Pattern Theory
and "Poverty of Stimulus" argument in linguistics

Yuri Tarnopolsky

Abstract

This paper continues the examination of language as a quasi-molecular system from the point of view of a chemist who happens to ask, "What if the words were atoms?" Any new word in the vocabulary must be at some time heard or read in order to be acquired. The new word practically always comes linked either with an observed image or with a few other words in a phrase or discourse, otherwise it would be meaningless at the first encounter. This halo of selective connections makes the morphemes and words recognizable as generators of Pattern Theory (Ulf Grenander), i.e., as atomic objects possessing a certain structure of potential bonds with preferences for binary coupling. In this way the word is typically acquired with a fragment of grammar. Metaphorically speaking, the generators of language carry bits of grammar on their bonds like the bees carry pollen on their feet. Regarding Pattern Theory as meta-theory for atoms and words, parallels between linguistics and chemistry are discussed.

CONTENTS

Introduction	3
1. Sets and Order	5
2. Laws of grammar and laws of nature	7
3. From sets to generators	10
4. Patterns	15
5. Notes on notation	21
6. Bonding	26
7. Acquisition of generators	29
8. Bond space	30
9. Acquisition of bond values	33
10. Locality	36
11. Some examples	37
12. Language and homeostasis	41
Conclusion	44
References	46
APPENDIX: The Chemistry of *The Three Little Pigs*	49

First draft, 2004. In the process of editing. Remarks appreciated: yuri@ids.net

Introduction

The idea that children acquire their native language in spite of the lack of either direct instruction or sufficient number of correct or correcting samples goes back to Plato. Starting with this well seasoned "poverty of stimulus" premise, Noam Chomsky postulated the existence of an innate universal grammar (UG), and the entire theory became two postulates, one on the shoulders of the other. Further postulates about the nature of UG (for example, principles and parameters) had to be added to the increasingly unstable cheerleader pyramid, so that the issue became complicated and hotly debated. Any general course of linguistics, as well as the Web, reflects the war of the words over the tiny piece of intellectual land [1].

It seems strange that the problem of language acquisition exists at all. Language is a notation of thought. Why is then mastering notation is separated from acquiring knowledge, logic, and mastering communication with the world? A possible reason is that we hear what children say but do not see what is going on in their minds. Circumventing this very large and complicated issue, I attempt to look at the bottom postulate of the disputed paradigm: the poverty of stimulus. This unaffiliated paper continues the examination of language as a quasi-molecular system from the point of view of a chemist who, inspired by Mark C. Baker [2], happens to ask, "what if the words were atoms?" The paper serves is an addendum to [3], without which some loose ends will hang in the air.

Speaking and writing is the manifestation of life we all engage into with visible and audible output. Why should a chemist's opinion matter more than anybody else's?

It is not universally remembered that chemists long ago invented, with the purpose of communication, the language of chemical nomenclature, which converts a non-linear object into a string of words. The parallel between DNA and text was captured right at the birth of molecular biology. Chemistry and linguistics share much more conceptual genes [4].

Chemistry may be cool but linguistics is hot. I realize that there are very few original ideas in this paper, but to review even major domains of enormous linguistic literature is a hopeless task. Philosophic literature on the subject, from Plato to Wittgenstein, would alone sink my ship still in the harbor. Many references are omitted, especially, when ideas have been popularized and are widely spread. Leading modern linguists (some are mentioned in [3]), write about language with gripping virtuosity and passion (Examples: [9, 19]).

There are good reasons why the linguistic literature easily overwhelms a chemist, used to the enormity of the chemical literature, and, I suspect, even linguists themselves.

First, modern linguistics is far from the habitual for chemistry consensus. Second, linguistics, is still very far from reproducing the human ability of rational communication, contrasting with the triumph of applied chemistry which can identify ("see") and reproduce ("say") any substance from scratch. Third, many linguists examine language in terms of language, while chemists examine molecules in terms of their graphic images and measurable properties, using an absolute minimum of words.

While many linguists appeal to the jury for a verdict beyond reasonable doubt, the chemists require a hard proof beyond any doubt. Amazingly, both deal with real and observable objects, which alone should clear the way for linguistics into the family of natural sciences. Moreover, molecules are invisible without instrumentation, while words can be heard and seen even by small children.

> I have no intention to snicker at linguistics. On the contrary, I intend to lovingly poke fun at the heavy, Puritan , tedious, and down-to-earth realism of my native chemical thinking. The chemist works like the farmer and proves his point by bringing a molecule or process

> to reproducible material existence, as if it were a pumpkin. This is, probably, too much of virtue for the flowery, fluid, sybaritic, and Bohemian habits of language speakers and tillers. But what can you expect from those who deal with mindless and speechless molecules? Even cows moo. And yet I think chemistry is one of two most romantic sciences on Earth. The other one is linguistics.

Coming from a non-linguist, the opinion expressed here cannot be an argument in a professional discussion. It might, however, illuminate the problem from a new angle for both professionals and fans. Chemists think about the world in very distinctive terms. To demonstrate this manner is my major goal and part of a larger program, see [4, 5]. I am interested in the export of chemical experience to cognitive and social sciences with Pattern Theory as meta-theory for all discrete complex combinatorial systems.

> As for my own language, which is not my native, I will use, with the ecumenical blessing of George Lakoff [6, 7], metaphors with all the self-indulgence of somebody in no need of a grant.

> All unfamiliar to non-linguists terms could be easily found on the Web.

1. Sets and order

If I asked only the question "***what if the words were atoms***" and stopped there, the answer would not go beyond a metaphor. When Mark C. Baker entitles his enjoyable book "*The Atoms of Language*," he assumes that there is more to words than their use as building blocks of a combinatorial Lego. I am going to encourage the timid interplay of two distant but related disciplines by asking the inverted question: "***what if atoms were words?***" I believe that both questions are equally legitimate. They will guide us toward the realm of complex discrete combinatorial systems, which is still a little explored pre-Magellan world where administration and legislation overrides navigation.

Pattern Theory is the first system of mapping which shows Linguarctica and Chemistralia as recognizable continents made of the same firm land and surrounded by the same ocean. In Pattern Theory (PT), both words of language and atoms of chemistry are **generators** and their deep kinship is more than just a metaphor. The best way is to go to *Elements of Pattern Theory* by Ulf Grenander [8], which carries a great inventory of seeds and fertilizers for intellectual farmers. Next I will try to approach some of the basic ideas of Pattern Theory through the back door where ID is not asked and nobody cares whether you are chemist, linguist, or just a chatterbox.

The most fundamental and unifying concept of mathematics is **set**: any collection of any elements which can be combined and recombined into new sets. Elements of the set have nothing but pure identities. Thus, from elements **A**, **B**, and **C** we can form sets **{A}**, **{B}**, **{C}**, **{A, B}**, **{A,C}**, **{B, C}**, **{A, B, C}**, and empty set **{Ø}**.

The **mental** operation of combination is **unconstrained**, requires **no effort**, and the order and relation of elements in a set **do not matter**. There are no ties of any kind between the elements of a set except for being thrown together between the brackets. Set elements can be compared to pieces of paper with names, thrown in a bag and offered for drawing a lot.

> In casting a vote, when a number of voters place their ballots, some of them identical, we deal with the **multiset (**also called **bag** in computer science**)**, for example, **{A, A, B, C,C,C}**, where order does not matter either.
>
> Strictly speaking, chemistry deals with multisets, so that when molecule **A** turns into **B**, there is still plenty of **A** around. This is not so in computers and the mind, where everything is represented in single copy, so that any destruction is irreversible. The analogy with chemistry, however, becomes striking if we note the function of memory: whatever happens, memory keeps hard copies, at least for a while, and so creates an effect of multiplicity.

A list of names in **alphabetical order** is a quite different object. The elements of the list cannot line up freely. They must stick together in a certain way defined by their **local** properties, namely, their first and subsequent letters compared with an arbitrary **global** alphabet: **A**, **B**, **C …Z**. The alphabet is just a mapping—one symbol to one

number—of the set of positive integers on the set of symbols. Such sets are **ordered:** for each two elements, one precedes the other.

> There are also partially ordered sets (**posets**), like the somewhat flexible list of our daily priorities or the hesitant subjective rating of beauty contestants. In such sets we are certain about the order of some but not all pairs of elements.

2. Laws of grammar and laws of nature

Suppose we collect a large number (**corpus**) of actually spoken and not just possible utterances and have to decide which are **correct** (grammatical, *lawful*) and which are not. This task is easy if we have clear criteria of correctness, or at least a corpus of definitive correct utterances. But what if the language purists are split on the subject and, moreover, the speakers do not listen to them? Then a possible solution is to select a sub-corpus of statistically most frequent variations with the same meaning and regard them as the standard against which irregularities could be measured. The purists may not agree with some entries, but they will be certainly satisfied with most of them. Some people, however, may disagree about meaning. Besides, spoken language is largely automatic, improvised, and heavily dependent on context, intonation, and facial expression.

My point is that the notion of correctness is like the survival of the fittest in biology: it is circular and, therefore, just a mental toy.

> Note that the real man-made lists can have irregularities and the number of deviations from the alphabetical order per a unit of length could be an overall measure of the irregularity. For large collections of linear sequences (strings) of elements, whether speech, texts or DNA, a metrics, like Hamming distance, can be established: the sequences differing in one element are closer than the sequences with two discrepancies. This can be generalized for any complex objects. This is how a natural statistical norm can be captured. Statistics is not the only instrument of the linguist comparable with the

heavy duty instrumentation of the chemist. There are probes and tests very similar to chemical ones, like the wug-test.

My intent in probing the laws of grammar is to compare them with the laws of physics and chemistry. There is no correctness or incorrectness in nature, to which language belongs. This is why the very idea of a natural grammar as a system of rules and parameters, or just rewrite rules, stored in the mind of a little child and not in a book or an adult mind, seems to me, a chemist, as unnatural as any alphabet.

Let us turn from artificial (mind-driven) to natural (mindless) processes without external human control, which we understand much better than human matters.

The atoms stick together for physical reasons and form more or less stable aggregates called molecules where the number and order of connections of the atoms of different kind have a decisive bearing on the individuality and **behavior** of the molecule. If the atoms were indeed words, we could say that the atoms could form some stable aggregates and resist forming some others because **they knew the grammar** (i.e., rules and parameters) of chemistry.

Atoms do not consult the textbook of chemistry, however, before assembling into aspirin. The possibly enlightening for a linguist chemical story is that they "know" the rules in a noteworthy manner: **given indefinite time**, the molecules assemble in such a way, that the aggregates with the lowest energy are much more abundant than those a notch up on the scale. This is the only universal natural rule, but there is a multitude of not rules but **properties** of atoms, which define the actual form of the aggregates. There is a wide-spread among linguists belief that language has an infinite generating power (Chomsky: "discrete infinity"), but chemists are more cool-headed. Theoretically, all possible assemblies of a given set of atoms will be present after an indefinite time, but only a few will be **in fact detectable**, and even less will be prevalent. Should we say that the minor versions are wrong?

