – words and short phrases made visible.

If one deals with the category $\mathcal{C}^{\hookrightarrow\downarrow}$ corresponding to an *annotated* library, or with a subdiagram $\mathcal{D}^{\hookrightarrow\downarrow} \subset \mathcal{C}^{\hookrightarrow\downarrow}$, then one attaches "vertical" rectangles along with "horizontal" ones, where the horizontal rectanglers are associated to the arrows $s'_1 \hookrightarrow s'_2$ and the vertical ones to the arrows $s'' \downarrow s'$.

Branching Entropy

Extensions of a *string-unit*, s, e.g. of a word, by short units t following next

[37] Our ad hoc bound *length* ≥ 5 serves to eliminate/minimise the role of "meaninglessly isomorphic" substrings, (e.g. of individual letters) where the same purpose may be implemented by a natural constrain on strings and gluing maps as we shall see later on.

after s in a library L define a *probability measure* on these t for

$$p_{\vec{s}}(t) = p_{L,\vec{s}}(t) = N_L(st)/N_L(s),$$

where $N_L(s)$ and $N_L(st)$ denote the numbers of occurrences of the strings s and of st respectively in L.

The collection of numbers $\{p_s(t)\}$ indexed by t serves as an indicator of variability of usage of s in the texts in the library L, where it seem reasonable to use not all t but only a collection T of unit-strings (words) t corresponding to roughly 10 (it may be something between 3 and 50, I guess, that need be determined experimentally) *largest numbers* among $p_{\vec{s}}(t)$.

The standard invariant of the probability space $\{p_{\vec{s}}(t)\}$ that reflects variability of p and regarded as an invariant of s is the *(one step forward) entropy*

$$\vec{e}nt(s;L) = -\sum_{t\in T} p_{\vec{s}}(t)\log p_{\vec{s}}(t).$$

Similarly, one defines $\overleftarrow{e}nt(s;L)$ via left extensions ts of s as well the corresponding invariants reflecting relative frequencies of "double extensions" of s that are t_1t_2s, t_1st_2 and st_1t_2.

Probably, such entropies (this is definitely true for more elaborate invariants of this kind) will be quite different for the strings *"birds-fly"* and *"pigs-fly"*[38] while *"bats-fly"* will be close to *"birds-fly"* in this respect.

Classification of Words and Partitions into Sentences. Segmentation of texts into strings with more than 2-3 words in them is impossible without preliminary syntactic classification of *basic units* – words and short phrases. But when such classification is performed and the number n of basic units u – this n is about 10^5-10^6 in English – is reduced to much smaller number $\underline{n}$ of classes $\underline{u}$, realistically with $10 \le \underline{n} \le 30$. Then a library with N basic units in it would allow one to reconstruct the rule of formation of strings of length about $\log_{\underline{n}} N$. For instance, if we classify with $\underline{n} = 20$, then a modest library with 10^9-10^{10} basic units in it[39] gives an access to 6-8 basic unit long strings, for $\log_{20} 1.3\cdot 10^9 \approx 7$ that may allow an automatic discrimination between *admissible* and *nonsensical* strings up to, maybe, 12 words in length. Then *generation* of *meaningful strings* becomes a purely mathematical problem.

Gross Contextual Segmentation. In the spoken language, utterings are divided according to when, where and who is speaking to whom, while texts in written languages are organised into paragraphs, pages, books, topics, libraries with a similar arrangement of pages on the web.

These partition structure are essential for making a statistical analysis of languages; conversely, texts can be classified/partitioned according to relative frequencies of short range structural patterns. e.g. basic units, present in them. vspace 2mm

Non-Textual Syntatctic[40] *Units.*[41] Languages, unlike non-linguistic arrays

[38] The two strings have comparable frequencies on Goggle

[39] There are about 100 basic units on a page, 10^4-10^5 of such units make a book, a 10 000 book library comprises $\approx 3\cdot 10^9$ units, while the world wide web may contains up to 10^{12} basic units of the English language.

[40] The word "syntactic" is understood in the present article as "characteristic of languages".

