SALT 2: INCREMENTAL EXTRACTION OF GRAMMAR BY SIMPLISTIC RULES

Yuri Tarnopolsky

2005

Abstract

This e-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?" Previously, a scheme of incremental language acquisition, based on very few and simple chemistry-inspired principles, was described on the example of Hungarian folktale *A Só* (Salt). In this e-paper, the principles are further applied to a sequence of acquisition steps. The process does not include any numerical calculations. The elementary acts of analysis and extraction are regarded as binary encounters of quasi-molecules: small linear sequences of "atoms" of language negotiate the outcome of the "collision." A concept of **natural** computing in language evolution and acquisition is discussed.

KEYWORDS: language acquisition, bootstrapping, pattern theory, chemistry, automatic translation, Turing test, Hungarian language, speech generation, semantics, robot-child, robot communication, incremental learning, grammar extraction, language acquisition device.

This e-paper is a further continuation of 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?" For the explanatory and introductory material, as well as the text and translation of the Hungarian tale *Salt*, see SALT [1], where more references could be found. The overall intent can be formulated as application of chemical ideas to subjects outside chemistry: mind, language, society, and technology. While physicists have been doing that with physical ideas for over one hundred years, some chemists are only now slowly and timidly coming to the realization that chemistry might carry its own extra-chemical message.

The purpose of SALT 2 is to see if there is some bread to SALT [1]. The latter is absolutely necessary for understanding SALT 2. In a preliminary fashion, I attempt to test the outcome of SALT experiment for any, however early, promise to be used for linearization of thought structure. The latter is understood in terms of Pattern Theory as a configuration characterized by content, connector, and their quantitative measure [2].

One of the main stimuli of this embryonic work is to develop basic principles that could represent the process of **individual** language acquisition by a **robot-child**, whether realistic or not, in all concrete detail, without being overshadowed by mathematical equations, graphs, numbers, and collective behavior. This is a typically chemical manner of investigation, embodied in the structural chemical equations. Ideally, they represent a detailed sequence of all stable and ephemeral states of the reaction in terms of **individual** atoms and bonds in participating molecules.

A possible application of this framework, if it turns out promising, is robotic communication based on grammar and lexicon **acquired** with minimal assistance and in conditions of **poverty of stimulus**. The **robot-child**, as the model can be called, has to go from infancy to the beginning of maturity when speech is mastered and active intentional learning becomes possible. Infants do not do that by reading *The Wall Street Journal.* Although the term bootstrapping, used in various meanings, is vague, it seems appropriate for the pre-learning mechanism. Its mechanistic and automatic nature resonates well with the term Language Acquisition **Device** (Chomsky).

While loose ends and questions hanging in the air can be clearly seen in this experiment, some excuse is that we all started with baby talk.

In SALT 2 I use larger fragments of input than in SALT , in order to faster accumulate the representation of a larger text, but shorter ones work similarly.

The problems of semantics are considered here least if all, although some initial idea will be put forward: semantics is **possibly** as presentable in triangles as syntax in triplets, i.e., squashed triangles.

I perform some easy operations, like input rewriting and haplology elimination, manually, while more cumbersome ones, like generating comprehensive tables of bonds and categories (CATS), are done with MATLAB codes (can be sent on request).

No explicit numerical calculations are involved.

The MATLAB output needs some simple manipulations with MS Word in order to convert it to tables in document format. Macros can be used.

The following six **steps** of acquisition are described with a diminishing degree of detail. The stressed syllables are capitalized.