Chemists, like businessmen, are not interested in indefinite time and they focus on the problem of relative speed of alternative concurrent transformations. The concept of transition state, i.e., the bottleneck of transformation, is the major, but still neglected, contribution of chemistry to the theory of complex systems. More about that, see [5, 4, 3].

> This, however, has little to do with language acquisition because the time of individual learning is negligible with the time of language genesis and evolution.

Molecules **behave** because of the inherent molecular motion. Any living, not pinned to the display box, language is also full of motion: utterances fly around, clash, and scatter the sparks of fragments, not to mention processes on the historical scale. There are especially hot areas of professional, sub-cultural, and child language where new mutants are born to survive or die. Some parts of everything that has been **really** said and written during the day settle down in written or recorded form and can be used to study the language variations and evolution in the same way a paleontologist studies the evolution of birds. Alas, not much was left before writing, but tribal languages store a lot.

> Language, which comes in populations and perpetuates by replication, mutation, and exchange of material, is a form of generalized life. Contrived examples are chimeras. Remarkably, chemists sometimes start with chimeras in their imagination and then successfully synthesize them and even use for practical purposes, as nanotubes illustrate, but one cannot learn chemistry from nanotubes alone.

Molecular systems and language are similar not only because they consist of atoms, but because they are natural dynamical systems driven by the constraints of **thermodynamics**. This reeking of hot engine oil term means, in fact, something very general and simple: there is a preferred direction of natural events, and we know what it is, and if we go against it, we have to pay a price in the currency of energy. Natural language is not an exception and this is why it is **always correct**, until some shock hits the society of speakers and the language finds itself in an uncomfortable **unstable** position on the hill slope and slides toward a new position of reduced social stress, as the preferred direction of natural events requires. The preferred direction of language evolution, recapitulated in individual language acquisition, is the optimization of communication as part of social life. This is how a chemist could paraphrase the perfectly natural idea that language is an adaptation (Steven Pinker in [9] and earlier with Bloom), although evolutionary thermodynamics of the open systems, like life and society, is today

formulated only in very general terms. An eloquent, I would say, beautiful discussion of the subject in non-chemical terms can be found in [9], where even the distant behind-the-scene voice of chemistry can be heard (Komarova and Nowak, in [9])).

> My opinion of an outsider is that truly systemic linguistics should treat language as a natural system with a generalized physics, chemistry, and even physiology. A similar direction was outlined by Christian Matthiessen and M. A. K. Halliday: "It [language] is a phenomenon that can be studied, just like light, physical motion, the human body, and decision-making processes in bureaucracies; and just as in the case of these and other phenomena under study, we need theory in order to interpret it" [10]. In [3] I put Joseph Greenberg and Noam Chomsky in the opposite corners of the linguistics Hall of Fame. Today a see, alongside Greenberg, Brian MacWinney [11], who works with language as a typical natural scientist.

3. From sets to generators

There are more probable (and, therefore, more stable—thermodynamics again) and less probable (and, therefore, less stable) molecules, which are sets of atoms under various constraints. Some atoms in a molecule are bonded, others are not, and the bonds have different properties.

The mathematical image of molecular structure is a **graph** in which atoms are points selectively connected with lines. In a simple graph, the positions of the points (**nodes**) do not matter and all the connections (**arcs**) are the same. In the colored graphs the arcs, and in the labeled graphs the nodes, can be different, which makes them suitable to portray molecular structure. In a real molecule, atoms commonly oscillate around fixed positions in 3D space.

Only a few molecules are linear, but linear polymers, for example, DNA, consisting of molecular blocks arranged like the words in a text, are the most conspicuous evidence of the mathematical kinship between language and chemistry, well recognized by both clans.

Figure 1 shows a few objects that can be obtained from simple sets by adding binary relations between elements. Some elements are circles with different fill patterns, to emphasize their individuality. Sets with connected elements represent what chemists and architects understand by **structure**.

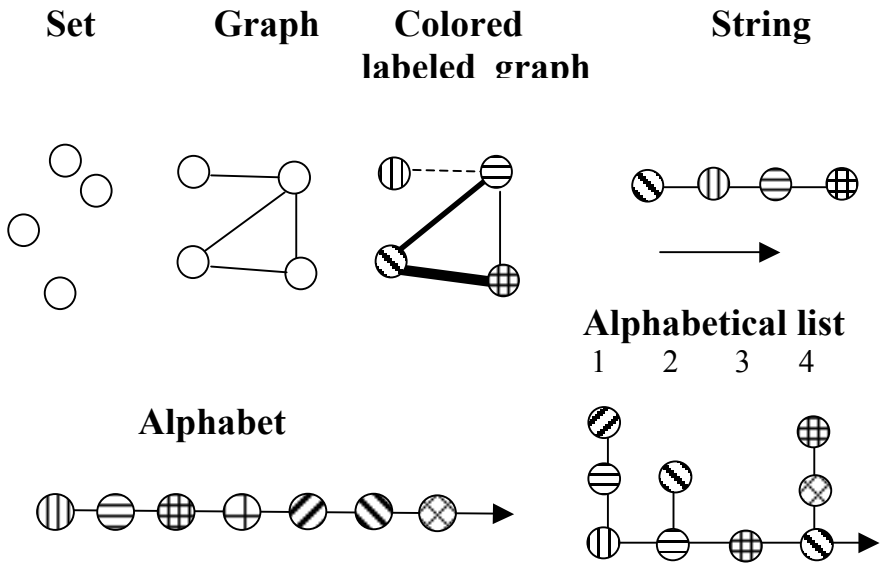

Figure 1. Evolution of sets

The concept of **mathematical structure** is something different. In a way, an ideal grammar is a mathematical structure: it separates right from wrong. Mathematical structure, simplistically, consists of terms, axioms, and operations, so that one can say which result of operations is right and which is wrong. **Algebraic structures** are a set of elements, a list of axioms, and one or more operations that turn two elements into a third. Obviously, this is what chemistry and linguistic do by connecting one block with another. This is why algebraic structures are in the foundation of mathematical linguistics [12]. The belief in the infinite language, probably, feeds on the computational idea of language as an obviously infinite set of strings generated by the operation of concatenation.

Generators and configurations also form an algebraic structure. From this point of view, algebraic structure and "structural" structure conceptually converge: it is all about binary operations. Chemical structures, however, have a lot of additional physical constraints that are beyond algebra.

Let us take an alphabetical arrangement: **LIST** = Andy—Bob—Pat—Ronny—Zelda. It is ordered according to the Latin alphabet:

A	**1**	**…**		**Q**	**17**	**….**	
B	**2**	**O**	**15**	**R**	**18**	**Z**	**26**
C	**3**	**P**	**16**	**S**	**19**		

LIST is a very simple artificial object, a result of my mental activity. ***If the words were atoms***, it would be reasonable to ask how **LIST** could originate from names without human participation.

We can describe the petite **LIST** in deceptively complex terms as a **configuration** of elemental objects: **generators** g from the generator space **G**. Generators are a kind of abstract atoms that can self-assemble, at least in or mind, into regular (*lawful*) configurations. Two examples of are shown in Figure 2.

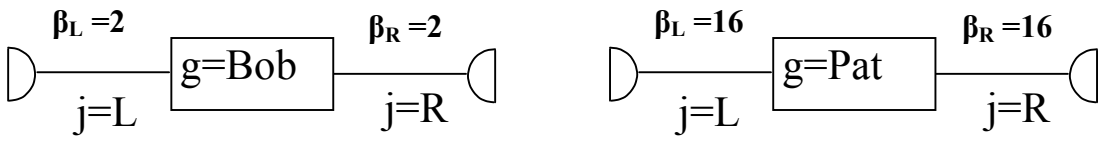

Figure 2. Generators of LIST

Each of the objects has an identifier g, i.e., label (name) and two bonds with coordinates j, labeled as L(eft) and R(ight). The bonds have numerical bond values **β** which are simply the positions of the first letter of the name in the alphabet. It would be reasonable, however inconvenient, to index **β** and j with the generator name:

$$\beta_L^{Bob}, \; j_R^{Pat}.$$

Let us apply the following rules to the **behavior** of these objects:

For g_1 and g_2, $\{g_1, g_2\} \subset G$ (i.e., for two generators from space G)

ρ= TRUE if $\beta_L^{g2} > \beta_R^{g1}$

Here we have the bond relation ρ that depends on the bond values β of two generators g. If ρ= TRUE, the two generators can be neighbors in the list. If FALSE, they are not fit to rub shoulders.

Note the **local** character of the rule. In order to check a compliance with the rule, only an examination within a tiny area of the string is necessary. The entire "behavior" of generators is local: the events consisting of acts of locking and unlocking (no, words *local* and *lock* have different origin) do not happen at a distance. Otherwise it would require a homunculus to control it. This property is important for any process of genesis of a complex system without a complex controlling mind. *Natura non facit saltum*, from this viewpoint, means that nature has neither mind nor algorithm, nor Random Access Memory to romp around all over. We can safely assume that humans started the language from scratch and were not taught by other beings how to deal, for example, with anaphora, i.e., the mental jump from the noun to its respective pronoun and back while we speak. It is only natural that humans are teaching the computers with RAM and ROM to do this trick and not to be lost in multiple nouns and pronouns. The simpleminded children acquire their simple pre-school language **without a tutor** because the acts of acquisition are local and, therefore, do not require a thinking mind. Later they start to chew on the bitter roots out of which the minds grow.

The word **behavior** adds a flavor of spontaneous activity to generators. Metaphorically speaking, we take a handful of generators g from the bag G and shake them in a box, so that they could stick to (or repel) each other according to their preferences, like the frequenters of a singles bar, only somewhat kinkier. Note, that

neither algorithm nor human control is needed for the spontaneous self-assembly of the above generators into a string—nothing but some random "shaking." We have to supply motion as the source of chaos, but a series of earthquakes over a million years would be as good for the purpose of shaking a box.

> The physical flavor that I attribute to generators is not to be found in Pattern Theory, although it can be easily inferred. I am taking some chemical liberties with mathematics. Nevertheless, in some sense, defined in PT, generators selectively **accept** (or **repel**, I would add) each other with varying enthusiasm. There is an **acceptor function** for the mutual affinity of two generators.

In a more general case, the bond value can be any number, regardless of any alphabet, and ρ could be a function of the two contacting β, so that for some pairs (bond couples) ρ is high (i.e., very true) and for others low or negligible (i.e., very false). Not only that, but probabilities or additive **weights** (not related to connectionist learning) can be attributed to the generators themselves, as in topicalization, i.e., putting the most important word first (fronting), and marking it either vocally or, in languages like Japanese, with a morpheme, in addition to fronting.

Finally, we would be able to calculate the probabilities of a string of symbols or a list of names by multiplying the probabilities of bond couples and the "weight" of names. We can also express the mutual affinity of generators and their weight in terms of generalized **energy**, which, unlike probability, is additive over the increments of bond couples. As result, we arrive to a kind of generalized chemistry of symbols, if not of everything. In PT it is called Pattern Synthesis.

Pattern Theory is about probabilities on structures. The system of this kind generates **thoughts** in Ulf Grenander's GOLEM [13], which, in my view, is an initial model of meta-chemistry where generators combine into configurations under predominantly local rules.