[41] *"Unit"* can be "defined" as *"everything worth giving it a name"*; it remans, in order to implement this "definition", to design a program that would understand what "worth" is.

of signals, "contains" units that are *not fragments of texts*. For instance, the groups of words, such as

{yes, no, maybe }, {we, us, our}, {big, large, huge}, {smelly, tasty, crunchy}.

are kind of "outlines" of such units. We shall explain later on automatic processes for locating of these and other "higher order" units, in texts and properly incorporating them into dictionaries.

2.5 Similarity, Coclustering, Classification and Coordinatization.

Organizing multi-branched (hierarchical) classifications of units is essential for developing "understanding dictionaries". where the resulting classes become "higher order" units.

Even the basic units – *the words* come up as *equivalence classes of strings containing these* words, rather than as the mere "spell-strings". For instance, the two collections of strings

[**bats-eat**]: *bat-with-flapping-..., bats-from-..., bats-are-present-..., vampire-bat, bats-catch-..., innoculation-of-bats, bat-captured...*

[**bat-hits**]: *training-bat, used-bats-on-sale, made their own bats, increase-your-bat-..., throws-his-bat, ...-bats-per-game, raised-a-bat...*

represent two different "bat" class-words.[42]

Our goal is formulating a *universal* classification rule(s) *a priori, applicable to all kinds of strings* (and, desirably to differently structured signals) that would essentially agree with the above division of "*bats*" into two classes.[43]

Classifications are often (but not always) achieved by means of *similarity* and/or *equivalence* relations R that, besides *similarity* and *equivalence*, reflect the ideas of

"sameness", "identity", "equality", isomorphism", "analogy", "closeness", "resemblance",

where such relations R are regarded as *higher order units* and are themselves subjects to further classification.

And not all similarities lead to what we call "classification", essentially, because the "equivalence axiom" $A \sim A$ is unacceptable in ergo-logic. (If the space in your head is filled with such staff you are brain-dead.)

In fact, similarity concepts S that are applicable only to small groups of objects, such as what brings together

{sweet, bitter, salty, sower, tangy},

and that do not *meaningfully*[44] extend to majority of words, *unite* their respective S-similar members rather than *divide* non-similar ones into classes.

[42] Non-existence of the string "*bats eat and hit*" shows how far apart the two classes are but ambiguous strings such as *"hit by a flying bat"* effectuate "linguistic bridges" between the two classes.

[43] There are more – about a dozen – different class-words spelled "bat", that are, essentially, subclasses of [**bat-hits**].

[44] The common idea of "meaning" is inapplicable *within* ergo logic but *ergo-meaningless* formalism, such as $A = A$, is non-acceptable either.

Another kind of groups of words having much in common that may or may not be regarded as true classes are those of *morphological word forms* such as

{works, worked, working}

or

{white, whiteness, whiten.}

On the other hand, traditional parts of speech: *verb, noun, adjective,...*, etc. represent typical *classes of words*; also division of words into "common" and "rare" is essential despite being ambiguous.

Certain type of classes (*categories*, as they are called by linguists) are often hard to extract from the depth of a language, but interesting and often unexpected results may be achieved by applying *universal classification schemes* to languages.

Basic Example: Coclusterization principle.[45] Two units u_1 and u_2 are regarded as *similar* and/or brought to the same class/cluster if they *similarly interact* with units v_1 and v_2 in-so-far as v_1 is *similar* to v_2.

To see how this seemingly circular "definition" works let, for instance, u and v be words that are regarded as *"interacting"* if v *often*[46] *goes next after* u. If we have about 100 000 words to work with and "often" means *"at least ten times"*, then such an "interaction" is described by a $U \times V$ matrix R with $\{yes, no\}$ entries of size $10^5 \times 10^5$, and a reliable evaluation of these entries needs a library of about $10^{11} = 10 \cdot 10^{10}$ words in it.[47]

But it may happen, and it does often (albeit approximately) happen in "real life", that this huge matrix is (approximately) determined by something mach smaller, say by a 300×300 matrix, where you need only 90 000$< 10^5$ entries to fill in and for which a a 10^6-10^7-word checking would be sufficient.