STEP 1

INPUT 1

```
P1=char ( 'volt', 'EGY', 'szer', 'egy', 'Ö', 'reg', 'KI', 'rály', 'PAUSE', 'és', 'HÁ', 'rom', 'szép', 'LE', 'ány', 'a', ...
'STOP', 'az', 'Ö', 'reg', 'KI', 'rály', 'SZER','et','te','VOL','na','mind', 'a', 'HÁ','rom', 'LE', 'ány', 'át', 'FÉRJ', ...
'hez', 'AD', 'ni', 'STOP', 'ez', 'nem', 'is', 'lett', 'VOL', 'na', 'NE', 'héz', 'mert', 'HÁ', 'rom', 'OR', 'szág', 'a', ...
 'volt', 'PAUSE', 'mind', 'a', 'HÁ', 'rom', 'LE', 'ány', 'á', 'ra', 'JUT', 'ott', 'EGY', 'egy', 'OR', 'szág', 'STOP');
```

'START' and 'END' are added to P1:

P1=char (**'START'**, 'volt', 'EGY', ……….. ,'OR', 'szág', 'STOP', **'END'**);

output:

1. Generators and triplets

Command: ms, dsgn .

ms (mindset) compiles structure **G** in which every generator enters only once.

dsgn (display generators) displays the triplets.

GENERATOR SPACE 1

P =72, G =44 (72 partly repeating and 44 different generators in input) **.**

The left and right neighbors are preceded by the number of their occurrences.

The occurrences of central generators are in Column 4.

G:	**GENERATOR SPACE 1**			
1	2	3	4	5
No.	LEFT NEIGHBOR	G	No. of entries	RIGHT NEIGHBOR
1	START	START	1	volt;
2	START; a;	volt	2	EGY; PAUSE;
3	ott; volt;	EGY	2	egy; szer;
4	EGY;	szer	1	egy;
5	EGY; szer;	egy	2	OR; Ö;
6	az; egy;	Ö	2	2-reg;
7	2-Ö;	reg	2	2-KI;

8	2-reg;	KI	2	2-rály;
9	2-KI;	rály	2	1-PAUSE; SZER;
10	rály; volt;	PAUSE	2	mind; és;
11	PAUSE;	és	1	HÁ;
12	2-a; mert; és;	HÁ	4	4-rom;
13	4-HÁ;	rom	4	2-LE; OR; szép;
14	rom;	szép	1	LE;
15	2-rom; szép;	LE	3	3-ány;
16	3-LE;	ány	3	a; á; át;
17	2-mind; szág; ány;	a	4	2-HÁ; STOP; volt;
18	a; ni; szág;	STOP	3	END; az; ez;
19	STOP;	az	1	Ö;
20	rály;	SZER	1	et;
21	SZER;	et	1	te;
22	et;	te	1	VOL;
23	lett; te;	VOL	2	2-na;
24	2-VOL;	na	2	NE; mind;
25	PAUSE; na;	mind	2	2-a;
26	ány;	át	1	FÉRJ;
27	át;	FÉRJ	1	hez;
28	FÉRJ;	hez	1	AD;
29	hez;	AD	1	ni;
30	AD;	ni	1	STOP;
31	STOP;	ez	1	nem;
32	ez;	nem	1	is;
33	nem;	is	1	lett;
34	is;	lett	1	VOL;
35	na;	NE	1	héz;
36	NE;	héz	1	mert;
37	héz;	mert	1	HÁ;
38	egy; rom;	OR	2	2-szág;
39	2-OR;	szág	2	STOP; a;
40	ány;	á	1	ra;
41	á;	ra	1	JUT;
42	ra;	JUT	1	ott;
43	JUT;	ott	1	EGY;
44	STOP;	END	1	END

Extraction of CATS and BONDS implements the following simplistic rules:

RULE 1: Adjacency A—B is registered as bond if {A—B } repeats two or more times.

RULE 2: If {A—B, A—C } or {D—A , E—A }, A is a generator.

RULE 3. Haplology is eliminated.

RULE 4. If {A—B, A—C }, then A is a RIGHT CAT with domain {B,C}. If {B—A, C—A }, then A is a LEFT CAT with domain {B,C}.

EXAMPLES:

1. Bonds **a—HÁ** and **mind—a** contain generator **a**, which is encountered also in doublets **ány—a** and **a—volt**. Therefore, they cannot be qualified as very stable blocks, but could remain as background weak bonds.