To make the next step toward PT , we need to further generalize generator by lifting the limit on the number of bonds and allowing non-linear configurations. They are

never present in speech, but nobody has ever managed to do syntax or semantics without either a tree or a Russian doll of brackets.

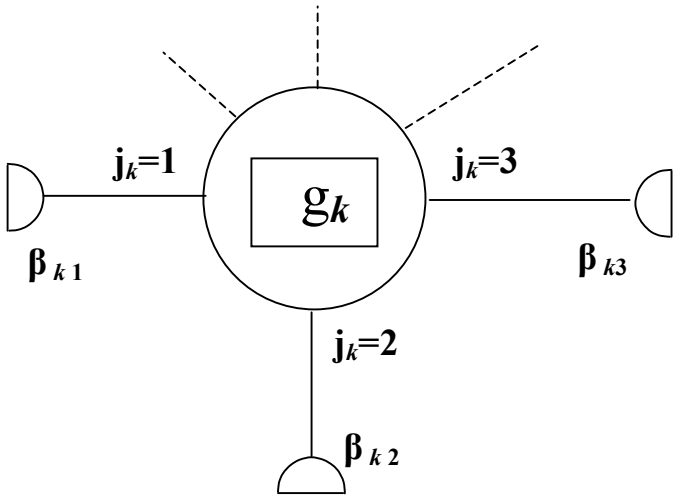

Figure 3. "Generic" generator

A generic portrait of an atom of everything is presented in Figure 3. It can have more than two bonds and can form a great variety of configurations, typically, non-linear. Abundant illustrations from large number of areas of knowledge can be found in [8] and more special and technical works in PT, where numerous shapes are analyzed and produced from a generator space and a minimum or no global constraints.

4. Patterns

The last key step from configurations to patterns is simple: pattern is a class of configurations. In PT it is a **similarity transformation** that generates one configuration from another within the same class. **Regularity** of a configuration, which plays the role of mathematical structure axioms, consists of generator space G, bond relations ρ,

similarity transformation S, and the type of connector Σ, for example, LINEAR or TREE . Regularity can be strict or relaxed.

Since major applications of Pattern Theory are designed for processing two-dimensional images, similarity transformations can often be expressed analytically in the form of equations, for example, for stretching, rotation, warping, etc. , or by non-trivial algorithms. In the discrete space of linguistics this is hardly possible. A similarity transformation can be simply a change of the word within a logical, morphological, syntactical, or, as in poetry, just phonetic category. The prototype of all similarity transformations in linguistics is the Laputian machine, witnessed by Gulliver, in which letters were permuted at random to produce a word of new knowledge.

Another way to describe the pattern is to formulate not what the similarity transformation changes but what it leaves unchanged.

Because of the historical origin of language patterns, it is not always possible to explicitly formulate the similarity transformation and it can be defined as just a list. Languages with genders (Russian, German) and noun categories (Bantu) are of this type. What seems logical aberration, like the neutral gender of **girl** in German (das Mädchen) must be just memorized by rot. There must be some evolutionary logic.

An interesting channel opens between PT and the domain of categorization in linguistics, with theory of prototypes (Eleanor Rosch [14]), inspired by Ludwig Wittgenstein's concept of family resemblances. The **template** of PT, which is a **typical** configuration, seems to be the exact counterpart of the prototype in linguistics. Pattern Theory was also inspired by Wittgenstein. The realistic aspirations of systemic and cognitive linguistics echo the realistic spirit of PT, still missing, as it seems, the heart of the matter with all natural systems: measure.

Since I am interested here only in what is necessary to test the Poverty of Stimulus Argument, I shall refer the reader to PT for more detail. I have only a few strictly personal remarks to be made on part of a chemist and not-mathematician.

A generator space $\mathbf{G}$, i.e., a set of generators with their properties, defines a combinatorial configuration space $\mathbf{C_G}$. Formally, it does the same job as the Cartesian coordinate system, which defines all possible points of Euclidean space, or the Chomskian principles and parameters, which, ideally, define all possible grammars, or the questions of the US Census, which define the essential profile of the US population. It does the job, however, without any explicit **global** list of coordinates for the multidimensional spaces, but by providing strictly **local** descriptions of generators and allowing some freedom of their mating.

We can use this curious system of coordinates, so practical for immaterial discrete systems in non-Euclidean spaces, if we tacitly count on the mystic ability of generators to find each other and stick together the right way or one among many right ways, some of them more right than others. This is a kind of mathematics that has some properties of physics and chemistry. This is how I became enthralled by PT. It is potentially a calculus of all **realistic** (however immaterial) **discrete** combinatorial objects because it attributes *a priori* some probability to their existence. It predicts what is **likely** to happen and explains why other alternatives are less likely.

In the case of alphabetic lists, the all-or-nothing system of rules defines all possible alphabetic lists and separates them from even larger space of all non-alphabetic ones.

In general case, Pattern Theory can partition the set of all configurations into regular (by the rules) and irregular ones, but even more generally, it offers a measure of regularity on a continuous scale.

The connection with linguistics can be seen here. Not only are grammatical structures regular and ungrammatical ones irregular, but some are more regular than others and others a less irregular. Moreover, there is a measure of stability, and some constructs are more stable—less ambiguous or difficult to understand—than others, provided the content of the mind is the same. Benjamin Worf's idea, as I understand (or misunderstand) it, was a relation between language and the content of the mind, which is defined at least partly by the environment.

Here are two color-coded examples. Examples are from [15]

The lines mean:

1. A Hungarian phrase.
2. The Hungarian phrase segmented into morphemes.
3. My interpretation, in a kind of pidgin, of the meaningful words and morphemes. The colors match the first line.
4. Linguistic glosses, i.e. meanings of words and morphemes, using abbreviations.
5. English translation
6. Literal translation

Example 1

1. A szobában ülö gyerekek **játszanak**
2. **a** szobá-ban ül-ö gyerek-ek **játsza-nak**
3. **the** **ROOM**-in_it **SIT**-doing_**CHILD**-many_of_them **PLAY-they_do**
4. the room-INESS sit-PRSPART child-PL play-3PL
5. 'The children sitting in the room are playing.'
6. Literally: The in-room-sitting children play.

Example 2

1. A gyerekek a szobában ülve **játszanak**
2. **a** gyerek-ek **a** szobá-ban ül-ve **játsza-nak**
3. **the** **CHILD**-many_of_them the_**ROOM**-in_it **SIT-while_doing** **PLAY-they_do**
4. the child-PL the room-INESS sit-VERBADV play-3PL
5. 'The children are playing sitting in the room.'
6. Literally: The children, in the room sitting, play.

I bet most readers can understand almost any Hungarian phrase presented like the third line. Moreover, any language can be understood in the same way. This property of language is the basis for of automatic translation. I added line 6 in order to be as close to the Hungarian text as possible, even at the cost of "correctness." Yet the correspondence between the phrases is somewhat loose. There is only one definite article a in the first example, but two the in the English translation. The Hungarian text does not give a clue regarding the English tense, because there are only Past and non-Past tense markers in Hungarian verbs. The Hungarian continuous tense is extinct. Instead, Hungarian has a number of verb forms untranslatable into English directly.

Both examples, honestly, mean the same and can be expressed in both languages also in other ways, e.g., "The children are sitting in the room and playing".

We understand the sentence in Hungarian or Japanese, if translated into an artificial inter-language, because it describes a situation with which both Americans and Hungarians are familiar. A phrase from a textbook of microbiology would not be universally understood. Our language is embodied (George Lakoff) in our human existence.

I have not discovered either America or Hungary with my examples. This subject has been intensely discussed in linguistics. I would formulate it this way: the patterns of thinking are universal because their invisible generators are identical, but the patterns of language are different because their visible generators are different. We master both because we acquire the generators bit by bit and this is all we need. Our mind is a flask where they assemble and our tong pours them out. The ready patterns of adult and peer speech catalyze some patterns of children at the expense of others. But we pick the generators from different fields and they could be different berries altogether.

As I suggested in [2], not claming any revolutionary insight either, because this is the very essence of language, the configuration of the internal source of the utterance is commonly non-linear and it must be linearized, sometimes in a tortuous way, to be verbalized. This where I see the essence of universal grammar, and as any essence it is utterly simple.

In terms of Pattern Theory, grammar is a collection of regular (not correct!) patterns of word/morpheme configurations. From this angle, UG is, most probably, just the innate ability of humans and animals to perform pattern analysis and synthesis, demonstrated not only in language but also in locomotion, perception, hunt, work, dance, rituals, play, investigation, politics, etc. The uniqueness of human language acquisition device, however, is undeniable. The speaking mind has to convert the nonlinear content into a linear message at one end of communication channel and reconstruct the content at the other end—a far cry from learning to dance or bake pizza by just watching how it is done and repeating the motions in the Euclidean space.

It is my personal impression that PT plays the role of mathematical physics of complex combinatorial systems to which all chemistry and manifestations of life on earth from the life of a cell to society belong. In this area not only deterministic equations are usually powerless but even the probabilistic theories get stuck in the mud for a simple reason: in evolution and history every global (defining) event is unique. It does not belong to a statistical ensemble, while local events do. It can be comfortably viewed, however, in local terms of breaking and interlocking bonds, which is the area of expertise of chemistry.

So much for the kinship of chemistry and linguistics. What about linguistics and biology? Speaking not as a chemist but as an adventurer, I would say that the language in the form of the second or third lines of the above example is, in my opinion, the closest we can have approximation to the most ancient pattern of tribal languages, well after Nean. As we had lost our tail and are now losing classical music, so English lost its cases, Hungarian lost its Present Continuous, and Russian its vocative case.

> What they all acquired was a great syntactic complexity of compound sentences to describe complex ides, situations, or just to show off, as is proper for a performance art such as circus.

With similar experience, we will understand each other whether we say

ROOM-in_it **SIT**-doing_**CHILD**-many_of_them **PLAY-they_do,** or

CHILD-many_of_them **PLAY-they_do** **SIT**-doing **ROOM**-in_it , or

many_of_them-**CHILD** **they_do-PLAY** **SIT**-doing in_it-**ROOM**-in .

Moreover, unless our life experience is radically different, we will understand each other even in Nean [2]:

child! child play! child sit! sit room! play sit !
child! child room! play child! play room!

I have an impression that we encounter difficulties with automatic translation not because the problem itself is complicated, but because our civilization is a real mess. What do you think the words *universe, magma, ring, loop, variety, envelope,* and *signature* means? They all are mathematical terms. *Atom* and *molecule* are terms of propositional logic. I suspect that tribal languages in their traditional pre-technological forms are the easiest to cross-translate if the subject is traditional, too.

5. Notes on notation

Formalization in chemistry is of little value. The rules of the chemical grammar can be easier described than formalized, especially, for a chemist.