Namely, think of R as a $\{yes, no\}$ valued function in two variables, $R = R(u, v)$, and *conjecture* that

• there are *reduction maps* $\underline{f} : U \to \underline{U}$ and $\underline{g} : V \to \underline{V}$ where the cardinalities of the sets $\underline{U}$ and $\underline{V}$ are ≤ 300;

• there exists a $\{yes, no\}$ valued *reduced relation (function)* $\underline{R} = \underline{R}(\underline{u}, \underline{v})$, such that

$$R(u, v) = \underline{R}(\underline{u}, \underline{v}) \text{ for } \underline{u} = \underline{f}(u) \text{ and } \underline{v} = \underline{g}(v).$$

Observe, that the existence of $\underline{R}$, $\underline{f}$ and $\underline{g}$, a priori, is *extremely unlikely* even if $R(u, v)$ is only approximately equal $\underline{R}(\underline{u}, \underline{v})$ and if such (even approximate) $\underline{R}$, $\underline{f}$ and $\underline{g}$ are found they would be *unique with a overwhelming probability.*

Also notice that description of R by means of $\underline{R}$ needs only

$$90\ 000 + (2 \log_2 300) \cdot 10^5 < 2 \cdot 10^6 \text{ bits}$$

instead of the original 10^{10} bits.

The above may be generalised and developed in a variety of directions, made more precise, more detailed, more specific (we shall do this later on). But the following question remains:

[45] It is hard to trace the origin of this idea without knowing how it was christened at birth.

[46] This is, purposefully, formulated ambiguisly. But within the range of possible "often" there is a "coclusterization stability region" that allows a resolution of this ambiguity.

[47] If you check one word per second - eight hours a day - five days a week, it will take more than 10 000 years to go through such a library.

What can one do, say, with functions $R(u_1, u_2, u_3)$, where each u-variables may take $\approx 10^6$ values (that makes 10^{18} u-triples) and where no kind of $\underline{R}$-reduction is available?

Our pessimistic answer is "*In general, nothing*": the human mind will be unable to understand the structure of such an R, unless... a miracle happens: somebody discovers a hidden regularity in such an R something like what we call "a law" in physics.

Biclustering for Words × Contexts. We reserve the word "biclustering" to the case of functions R in *two* variables as in the above $U \times V$, where we want to look now at another kind of example where $u \in U$ are *words* while $v \in V$ are *books* and the function R encodes presence/absence of a u in v. Here biclusterization serves to classify books by topics according to their "key words" while the words themselves become classified by configurations of subjects, such as: chemistry of plants, animal foods, etc.

Such a clusterization may be also applied, besides $R(u, v)$, to the entropy function $\vec{ent}(u; v)$ defined in the previous section and it may go along with that for pairs of words (u, u') according to their *systematic appearance in the same book* and/or with *tri-clusterizatuion* for $R(u, u', v)$ encoding the *insertions of words u and u' in the book v*. The resulting three classifications may be different and they must be incorporated into different facets of the "I understand" program/structure.

Iterative Clustering. Coclusterization may be computationally hard in general, but the following simple algorithm for biclustering words illustrates a possible resolution of this problem.

Evaluate the above $R(u, v)$ for v taken from the set of 100 most common words in English. Observe that this will need a modest library only with 10^8 words ($\approx$ 1000 books) in it.

Represent the function $R(u, v)$ by a map, say R_* from words u to $\{yes, no\}$ functions $r(v)$ for

$$u \overset{R_*}{\mapsto} r(v) = R(u, v).$$

Now, let us endow our space of 2-valued functions on the 100-element set by some metric, where the simplest one is *the Hamming metric* and clusterize words u using the induced metric on them. (Assume, some unambiguous clusterization exists.) Take this for the preliminary classification/reduction of words u.

Suppose, you thus divided words into 3000 classes $\underline{u}$. Select 300 most common $\underline{v}$ among $\underline{u}$ and apply the above procedure to $\underline{u}$ (that would need less than 100 books), Then, after possibly doing it yet another time, you will arrive at the desired classification.