I have a subtle grudge against mathematical formalism in its dominating form: it is based on the axiom of closure, which, coming from the Aristotelian requirement of the permanence of the subject, means that the set of terms during the discourse remains unchanged. It efficiently eliminates any ability of mathematics to formalize the phenomenon of novelty and evolutionary invention. As far as I know, only Bourbaki [16]

> in the concept of the scale of sets attempted to cover, albeit in a skeletal way, the unusual subject of novelty and, therefore, evolution. Acquisition is a particular case of evolution. How can you acquire something that is already in your bag? Until I am seriously rebuffed by a mathematician, I swear never to miss a chance of drawing attention to it.

Although my aim here is to outline some ideas in an informal manner, a notation can help clarify them and to show how language is truly embodied in reality and what chemists mean when they speak their lingo.

I use figure brackets and three other kinds of symbols: letters, simple lines, and special arrows $\Rightarrow$ or $\Leftrightarrow$. The letters can signify sound, word, phrase, sensation, image, trace in memory, idea, etc. The lines are binary relations (if directed, the line turns into a simple arrow) , and the special arrows mean causation. The signs, including the brackets, suggest, in sufficiently vague terms, that we deal with real world objects and processes displaying in real time and topological space. The underlying nature of these processes is completely beyond the scope of the current discourse. The chemical parallels, however, are clear: letter is an atomic object, line or arrow is a bond, and the special arrow is a transformation. Symbols of chemical reaction $\rightarrow$ and $\rightleftarrows$ are the verbs of the chemical language, if you will, and $\Rightarrow$ and $\Leftrightarrow$, which I use outside chemistry instead of them, are also verbs of a kind. As I will try to show, the chemical formalism, while reflecting chemical ideas, points to much more general concepts, rising some strange questions that never bother either mathematicians or linguists or computer scientists.

$\{A,B\}$ means not just the set with two elements, A and B. This is more like a topological neighborhood: the elements of the same set in some sense are **close** in space or time. More specifically, the two elements appear together in some situation. In case of language acquisition, $\{A,B\}$ means that A and B are words, objects, or traces in memory perceived or recalled within a relatively short span of attention. They are, so to speak, pushed by chance or intent to face each other—a common thing in psychology and neurophysiology. In case of a chemical process, they are on collision or within a close spatial range or just are dashing around in the same flask, of course, not without the chemist's hand in it.

A—B means that elements A and B are **bonded**. AB means the same.

A→B may denote a directed bond, which we find in configurations of thought and many chemical bonds.

$\{A, B\} \subset \mathbf{G}$, $\{A, B\} \Rightarrow AB$, $AB \Rightarrow C$, $\{A, B, C\} \subset \mathbf{G}$,

means a chain of transformations leading to the expansion—this is synonymous with acquisition—of the generator space $\mathbf{G}$. First, A, B, elements of $\mathbf{G}$, brought together, form bond AB. Configuration AB generates a new element $\mathbf{C}$, which also enters $\mathbf{G}$ as a generator. C is the sign that denotes doublet AB and it can use instead of it. This relation is reversible: $AB \Leftrightarrow C$.

This is the tricky point where the strange questions arise. For support I can only turn to the authority of Bourbaki, who builds the scale of sets in this way. Nothing like that can really happen in nature, where matter has no double existence, but is natural in the mind. **The mind is the second existence of the world and the language is the second existence of the mind.** When we talk about really profound subjects, things are never completely clear and, as Niels Bohr once noted, opposite statements are both true. Only a chemist can confidently say that A and B **reversibly** combine into a very stable C, which does not exclude the existence of free A and B. For the chemist, AB and C are **in equilibrium**. For the rest, it is just a far-fetching metaphor. In cognition, however, we can find a more sympathetic reception: C is a **sign** of the **category** to which AB belongs. When we think about **cat**s and **dog**s, **pet**s are kept in mind, and if we think about **pet**s, **cat**s and **dog**s pop up.

In PT we are like fish in water: C is the identifier of a composite generator. Two bonded generators (doublet AB) can be regarded as a new generator and assigned a separate symbol (C), which does not erase either A or B from generator space. The bonds of C are whatever wholes are left after AB is bolted together.

We could spend a lot of time hairsplitting over the relation between set theory and Pattern Theory, mathematics and the world, theory of meaning, but I am least of all qualified to do it. What we are talking about is a very generic and universal thing: the hierarchy of building blocks, with which linguists, chemists, and engineers deal every day. The existence of dining tables is no threat to the existence of either the table boards or the legs in the inventories and storages. The existence of grammar does not jeopardize the prosperity of either syntax or morphology.

> The chemists practically always deal with multisets in the sense that atoms and molecules are present in enormous number of copies. I believe that the ability to build a hierarchy of signs is as profound property of mind as physical aggregation is a property of matter, but symbols, signs, and shortcuts do not belong to matter. This is how I would formulate the strange problem, which a mathematician could, probably, clarify: is the scale of sets or, to take a simpler example, the set of all subsets (**power set**) a multiset? If yes, than there is a non-trivial set which is also a non-trivial multiset. If no, then any set of sets is always their union. All I could find was that obvious statement that the power set of multisets is a multiset and that in computer science it is convenient to regard a power set a multiset, but the references were not reliable.

In connection with the strange problem, I would refer to memory—the crucial part of cognition and, more generally, life. Mathematics and the universal Turing machine assume an infinite memory. We recognize and/or memorize NEW and recognize or forget OLD. In human mind, if we do not use $\{A, B\}$ anymore, it is forgotten, but C can remain, and *vice versa*.

> Formally, $AB \Rightarrow C$ reminds the composite arrow of **category theory** (CT) [10] , which exerts an unclear to me and some others influence on linguistics, starting with Chomsky. Generators, obviously, are different mathematical objects, but both objects are associative.

I am determined to shun any discussion of what the terms **close**, **appear, bond**, **stability**, **reasonable**, **result**, etc., could mean: their meanings follow from their use, as

Wittgenstein believed, and this is why we sometimes cannot understand each other. What is important, all such terms can have measures. If close, one can ask, then how close? If something is stable, than is this more stable than another? If it results, transforms, or appears, then how fast? To ask such questions is a deeply ingrained habit of the chemist and the natural scientist in general. The remarkable aspect of PT is the ability to provide the framework for answering them regardless of the particular subject. Thus, approaching a speech generation problem from this typically chemical angle, we might decide that not the most grammatically and semantically correct, but the fastest to generate utterance will be produced and, probably understood in context. Similarly, in the social and political matters, not the most reasonable in the long run but the easiest to implement decisions are most often taken, falling into the range from symbolic to violent actions.

To conclude this session of a self-examination of the chemical mind, I would like to touch the evolutionary nerve of a chemist.

While the general principles of evolution are a separate topic, far from consensus and not to be discussed here, the chemist's view of evolution is more settled. It shapes the overall chemical attitude toward building any complex system. In a few words, in the style of *Poor Richard's Almanack* , it is as follows.

1. **Easy does it.** Complex systems are built from the simple ones in simple steps. This is the most axiomatic statement from which the other two are partly deducible.
2. **Rome wasn't built in a day.** Therefore (see 1), the building of a complex system starts from a simple systems.
3. **Do not change horses in the middle of the stream.** The steps are similar throughout the evolution because when the step is simple (see 1), there is not much margin for variation.

In Poor Richard's words, "Haste makes waste," which is, actually, a nice definition of generalized temperature.

From the PT standpoint, which is directly translatable into the chemical mindset, the acquisition of language consists of pattern analysis of speech, i.e., identification of generators, partition of generator space into classes, and selection of **stable** (regular) patterns that partition the configuration space. What is called Pattern Synthesis is the actual production of configurations, which is of no interest for us here, but should be for those working in speech generation.

The steps will be described below. The reader should keep in mind that this is not a linguistic discourse but just a series of variations on the theme of "***If words were atoms.***"

6. Bonding

The concept of atomism is usually presented to schoolchildren as the granular structure of matter, but Lucretius, following Democritus, saw in bonding an intrinsic property of atoms:

> But now
> Because the **fastenings** of primordial parts
> Are put together diversely and stuff
> Is everlasting, things abide the same
> Unhurt and sure, until some power comes on
> Strong to destroy the warp and woof of each:
> Nothing returns to naught; but all return
> At their collapse to primal forms of stuff.
>
> (*On the Nature of Things*)

Chemical bond is well understood today. Formation of a bond, however, is a very general property of the world, extending far beyond the inanimate matter. Bonding, of which both Pavlov's dog, salivating at the sound of the bell, and the inseparable Romeo and Juliet are two quintessential examples, is neither specifically human, nor specifically linguistic phenomenon.

My central idea of language acquisition (most probably, not new) is: the **new** word never introduces itself alone.

When a young monkey hears the warning cry of another monkey for the first time, the sign consists of a single "word," but it associates with the subsequent sight of the predator and either a specific collective behavior or just a general commotion. If the predator never appeared after the cry, the effect would fade away.

A word heard by a child for the first time comes either with another word or an object, gesture, action, sensation, appearance, etc. , i.e., as $\{A,B\}$, which leads to:

$\{A,B\} \Rightarrow AB$, which implies $AB \Rightarrow C$:

$$\{A,B\} \Rightarrow AB \Rightarrow C; \quad \{A,B,C\} \subset \mathbf{G}$$

Two sensations that are close in time or space develop a bond between their representations. For example, the audible word "dog" and a visible particular **dog** can form a link.

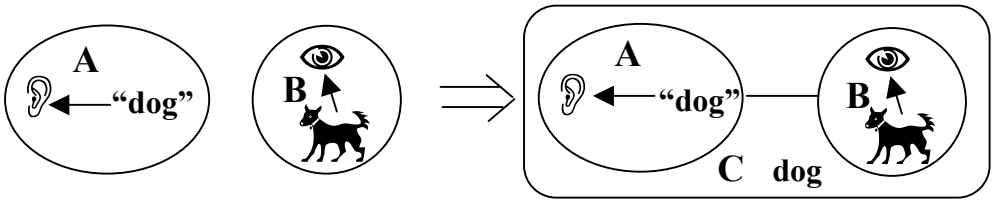

Figure 4. Linking of sound and sight and generation of the idea of dog

A and B can also be two words or just any sounds and their combinations, or an idea and a sensation.

In short, if two sensations belong to the same temporal or spatial set, the bond between their representations "may follow," which the symbol $\Rightarrow$, similar to the chemical $\rightarrow$, signifies. "May follow" means that this is possible, but how probable, it remains to be investigated. Moreover, the bond can fade away without repetitive stimulus.

Chemistry is very unenthusiastic to the distinction between true and false, but takes a great interest in the questions like "how much? how soon?" even if the target of the question seems false.

I am greatly tempted to extend the chemical analogy even further. All chemical reactions are by their very nature reversible. Only because of some special circumstances, like the irreversible escape of carbon dioxide, baking soda and vinegar cannot be reconstructed from the remaining products of their mixing. In a closed steel tube the equilibrium would be reached.

The correct symbol for chemical reaction is $\rightleftarrows$, what is usually meant by the single arrow. If we modify our $\Rightarrow$ into $\Leftrightarrow$ (chemical $\leftrightarrow$ means a very different thing than $\rightleftarrows$), we will suddenly find ourselves facing the beautiful idea that if $AB \Leftrightarrow C$, then whether we somehow get in the focus of attention a **single** A, B or C, all three generators, A, B, and C will float there. It is only because of the limited capacity and speed of our mind that any word does not bring into memory at least one tenth the entire *Webster's II New Riverside University Dictionary*. ***If words were atoms***, this is what would happen in an infinite time.

Whether we should embrace the idea or not, I would spare it for a separate discussion elsewhere. It is related not to the subject of the poverty of stimulus but to the more general subject of deep analogies between all **natural** discrete complex dynamic systems and to the naïve but deep questions like what is the difference between the dog and the word "dog " and why there is no tangible thing called "animal" in the world.

If a word or sound is not tied to another word or sensation or something else, it is meaningless. All theories of meaning agree on it. Only some words are signs of external reality, but all of the atoms of language are meaningful, as a monolingual dictionary testifies.

7. Acquisition of generators

The visible delimiter, i.e., the space between words, is absent from Chinese, Thai, and ancient texts, not to mention speech. Japanese gives some good queues. It is not seen and mostly not heard by children. Morphemes are embedded into words, too. If the child, ignorant of any theory, identifies verbal generators in the input and uses them, there must be a simple procedure to identify sound bites and their sequences as generators.

The identification of generators is the purpose of pattern analysis. How do we know that a sign (word, morpheme, or phoneme) is a generator? The definition of an atom based on indivisibility is a negative one. There is no way to test the indivisibility *a priori*. Lo and behold, the atoms became indeed divisible. For a dog, *Good boy* is probably atomic. On the contrary, divisibility of speech can be easily established before the entire language has been acquired.

Let us try the following simple rule based on divisibility:

The word or morpheme is a generator if it enters at least two different configurations. This is how the rule looks in our quasi-chemical notation, defenseless against any mathematical criticism:

$$\{XB, XC\} \Rightarrow X \in \mathbf{G}$$

Or: if $\{XB, XC\}$, then, probably, $X \in \mathbf{G}$.

In other words, if at least two generators from a generator space **G** can form bonds with X, then X is a generator and it belongs to **G.** To put it even simpler, generator is what can bond with other generators but not what cannot be split into generators.

By the same token, if a configuration can bond with other configurations or generators, it is also a generator. This is especially obvious for linear sequences. In writing, the hierarchy of generators is usually portrayed by using brackets of various types. Naturally, formal linguistics uses tree diagrams.

The ultimate simplicity of local rules like the one just described, hypothetically, requires simple innate physiological mechanisms common for all species with nervous system. It does not require learning.

8. Acquisition of bond space

The sophisticated educated language is acquired by different means involving analysis, sometimes slow, of complex sentences and rhetorical devices as well as contact with complex subjects and situations. To learn chemistry, for example, means to learn what you can say about it that will be grammatically, contextually, and factually acceptable, although not necessarily true. I believe that it is crucial for understanding language acquisition to remember that even a complex literary, philosophical, or scientific text could be told in a simplified childish syntax by cutting the sentence into simple segments.

Here is a single sentence form *The Portrait of a Lady* by Henry James:

> The large, low rooms, with brown ceilings and dusky corners, the deep embrasures and curious casements, the quiet light on dark, polished panels, the deep greenness outside, that seemed always peeping in, the sense of well-ordered privacy in the center of a "property"--a place where sounds were felicitously accidental, where the tread was

> muffled by the earth itself and in the thick mild air all friction dropped out of contact and all shrillness out of talk--these things were much to the taste of our young lady, whose taste played a considerable part in her emotions.

In the beginning, the conversion would go smoothly:

> There were rooms. There was a lady. The embrasures too. The light too. The lady was young. The rooms were large. The rooms were low. The ceilings were brown. The corners were dusky. … The lady liked the rooms. ...

But we would soon run into problems with *these things were much to the taste of our young lady*, which still can be dealt with, but *sounds were felicitously accidental* , and *whose taste played a considerable part in her emotions* are absolutely beyond a child's experience. They are not observable by anybody but the author, in accordance with his esthetic position. To say and understand something like the above sentence requires not only a significant life experience but also an experience in reading literature. Moreover, it requires time to compose and optimize it. The average pre-school child naturally acquires language as a tool to study the higher floors of the edifice at school where the artificial language of the civilization dominates.

I believe that **natural** language acquisition ends with the ability to say who does what in connection to what or whom and in what fashion. The rest, starting with mastering compound sentences, is acquired by learning, analysis, synthesis, and conscious mimicking. It was my personal impression that the Russians with up to seven years of school, especially, in the countryside, hardly ever used compound sentences with more than one clause, which was not to the detriment of content.

Many technicalities of general PT become highly simplified for linear configurations, even more so if we speak only about acquiring a language sufficient for a child to maintain balance with the limited social environment.

The generator of utterance has only left and right bonds. We can assume that they have bond value spaces B_L and B_R on each side. All β are various, possibly, nested tags (*modalities*, along Ulf Grenander [13]) of the generator that signify its multiple categorization: dog is noun, animal, direct or indirect object, etc. on the left, subject, noun, animal on the right, etc. Its grammatical tags can be expressed as morphemes or even by capitalization, as in German, but I am interested how the bond space can be acquired in childhood and not analytically.

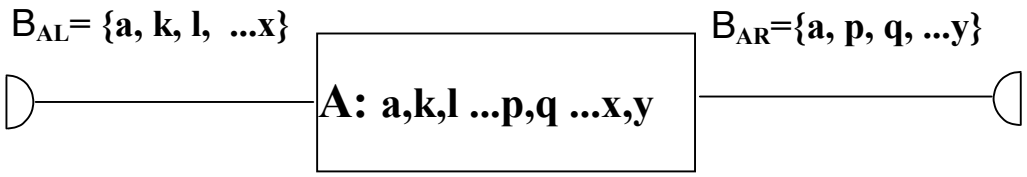

Figure 5. Language generator A

In Figure 5, generator A has some bond values twice on both sides, which for example can happen with adjectives or listed nouns. The union of the left and right sets is a complete list of all bond values, which may be enough for languages with loose word order.

For the left-to-right doublet **AB**, bond relation ρ = **TRUE** if

$$\beta_{AR} \in B_{BL} \text{ , and } \beta_{BL} \in B_{AR} \text{ ,}$$

i.e., if the two generators have the same bond values at the bonds in contact (Figure 6). Thus, "Dog bites man" and "Man bites dog" are equally grammatical. What contradicts the factual content of the sentence, however, decreases the **stability** of the wrong version. This is why some politicians love meaningless clichés.

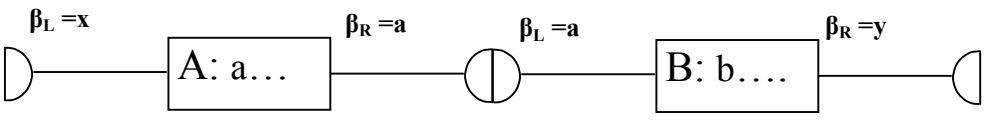

Figure 6. Regular doublet

It all looks much simple for a chemist who would just say that A and B form a donor-acceptor couple. Even a linguist could not find anything new in the statement that there are grammatical and ungrammatical adjacencies between morphemes, words, and phrases, which is a fair way to put it. I only paraphrase common linguistic knowledge, although I suspect that for any common knowledge in non-experimental linguistics there are two conflicting views. What I claim as new (what can be new after Lucretius, anyway?) is that the rules of grammar, whether universal or specific, extant or extinct, do not need to be stored anywhere in the mind as a book on a shelf. They could be contained—and this is a hypothesis in need of a test—in the properties of generators, similarly to the way molecules assemble not by principles and parameters but by the properties of the atoms.

The knowledge of regular grammar, from the point of view of a chemist, is distributed among generators, up to a certain level of language evolution. Philosophers, scientists, and writers have inflated the language to such excess that poor children have to study their native tongue for many years at school, picking up irregularities from the peers.

9. Acquisition of bond values

The next important question on the agenda of language acquisition is generator classification. How are the bond values, which are signs for classes, acquired? Here is an extremely simple and local rule:

$$\{AB, AC\} \Rightarrow \{B,C\} \subset \mathbf{G'} \subset \mathbf{G}$$

It means that if two generators B,C combine with the third A, they have the same bond value, acceptable by A. In other words, B and **C** belong to the same subset

G′ of **G**. It means that the partition of generator space into classes is done by the generators they can share. "Tell me who your friends are and I will tell who you are" is invertible: "Tell me who you are I will tell who they are." The newly formed class can expand in the same way or disappear if the juxtaposition of AB and AC was accidental. This is how language can be acquired in a quite mechanical way by children who would give very little thought to anything but fun and who will rise to higher levels of language only when they develop abstract thinking that takes time and is not automatic.

Of course, this is only a hypothesis. Probably there are some supporting or contradicting works in linguistics literature.

Remarkably, every act of juxtaposition of two doublets with a shared element works both ways: (1) identifying a new generator and (2) identifying a class:

$$\{A—B\ ,\ A—C\} \Rightarrow\ \ A \in G$$

$$\{A—B,\ A—C\} \Rightarrow\ \ \{B,C\} \subset \mathbf{G'} \subset \mathbf{G}$$

Next, $\{B,C\} \Rightarrow D$! The class acquires its sign.

The effect depends on whether A or the bond with A is **new**. The notion of **novelty**, absent from mathematics and physics, but not from chemistry, where a molecular structure does not exist unless ***a posteriory*** and ***de facto***, is of cardinal importance for the formalization of evolution, which is exemplified by language acquisition. What the original UG concept seems to say is that there is nothing **new** in language acquisition but just selection from a timeless menu.

> A stone falls from the Tower of Pisa? No big deal, all possible trajectories and acceleration existed before the fall. The stone has chosen its trajectory, which is not so strange an idea for classical optics and quantum physics.

Nevertheless, formal linguistics, driven by Chomsky himself, has been undergoing such an involution toward simplification (X-bar) and quantification (optimality), that sooner rather than later it is going to fuse with the principles of PT even,

like Monsieur Jourdain, without realizing it. I draw attention to PT not because I do not believe that linguistics cannot find its own way to consensus, but because it illuminates the place of linguistics among other natural sciences, where my native chemistry dwells nearby.

I try to show in Figure 7 how the X-bar concept can be converted into its PT form by excluding **imaginary** XP and X'. The resulting generator has arity 3 (number of bonds) and is not good for building strings. This is why formal linguistics in all its forms has to bend over backwards to find **imaginary** movements and linearize the typically non-linear trees.

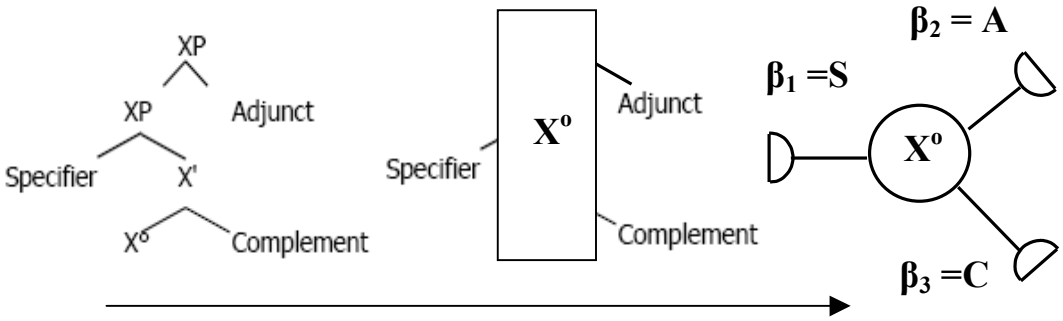

Figure 7 . From X-bar (left) to generator (right). A: Adjunct, etc.