Trees and Coordinates. These are two most common classification structures in life , and, typically classifies contain both trees and coordinate in them that are presented as **1** and **2** below.

1: Classification as a Tree. This is may be seen as a sequence of partitions of units u into smaller and smaller classes, where the rule defining the $Part_i$-class of a unit u depends $Part_{i-1}$ of u.

A linguistically rather artificial instance of that, is classification/positioning of words in alphabetically organised dictionaries.

More significant example is where $Part_1$ divides words into the two classes:

A. Class of content words: {*nouns*, (most) *verbs*, *adjectives*, *adverbs*.}

B. Class of function words: {*articles, pronouns, prepositions, etc.*}

And $Part_2$-classes are obtained by further subdividing words into "parts of speech".

2: Classifications by Coordinates. These are given by several *coordinate functions* $c_i(u)$, where determination of $c_{i_0}(u)$ is essentially independent of $c_{i\neq i_0}(u)$ and where the set $I \ni i$ is not necessarily ordered. Different classes are formed by assigning particular values to some coordinates.

For instance, one may have the following functions $\mathbf{c}_1$,$\mathbf{c}_2$, $\mathbf{c}_3$,$\mathbf{c}_4$ on phrase-units u.

$\mathbf{c}_1(u)$ takes values *long*, *medium*, *short* depending on whether u has at most 4, between 5 and 8 or more than 8 word-units in it.

$\mathbf{c}_2(u)$ takes values *yes* or *no* depending if u contains a *content verb* in it.

$\mathbf{c}_3(u)$ assigns the *key word* w in u to u.

$\mathbf{c}_4(u)$ is the expected minimal age (in years) of a child able to understand the phrase u.

We expect our program *PRO* will be automatically generating this kind of functions $\mathbf{c}(u)$ in the course of learning a language.

3 Combinatorics of Libraries and Dictionaries.

We want to represent languages and dictionaries by *multicolored networks* U as follows.

• *Linguistic units* u are represented by *nodes* or *vertices* in such a network.

• *Connections between units* are depicted by *connective edges* between these nodes. We do not exclude a possibility of *several* (or none) connectives joining some nodes.

• Besides connectives, there are *directed edges* between certain nodes; these are depicted by arrows, such, for instance, as *syntactic insertions* between strings and *reductions* $u \mapsto v = class(u)$. A relevant feature of such arrows is that some of them are *composable* as in $u \mapsto v \mapsto w$.

• Nodes and edges carry certain *colours* where such a colour depicts "essential properties" of the node or edge it is assigned to. For instance a colour on a node u may say: "***short string***", a colour on a connective may say among other things, "***similarity***" and a colour on a reduction arrow may refer to a particular classifier algorithm defining/producing this arrow.[48]

From another perspective, colours appear as *descriptors* of *structural/logical units* involved in organisation/construction networks U. The ensembles D of such descriptors are smaller than U:

> a dictionary may have 10^5-10^8 nodes (libraries are even larger) while the number of colours is in U, that is the number of units in D, is somewhere between 50 and 500.

But multicoloured network-like structures carried by these D are more sophisticated than the structures in U. (We shall explain this later on.)

[48] One should think of a colour as a simple combinatorial entity, e.g. a little tree, rather than a word or phrase.

Collections, Ensembles, Sets.

We *do not* regard *collections/ensembles* of nodes, and even less so of edges, as *mathematical sets* for the following reasons.

1. The presence of a particular node in a network, e.g. of a particular phrase, in the long term memory of a learner is often ambiguous.

2. Basic set theoretic constructions, such as the union $X_1 \cup X_2$ and the Cartesian product $X_1 \times X_2$, can not (and should not) be unrestrictedly performed in our networks.

The set theoretic language may lead you astray;[49] yet, we use fragments of this language whenever necessary.

Colouring Descriptors. A conceptually most difficult problem in building network models of learning[50] is assigning colours to descriptors and connectives between them. We think of these colours as *formalised expressions* of (simple combinations of)

fundamental/universal principles of learning.

Identification and formalisation of these principles is our main task.