If the notation {A,B}, as I said, means that A and B are not just elements of a set but are **close** to each other in some realistic sense, the transformation sign $\Rightarrow$ in $\{AB, AC\} \Rightarrow \{B,C\}$ needs clarification. If the elements are on the left of $\Rightarrow$, they are close in perceived reality. We say that they are in the same topological neighborhood in time and/or space and form a distinct cluster. There must be some physical or physiological reason why we assemble them in the brackets. Coming back to the analogy with the singles bar—a refreshing step away from X-bar—the brackets on the left of $\Rightarrow$ represent the singles bar, say, at 8 PM and on the right we see the same bar at 10 PM.

The closeness does not mean Euclidean distance. A good example is a circle of friends who may be separated by the entire continent or just a street but call each other over the telephone with the same minimal effort. A better term is channel of communication, for which the maximum distance between humans we know is from the Earth to the Moon. At the other end of the human scale are communicating neurons. The physical counterpart of communication is interaction and the chemical one is collision.

Human intellectual functions shuttle between the mind and the real world as—you've guessed it right—the bees shuttle between the flowers and the beehive.

Expressions like $\{A,B\}$ and $AB \Rightarrow C$ suggest another obvious idea: the concept of a strong bond is an expansion of the concept of the set. In real world and in the real mind, elements are placed into the same set for a reason. They **might** form a bond if they stayed in the set long enough, and fuse if when they stayed even longer. We can say that $\{A,B\}$ and C are at the two ends of a continuous scale with **AB** somewhere in the middle.

10. Locality

Locality in current context means that in the process of acquisition and identification we look only within a 1-neighborhood (immediate topological neighborhood, see Figure 7) of the generators and not any farther than that. We do not need to consult either a grammar or an algorithm. No long term memory is needed to keep intermediate data because they all are on hand.

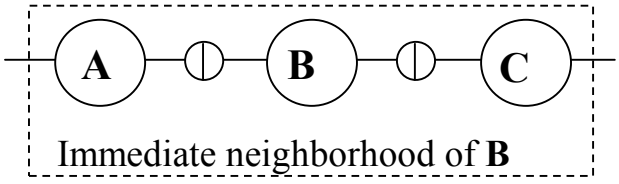

Figure 8 . The substrate of local operations

The target for both identification and classification is the same: two generators in the neighborhood of the third one. The identification of the generator (if generator combines with two others... etc.) is simply seen from the other side: if two generators form bonds with the same third one, they belong to a subset (class) of **G**. Therefore, the class is defined by the affinity (or aversion) of all its elements to a common generator.

As I believe, what the child gradually acquires is not any grammar as the list of rules, like the basic word order SVO, but the hierarchical partition of sounds into morphemes, words, word groups, phrases, and stylistic devices that constitute the generator space in which generators have specific bond structures, so that SVO order in English comes out automatically as soon as the abstract generators S, V, and O are formed. Of course, the pre-school child has no idea about syntactic categories. When the speech is generated, the content and form are reconciled in the process of linearization toward the minimization of stress [3].

11. Some examples

This is an imaginary way how the high level generators and patterns can be acquired in a child-robot:

{eat—apple , eat—carrot } $\Rightarrow$ {apple, carrot} $\subset G_1 \subset \mathbf{G}$.

{eat—apple, take—apple} $\Rightarrow$ {eat, take} $\subset G_2 \subset \mathbf{G}$

Pattern: G_2—G_1

{Mary—eat , Mary—take } $\Rightarrow$ Mary —G_2

Pattern: Mary—G_2—G_1

{Mommy—eat, Mommy—take} $\Rightarrow$ Mommy—G_2

{Mary—G_2 , Mommy—G_2 } $\Rightarrow$ G_3 = (Mary, Mommy) $\subset G_3 \subset \mathbf{G}$

Pattern: G_3—G_2—G_1

Much later the child learns at school that G_3 , G_2 , and G_1 are terms of syntax and discovers that he or she had been guided by the invisible hand of the Grammar.

Here is another imaginary example, inspired by observations on acquisition of German noun gender by native children [10A] acquisition of the German case-gender-number nominal marking system.

Vocabulary: Hund: dog; Hundchen: puppy dog, Mäd: the word does not exist; Mädchen: girl; -chen: suffix of diminutive form, marker of Neutral Gender.

$\{\text{Hund, Hundchen}\} \Rightarrow \text{Hund-} \in G$

$\{\text{ Das\_Hundchen, Das\_Mädchen}\} \Rightarrow \text{ -chen} \in G$

$\{\text{Ein\_Hund-chen, Das Hund-chen}\} \Rightarrow \text{ -Hund-} \in G$; similarly,

$\{\text{ Hundchen, Mädchen}\} \Rightarrow (\text{-Hund- , -Mäd- }) \subset G_{\text{-chen}}$

$\{\text{Das A-chen, Das B-chen}\} \Rightarrow \text{Das—}G_{\text{-chen}}$; $(A, B) \subset G$

Next, consider the basic word order of subject, direct object, indirect object, and verb. For English Imperative, the triplet in Figure 9 serves as a template.

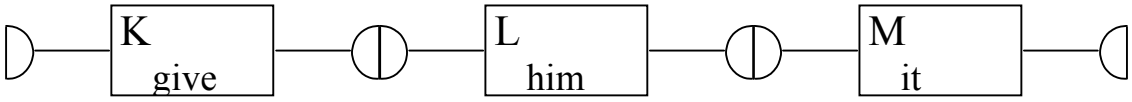

Figure 9. Pattern template

Here K and M are in the immediate neighborhood of L. K, M, and L represent whole classes of generators.

The triplets constitute already a transition phase from Nean, the language of doublets, to the advanced grammaticalized language, in which the alternatives of word order start branching, leading to the abundant, but not unlimited, variety of languages.

The probabilistic approach to syntax in [18] comes very close to the idea of Nean as the language with doublet or triplet syntax resulting from haplology [3]. Moreover, this work

suggests another idea: the sentence is generated from overlapping short fragments. I would call it **tiling**.

The advanced natural evolution of syntax when there was still no Henry James and William Faulkner in sight is a separate problem. As a mental experiment, we can try to guess why, under the pressure of linearization, in polysynthetic languages like Mohawk the **L—M** doublet develops into a verb prefix, in highly inflected Slavic languages, like Russian, the same doublet takes form of case endings, and in English it mostly evaporates, leaving only the ephemeral word order:

Jack he-her-like Mary (Mohawk)
Jack like-he Mary-her (Russian)
Jack like-he Mary (English)

All three syntactic constructs contain all necessary information to avoid ambiguity.

With an indirect object, "Give Alex the toy" would look like "You-him-it-give toy Alex" in a strongly polysynthetic language and "Give-you toy-it Alex-to" in a highly inflected language. Nevertheless, I could not find in my limited corpus of Mohawk, collected from the Web, anything like "you-him-it," which would generate up to a thousand of different verb prefixes.

Sak wa'-ku-hsvn-u-'.
Sak fact-1sS/2sO-name-give-punc (linguistic glosses)
Sak indeed-I-you-name-give-this_moment
'I (hereby) give you the name Sak.'

Here (the example is from [19]) *punc* (' , which is a sound) means that the action is one-time and *fact* (wa') means the factual mood. Morpheme *ku* means "I (subject) [do it to] you (object)" and "Sak" is not marked by any morpheme. The distinction between direct and indirect object, sharpened by word order or preposition in English, is blurred here, but the content is absolutely clear. This confirms to me that the distinction between

syntactic categories like direct and indirect object is rather artificial. In the configurations of thought [3] only bond couples are real. Moreover, I think, together with some linguists, that the grammatical categories like verb and noun are not without borderline fuzziness, especially in the non-Indo-European languages, but the Indo-European origin of grammar strongly influences our thinking.

In the above Mohawk example [19] , "name" (*hvsn* **; v** is a nasal vowel) immediately precedes "give**" (u** , another nasal vowel**)** creating the compound verb "namegive" (hsvnu), in a manner common also to German. But the Russian language does not mind scattering the three endings all over the phrase. The diluted medicine is less bitter, but the body of the Russian sentence can grow bloated.

> We can speculate about why the extraordinary redundancy of Russian and Polish (shared, it seems, by Swahili) has survived thousand years, satellites, and perestroika and why the inflective redundancy of Old English was so fragile. The tribal languages, it seems, aimed at rendering the nonlinear thought in the most straightforward way, making clear who did what to whom, without requiring any guessing. This holistic property of tribal languages—and they can have an astonishing complexity—was, probably, perpetuated by a limited number of situations meaningful for the tribal society. We can find the most generic ones in folk tales. This became an atavism after entering the modern era with literacy as the most powerful stabilizing factor.
>
> Unlike the English nation, the Russians did not know intense ethnic mixing. The Mongol invasion and 240 years of domination added some words to the vocabulary but the two people did not actually mix. The Mongols either lived a separate life or accepted the Russian one.

With PT approach we might find that the excruciatingly thorny problem of the evolution of language is not hopeless. Like in chemistry, we may find out what is "more true" and more expedient **at given conditions**, which are accessible to archeology. I can imagine a research—no doubt, very difficult—of relation between the way of life and the structure and vocabulary of tribal languages, but I am not aware of it. To separate language from conditions of life and culture would be equivalent to ignoring conditions

of chemical reaction, i.e., temperature, pressure, agitation, catalysts, acidity, irradiation, etc., by a chemist. Whether language defines culture, I have no opinion. But if culture defines language the observable result will be the same. This is yet another case of circular logic, which can be resolved only in one way: the less stressed complex of culture and language survives.

> Linguists cannot forget the debates over Benjamin Worf's denial of category of time in Hopi. It turned out that he was partly right. Some strange things with tenses happen also in Mohawk [19]. There is a sophisticated system of Future Tenses in Maya. In my view it is only natural that the perception of time in a tribal pre-industrial society could be very different from ours, shaped by timetable, heat engine, clock, and fertility of imagination. I feel comfortable with the idea of the past as what will **always** be in the future, like the death of a relative, and the idea of the future that is a pure **possibility, intent** (I will eat)**,** or just **present**—all without any guarantee of realization. The present can be expressed not only by *I'm going to eat* , *je vais manger* , and *eszni fogok* (Hungarian), but also by simple Present used in Russian and Hungarian as Future with an adverbial modifier of time. The Bulgarian Future Perfect, with two auxiliary verbs and an unusual pattern of change, is truly remarkable from this viewpoint, as the entire florid Bulgarian verb system is. On the other hand, the ghostly Future in Japanese is another convincing illustration of the idea that, come to think about it, there **is** only Past and non-Past in the naïve physics of the world. We can talk about future, but it certainly does not exist. Every philosopher starts with inventing his (this is a truly manly occupation) own language.

12. Language and homeostasis

From the PT point of view, the language generation amounts to the problem of **pattern synthesis**, which is the production of regular configurations of the same pattern.

Our next and last step is pattern synthesis during language acquisition. How to ensure that the acquired language takes a grammaticalized shape and not just any shape,

but that of the surrounding language? How are patterns selected? In general, the generator space does not guarantee a unique way of self-assembly of generators into configuration neither in chemistry, as isomers demonstrate, nor in linguistics, as the languages of loose word order and even the two forms of indirect object illustrate.

Unfortunately, as I believe, chemistry can tell us at this point very little, if anything at all. This pessimism is not shared by the whole school of evolutionary linguistics that simulate language evolution within the framework of competition and selection, based on groundbreaking models of Manfred Eigen and his group. This direction is represented in linguistics by Martin Novak, whose background is in mathematics, biology, and evolutionary dynamics [20]. Instead of criticizing these highly valuable and insightful models, under the spell of which I have been for over 20 years, I will make a constructive (I am sure, not new) suggestion of an ultimate simplicity.

The language acquired by a child differs very little from the ambient one because of social homeostasis. Speaking a non-standard and peer-challenging language and suffer mutual misunderstanding as consequence would create a stress, which most human and animals, except some born leaders and troublemakers, would avoid with any available means. In psychology it is known as theory of balance.

The patterns of speech based on the acquired generator space will be selected not so much by individual selective advantages as by the stability of the whole. Here I have little to say but to refer to [3] and [4].

I cannot resist a temptation, however, to generalize this principle over biological evolution, the area which, in spite of Darwin and molecular biology, is as far from consensus as linguistics. Evolution of species is not (I say it arrogantly, without any "not only") the survival of the fittest, because the fittest is always the one who survives, but the homeostasis of the biosphere subject to external (for example, climatic), internal (for example, cyclic or catastrophic non-linear fluctuations), or just random perturbations.

A non-equilibrium dissipative system, to which all life (*biosphere*) and its manifestations (*noosphere*) belong, searches and finds a way to end the stress of the perturbation. I believe this idea follows from the ideas of Ilya Prigogine [21] and William Ross Ashby. [22] The concept of punctuated equilibrium [23] is the closest to

it. One of the cardinal insights of this entire approach is that the disturbed complex dissipative system returns not to the previous state, which is hard to find among enormous number of possible states and pathways, but to a **new** and more stable one. This is what makes dissipative systems so different from common chemical systems which automatically find the point of equilibrium.

Homeostasis is a highly natural way of thinking for a chemist even though very few chemists deal with dissipative systems.

What do you think happens when you disturb the universe by mixing baking soda and vinegar? It dissipates carbon dioxide and comes not to the previous state but to a new one, from which there is no way back to the previous one. Moreover, nothing else can happen there on its own.

You can heat up and cool down a flask with chemicals millions of times with the same result, but the Sun warmed up the Earth and left it to cool down millions of times until life stepped out because the Earth was an open system.

The live dissipative system can change many times without the interference of the chemist or, for that matter, anybody else, while the supply of solar energy lasts.

> To illustrate this idea, the discovery of the mineral fuel and heat engine was a great disturbance of the previous civilization. It first occurred locally, between Manchester and Birmingham. Today it brings global civilization in turmoil. To recuperate, we are burning the mineral fuel in increasing amounts until the homeostasis will be, hopefully or woefully, restored in a new civilization, which may not welcome humans as we know them at all.

Conclusion

Why in the world are we speaking about thermodynamics? Isn't chemistry far enough from linguistics? And isn't it obvious that I cannot **prove** a word. No, I can't. I do not consider myself an amateur linguist. I am only a chemist. But I would like to plant a seed of something other than just a doubt in Plato's idea.

Why the verb in Mohawk, Inuit, and other polysynthetic languages is so loaded with short morphemes indicating moods and aspects, not to mention major syntactic functions, why the verb in Japanese is practically naked and so is the attributive adjective in the noun- and verb-overdressed Hungarian, why French has a fair number of tenses but the otherwise much sparser English is not too far behind, why the Titianesque Russian has no article and the noun-frugal Bulgarian slaps it onto the noun from behind, why English lacks diminutive, derogative, and affectionate suffixes, present in Italian, essential in Russian, and some showing up even in the stern German, and why Yiddish has no simple Past Tense—such questions could be answered if a function similar to energy of a molecule or some other measure of **stability** could be found for any segment of speech—which is just a thought, crudely squashed (but don't **mince** your words!) and drawn through a narrow hole regardless of the word segmentation.

Of course, we speak as our forefathers did, but there was some reason why they had departed from their non-speaking forefathers. If there are some laws of nature, they apply to both our forefathers and little children.

Figure 10. Bees carry pollen, words carry grammar

I believe linguistics could move closer to the status of natural and consensus-based science if one fine morning it discovered in the alien chemistry its own reflection in a gritty, wavy, cracked, but still a mirror. I hope chemists, on their part, could someday realize that they can give something else to the world except pollution, side effects, and genetic danger: the flowers of **new** universal ideas.

As for the language acquisition, see Figure 10.

REFERENCES

1. Mason, Timothy. *Could Chomsky be Wrong?*
http://perso.club-internet.fr/tmason/WebPages/LangTeach/CounterChomsky.htm

2. Baker, Mark C. *The Atoms of Language*. New York: Basic Books, 2001.
Mark Baker's publications: http://ling.rutgers.edu/people/faculty/baker.html

3. Tarnopolsky, Yuri. *Tikki Ttikki Tembo: The Chemistry of Protolanguage*, 2004
http://users.ids.net/~yuri/Nean.pdf

4. ———. *Molecules and Thoughts: Pattern Complexity and Evolution in Chemical Systems and the Mind* , 2003.
www.dam.brown.edu/ptg/REPORTS/MINDSCALE.pdf
Or: http://users.ids.net/~yuri/mindscale.pdf

5. ———. *Transition States in Patterns of History*. 2003.
http://users.ids.net/~yuri/HistMath1.pdf

6. Lakoff, George and Johnson, Mark. 1980. *Metaphors we live by*. Chicago : University of Chicago Press, 1980

7. Lakoff, George.. *Women, fire, and dangerous things : what categories reveal about the mind.* Chicago : University of Chicago Press, 1987.

8. Grenander, Ulf. *Elements of Pattern Theory*. Baltimore: Johns Hopkins University Press, 1995.
Advanced works:

———. 1976. *Pattern Synthesis. Lectures in Pattern theory*, Volume 1. New York: Springer-Verlag, 1976.

———. *Pattern Analysis. Lectures in Pattern theory*, Vol. II. New York: Springer, 1978..

———. *Regular Structures.Lectures in Pattern theory*, Vol. III. New York: Springer, 1981.

———. *General Pattern Theory. A Mathematical Study of Regular Structures*, Oxford, New York: Oxford University Press, 1993.

9. *Language Evolution*. Edited by Morten H. Christiansen and Simon Kirby. Oxford: Oxford University Press, 2003.

10. Matthiessen, Christian and Halliday, M. A. K. *Systemic Functional Grammar: A First Step into the Theory*. http://minerva.ling.mq.edu.au/resource/VirtuallLibrary/Publications/sfg_firststep/SFG%20intro%20New.html

11. Brian MacWinney's home page with a library of papers. http://psyling.psy.cmu.edu/brian/

12. Pullum, Geoffrey K. and Kornai, András. *Mathematical Linguistics* http://www.**kornai**.com/MatLing/matling3.pdf

13. Grenander, Ulf. , *Patterns of Thought*. www.dam.brown.edu/ptg/REPORTS/mind.pdf

14. Rosch, E. *Human Categorization*. In N. Warren (ed.) *Studies in Cross-cultural Psychology*. London: Academic Press, 1977 , vol. 1, pp. 1-49.

———. *Principles of categorization*. In E. Rosch and B. B. Lloyd (eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum, 1978, pp. 27-48.

15 Megyesi, Beáta. *The Hungarian Language A Short Descriptive Grammar* . http://www.speech.kth.se/~bea/hungarian.pdf

16. Bourbaki, Nicolas. *Elements of Mathematics: Theory of Sets*, Boston: Addison-Wesley, originally published by Hermann (Paris), 1968, p.259-382.

17. MacWhinney, B. J., Leinbach, J., Taraban, R., & McDonald, J. L. Language learning: Cues or rules? *Journal of Memory and Language, 28*, 255-277 (1989). http://psyling.psy.cmu.edu/papers/cues.pdf

18. Lafferty, John, Sleator, Daniel, and Temperley, Davy. *Grammatical Trigrams: A Probabilistic Model of Link Grammar.*

http://www.cs.cmu.edu/afs/cs.cmu.edu/ project/link/pub/www/papers/ps/gram3gram.ps

19. Baker, Mark and Travis, Lisa. *Mood as Verbal Definiteness in a "Tenseless" Language* , *Natural Language Semantics, 5*: 213-269 (1997). http://ling.rutgers.edu/people/faculty/baker/mohawk-mood-prt.pdf

20. Nowak, Martin, *From Quasispecies to Universal Grammar,* Z. Phys. Chem. **216** (2002) 5–20. http://www.ped.fas.harvard.edu/pdf_files_old/ZPhysChem02.pdf

Martin Novak's publications:

http://www.ped.fas.harvard.edu/publications.html

21. Prigogine, Ilya and Stengers, Isabelle. *Order out of Chaos* . New York: Bantam, 1984. Also: Nicolis, G. and Prigogine, I. *Exploring Complexity*. New York: W.H.Freeman, 1989. Stengers, I. and Prigogine, I. *The End of Certainty : Time, Chaos, and the New Laws of Nature*, New York: Free Press, 1997.

22. Ashby, W. Ross. *Design for a Brain: The Origin of Adaptive Behavior*, 2nd Ed., New York: Wiley, 1960. Originally published in 1952.

———.*An Introduction to Cybernetics*, London: Chapman & Hall, 1964. Originally published in 1956.

23. Eldredge, N. and Gould, S.J. *Punctuated equilibria: an alternative to phyletic gradualism*, in *Models in Paleobiology*. T.J.M. Schopf (ed.). San Francisco: Freeman, Cooper, 1972, pp. 82-115.

APPENDIX

THE CHEMISTRY OF THE THREE LITTLE PIGS

As a **very preliminary** illustration of some ideas expressed in this paper, I will explore a fragment of a text as a substrate for elementary local operations of generator and bond space acquisition. It is by no means a description of the mechanism itself because the real perception and processing of the text is diachronic, while my examination is going to be synchronic. The text here comes into view not bit by bit, as it should, but as a chunk.

A simulation of language acquisition, as I believe, cannot be done with any single compact segment of perceived speech as input. A minimal requirement of a realistic simulation is a long series of language stimuli, coming in packages, like the Internet content, over an extended time, and against a background of realistic interactive content. This is a difficult task, remotely comparable with designing a game like the Sims, http://thesims.ea.com/us/.