3.1 Library Colours.

What we call *a library L* is a a collection of *string-units*, where such a string may be, a priori, anything starting from a fragment of a word to a paragraph with a few dozen words in it.

These strings are "colored" according to their size, say into three basic[51] "colours": ***short, median, long***, where each of these colors may be "subdivided" into three "subcolors":

$\boldsymbol{short}_{short}$, $\boldsymbol{short}_{median}$, $\boldsymbol{short}_{long}$, $\boldsymbol{median}_{short}$, ..., $\boldsymbol{long}_{long}$[52].

(Later on, this structure will be *parsed* by identifying "significant strings" , such as ***word***s, ***phrase***s, etc. and disregarding insignificant ones, with possible subcolors such as $\boldsymbol{word}_{rare}$, for instance.[53])

There are several facets to a *learner's perception* of *geometric structure*(s) that underlines L and that is involved in formation of string-units and which is also needed for defining connectives between strings.

Simple Linear Order. A mathematician may be conditioned to think of a library as a string of symbols indexed by integers $i \in \mathbb{Z}$ with all of geometry of L derived from what he/she knows of $\mathbb{Z}$:

... ••

••• ...

But the "naive concepts" in the mind of our learner – think of a baby learning to talk – are more general, more powerful and more flexible, something like:

---

[49] Bringing forth *random sets* and/or *fuzzy set* may only aggravate the problem.

[50] Possibly, our kind of networks have little to do with "true learning", but one can not rule them out at the present stage.

[51] Digital number bases *two* or *four* are also possible(?) but *three* is preferable.

[52] These can be depicted as nine leaves of a rooted (and ordered) triadic tree with 12 edges and with a single colour ***length*** spread all over it.

[53] An essential (but not the only) intrinsic motivation for doing this by an ergo-learner is economising the memory space.

close-one-to-another, ***far-one-from-another***, ***next-to-each-other***, ***in-between***, ***begins-with***, etc., where these "colourful concepts" come in several subcolor-flavours similarly to (yet, differently from) how it is with lengths of strings.

Notice that all of these are binary relations except for the ternary ***in-between***.[54]

Besides these relations the large scale geometry in L is reflected in the presence of (relatively large) *contextual units* such as ***page***s and ***book***s, for instance.[55]

Then *closeness* between two strings can be seen as *simultaneous containment* of these *string-units* in the same *context-unit*.

This is essential for enlisting and keeping in memory (pre)syntactic insertions between strings, in particular all pairs of identical words w in L. The number of such pairs is uncomfortably large being *quadratic*[56] But identifying identical words say on a single page goes *linearly* in time.

Among colours carried by *insertions* of string-units into context-units we indicated here only the two: ***frequent*** and/or ***rare***. These are used, both for identification of (statistically homogeneous) context units as well as for classification units with a use of biclusterization.

It is not difficult to complete the above combinatorial description of our library L – one needs a few dozen more colours, about 50 of them all-to-gether, that represent *basic types* of units and of connectors between them in L.

But the principal issue is not so much L per se but a construction of the corresponding descriptor network D on the basis of few, probably 4-8, *"general rules"*, where, recall, the colours-descriptors of L are taken for nodes in D and *"general rules"* are used as colours assigned to nodes and connective edges in D.

[54] ***"begins-with"*** *formally* defines simple linear order that implies all other relations. But our baby-like learner is unaware of formal logic, for him/her these "colorful-relations", albeit interconnected, do not "logically imply" one another. If a human baby were born with a logical mode of thinking in his/her head, he/she would learn preciously little of essence in this world.

[55] Originally, these units are classified/coloured by their size, where different classes must roughly fit into the corresponding frames of the *short-term, medium-term* and *long-term memory*. Then the concept of "context" is modified and refined in the course of learning (not quite) similarly to how it happens to strings, where the true ***page***s and ***book***s must be either sufficiently *statistically homogeneous*, or *structurally unified* or to have *pronounced boundaries*.

[56] Squares are unacceptable except of small quantities. We happily live through *million* seconds that make less than 12 days of our lives. But *trillion seconds*, that is million squared, stretch over more than 31 000 years.