There are scores of works on child language acquisition, and corpuses are available, but direct observations of children are intrusive and difficult to realize on a large and comprehensive scale, as even the works on chimp language testify.

> We have numerous theories of the origin of the universe, life, and language because we cannot observe the origins. Nevertheless, most scientific theories work near perfectly every day. **The bulk** of linguistic theories can be tested by building a talking, writing, and translating machine that develops its abilities in human environment, from scratch, and without any algorithm. This is still easier than to create the universe.

The target text is a compact modified fragment from the tale of *The Three Little Pigs*.

Source: **Jacobs, Joseph. "The Story of the Three Little Pigs."** ***English Fairy Tales.*** **London: David Nutt, 1890.** http://www.surlalunefairytales.com/index.html

The target text P is a character array of 130 words, given here in the form of MATLAB input:

P = char ('_', 'there', 'was', 'an', 'old', 'sow', 'with', 'three', 'little', 'pigs', 'and', 'as', 'she', 'had', 'not', 'enough', 'to', 'keep', 'them', 'she', 'sent', 'them', 'out', 'to', 'seek', 'their', 'fortune', '_', 'the', 'first', 'that', 'went', 'off', 'met', 'a', 'man', 'with', 'a', 'bundle', 'of', 'straw', 'and', 'said', 'to', 'him', '_', 'please', 'man', 'give', 'me', 'that', 'straw', 'to', 'build', 'a', 'house', '_', 'which', 'the', 'man', 'did', 'and', 'the', 'little', 'pig', 'built', 'a', 'house', '_', 'presently', 'came', 'along', 'a', 'wolf', 'and', 'knocked', 'at', 'the', 'door', 'and', 'said', '_', 'little', 'pig', 'let', 'me', 'come', 'in', '_', 'the', 'pig', 'answered', '_', 'no', '_', 'the', 'wolf', 'then', 'answered', 'to', 'that' , '_' , 'then', 'I', 'll', 'puff', 'and', 'I', 'll', 'blow', 'your', 'house', 'in', '_', 'so', 'he', 'puffed', 'and', 'he', 'blew', 'his', 'house', 'in', 'and', 'ate', 'up', 'the', 'little', 'pig', '_')

We apply to the target the following transformations written as "chemical" reactions, in which X is a variable. By "chemical" I mean, actually, "pattern-theoretical," but I cannot use the latter term because the ideas are not explicitly formulated in PT. I infer them perhaps incorrectly. **Equilibrium** is a "chemical" counterpart of **equivalence** and **association** in cognitive sciences. It means, half-seriously, that if one thinks about three little pigs (A), the wolf (B) promptly comes to mind because the entire story (C) is remembered. The story is in equilibrium with all its components, which is pretty close to the chemical idea of equilibrium. I cannot invade the heartland of cognitive sciences, but from a distance I would repeat again the parallel between the chemical flask and the mind.

$\{A,B\} \Leftrightarrow \{AB\}$	bonding equilibrium	(1)
$\{AX,BX\} \Rightarrow X \in G$,	generator identification	(2)
$\{A,B\} \Rightarrow C$	generator categorization	(3)
$\{A,B\} \Leftrightarrow C$	representation equilibrium	(4)
$\{AX,BX\} \Leftrightarrow CX$	bonding categorization and its representation equilibrium	(4)

Using a simple program, a **vocabulary** of 71 words, including space (_), was extracted from P and the words were analyzed for their left and right neighbors in P. The results are in the Table:

Table : Vocabulary and neighborhoods of *The Three Little Pigs*

Left neighbor	No.	Word	Right neighbor
fortune him house house said in answered no that in	1	_	the please which presently little the no the then so
_	2	there	was
there	3	was	an
was	4	an	old
an	5	old	sow
old	6	sow	with
sow man	7	with	three a
with	8	three	little
three the _ the	9	little	pigs pig pig pig
little	10	pigs	and
pigs straw did wolf door puff puffed in	11	and	as said the knocked said I he ate
and	12	as	she
as them	13	she	had sent
she	14	had	not
had	15	not	enough
not	16	enough	to
enough out said straw answered	17	to	keep seek him build that
to	18	keep	them
keep sent	19	them	she out
she	20	sent	them
them	21	out	to
to	22	seek	their
seek	23	their	fortune
their	24	fortune	_
_ which and at _ _ up	25	the	first man little door pig wolf little
the	26	first	that
first me to	27	that	went straw _
that	28	went	off

went	29	off	met
off	30	met	a
met with build built along	31	a	man bundle house house wolf
a please the	32	man	with give did
a	33	bundle	of
bundle	34	of	straw
of that	35	straw	and to
and and	36	said	to _
to	37	him	_
_	38	please	man
man	39	give	me
give let	40	me	that come
to	41	build	a
a a your his	42	house	_ _ in in
_	43	which	the
man	44	did	and
little little the little	45	pig	built let answered _
pig	46	built	a
_	47	presently	came
presently	48	came	along
came	49	along	a
a the	50	wolf	and then
and	51	knocked	at
knocked	52	at	the
the	53	door	and
pig	54	let	me
me	55	come	in
come house house	56	in	_ _ and
pig then	57	answered	_ to
_	58	no	_
wolf _	59	then	answered I
then and	60	I	ll ll
I I	61	ll	puff blow
ll	62	puff	and
ll	63	blow	your
blow	64	your	house
_	65	so	he
so and	66	he	puffed blew
he	67	puffed	and
he	68	blew	his
blew	69	his	house
and	70	ate	up
ate	71	up	the

The following is a kind of chemical analysis of the table.

We encounter some doublets and triplets of high occurrence in everyday speech, for example:

2	**there**	**was**
15	**not**	**enough**
54	**let**	**me**
39	**give**	**me**

she	**14**	**had**	**not**
she	**20**	**sent**	**them**

In chemical language, if used frequently, the doublets and triplets can crystallize and form composite generators, provided the abstract **temperature**, which is the level of chaos, is low enough. In P, however, the statistics is meaningless because of the small size.

The following is a series of examples of what can "chemically" happen with P as a substrate.

1.

25	**the**	**first man little door pig wolf little**

THE creates the tentative class of all words right of THE. The classification may diachronically survive or fall apart. We need a name for the class, and **THE-X** is a natural one.

Class THE-X: X= {first, man, little, door, pig, wolf}

We know that **THE-X** includes both nouns and adjectives, but the child-robot does not know grammar.

2. Similarly:

31	**a**	**man bundle house house wolf**

Class A-X: X= {man, bundle, house, wolf }

These two classes can be expressed in terms of the vocabulary entries.

Since the class is in equilibrium with its entries, MAN is in equilibrium with its class (possibly, one of many).

Class X-MAN : **X = {a, the}** and, therefore, **X= A-X , THE-X**.

Confirmed by many occurrences, this classification will, most probably, survive.

But then A and THE form also a class, for which we are out of the names other than cumbersome A_THE. Of course, we now know the current name of the class: **article**, by the way, absent in many inflected languages.

45	**pig**	**built let answered _**

3. Similarly:

an	**5**	**old**	**sow**

and

three the _ the	**9**	**little**	**pigs pig pig pig**

would allow for inferring the distinction between nouns and adjectives, not quite reliable yet:

X-Adjective-Y: X=Article, Y= Noun

There is not enough data to form the class of nouns, however, but we can easily imagine that with enough verbs.

The above examples could make us feel what a little child feels when acquiring the knowledge of the world: what we know and what seems elementary and obvious must to be retrieved from the formless mass like the statue of David from the block of marble. Unlike the sculptor, who cannot make a big mistake with the stone, the child's mind works like a scientist—or living nature—creating, testing, and rejecting hypotheses.

The overall picture of language acquisition—and, therefore, of language genesis—becomes the field for competition of patterns, which are counterparts of biological species, and not individual sentences. When the starting pattern is as simple as doublet or triplet, further mutations can generate the largest variety of grammars, which explains why the languages are on the surface so different. The mutations of developed grammars are, of course, less radical.

The German separable verb prefixes seem to contradict the principle of locality, but if we start with simple situations and short phrases, German is no more strange than Japanese with its verb invariably at the end.

We can hope to reconstruct the process of linguistic genesis for two reasons: (1) we can understand the world of the first speakers where somebody does something to somebody or something, (2) we have only two choices for adding a new generator (morpheme): left and right of the old one.

The short fragment illustrates only the main principle: **if the words were atoms**, **there would be a chemistry of words**. Linguists can easily see some parallels with the widely used connectionist models, methods of statistical inference, Bayesian categorization, and the so-called memoryless learning algorithms, when the next entry either confirms or contradicts the already formed rule, but the data are not stored. Language acquisition fits into the fast growing area of unsupervised learning. It may turn out that there is much more consensus in linguistics than it appears, but the various areas do not have a *lingua franca*. The cobbler walks barefoot.

It is obvious that the "chemical" view of language has strong parallels with Hopfield Neural Nets with its concept of energy and Kohonen Self-Organized Maps with their triplets. Numerous references can be found on the Web, with one especially relevant:

> Timo Honkela, Ville Pulkki and Teuvo Kohonen, *Contextual Relations of Words in Grimm Tales Analyzed by Self-Organizing Map* (1995) . Proceedings of International Conference on Artificial Neural Networks, ICANN-95, F.Fogelman-Soulie and P.Gallinari (eds.), EC2 et Cie, Paris, 1995, pp.3-7. The paper can be downloaded through CiteSeer.

I will also refer here to the close in spirit and crisp in ideas work of Sylvain Neuvel and Sean A. Fulop, *Unsupervised Learning of Morphology Without Morphemes*, http://arxiv.org/abs/cs.CL/0205072 , where the sign $\leftrightarrow$ in

$$|X|_{\alpha} \leftrightarrow |X'|_{\beta}$$

means "bi-directional implementation," which is very close to what I want to express with my sign $\Leftrightarrow$ and what is the closest parallel to chemical equilibrium. The algorithm for morphological analysis, i.e., identification of morphological generators, as I would say, based on this principle, works on a POS-tagged lexicon. POS here means not *poverty of stimulus* but *part of speech*. What I expect from the three little pigs, however, is the POS-tagging without any training typical for neural nets.

The comparison of "chemical linguistics" with current approaches is a separate topic to be discussed elsewhere.

Being strictly local, the "chemical" or "unsupervised" mechanisms seem to lead toward the distributed intelligence, working at many levels, from morphemes to phrases, creating, literally, a distributed grammar. It may open way to new types of hardware based neither on linear nor on parallel but on commutation processors [4] that imitate the chemical reaction vessel. In such hardware, an artificial molecular chaos must be maintained. The World Wide Web is a prototype of such a machine. The WWW is a big distributed intelligence, but the remaining problem is how to turn human minds into extremely simple automata with the properties of neurons without the properties of agents.

The examples in **Appendix** illustrate nothing but a vague guess. Its further development, as well as comparison with other linguistic models of acquisition should be better left to those off-beat bees who might become attracted by the chemical smell of strange flowers. The entire direction of Darwinian linguistics, started by Manfred Eigen and continued by Martin Novak may then look like a blooming meadow. But as a chemist, I cannot resist my addiction to chemical smells.