Doublets **a—HÁ** and **a—volt** form right category (RIGHT CAT) **a—{ HÁ, volt}** ; doublets **mind—a** and **ány—a** form left category (LEFT CAT) **{ mind, ány}—a .**

2. In the following bond sequences the middle doublet is removed to eliminate haplology:

{Ö—reg , reg—KI, KI—rály} → {Ö—reg , KI—rály}

{HÁ—rom , rom—LE , LE—ány} → {HÁ—rom, LE—ány}.

3. **KI** and **rály** in INPUT 1 occur twice and **only as a doublet**. This is why they are qualified as a bond **KI_rály** , see below. The block further becomes a generator.

Under these circumstances **RULE 2** means that **CAT** is a generator. The difference is that **RULE 2** is applicable to **new** input, in which only **B** and **C** are known, while **RULE 4** applies to known generators. More important, **RULE 2** can be applied to levels below syllables. This difference is subtle and both rules can be combined.

2. Bonds

command: cblr (cats, bonds, left, right) ; it extracts bonds and CATs (categories) .

BONDS 1 output		
6 7	1	Ö_reg
7 8	2	reg_KI
8 9	3	KI_rály
12 13	4	HÁ_rom
13 15	5	rom_LE
15 16	6	LE_ány
17 12	7	a_HÁ
23 24	8	VOL_na
25 17	9	mind_a
38 39	10	OR_szág

→

BONDS 1 edited
Ö_reg
KI_rály
HÁ_rom
LE_ány
VOL_na
OR_szág

Haplology is eliminated. BONDS with PAUSE and STOP are ignored in this experiment, although they can be meaningful.

3. CATS (CATs, categories)

LEFT CAT has its domain on the left, **RIGHT CAT** has its domain on the right.

LEFT CATS 1		
	domain	CAT
1	START, a	volt
2	ott, volt	EGY
3	EGY, szer	egy
4	az, egy	Ö
5	rály, volt	PAUSE
6	a, mert, és	HÁ
7	rom, szép	LE
8	mind, szág, ány	a
9	a, ni, szág	STOP
10	lett, te	VOL
11	PAUSE, na	mind
12	egy, rom	OR

RIGHT CATS 1		
	CAT	domain
1	volt	EGY, PAUSE
2	EGY	egy, szer
3	egy	OR, Ö
4	rály	PAUSE, SZER
5	PAUSE	mind, és
6	rom	LE, OR, szép
7	ány	a, á, át
8	a	HÁ, STOP, volt
9	STOP	END, az, ez
10	na	NE, mind
11	szág	STOP, a

CATS with “mute” generators START, END, and STOP (highlighted) are erased in this experiment) if there is only one more generator except the mute one; if more (see line 8 in **RIGHT CATS 1**), the mute one is erased.

LEFT CATS 1 (edited)		
1	ott, volt	EGY
2	EGY, szer	egy
3	az, egy	Ö
4	a, mert, és	HÁ
5	rom, szép	LE
6	mind, szág, ány	a
7	lett, te	VOL
8	egy, rom	OR

RIGHT CATS 1 (edited)		
1	EGY	egy, szer
2	egy	OR, Ö
3	rom	LE, OR, szép
4	ány	a, á, át
5	a	HÁ, , volt
6	na	NE, mind

This concludes the analysis of **OUTPUT** from **INPUT P1**.

To prepare **OUTPUT 1**, for the next input **P2**, **P1** is rewritten (compacted) into **PP1** in accordance with the table of **BONDS.** The final 'END' is removed.

OUTPUT 1, compacted

```
PP1 = char('START', 'volt', 'EGY', 'szer', 'egy', 'Öreg', 'KIrály', 'PAUSE', 'és', 'HÁrom', 'szép', 'LEány', 'a', ...
'STOP', 'az', 'Öreg', 'KIrály',  'SZER', 'et', 'te', 'VOLna', 'mind', 'a', 'HÁrom', 'LEány', 'át', 'FÉRJ', ' hez', ...
'AD', 'ni', 'STOP', 'ez', 'nem', 'is',  'lett', 'VOLna', 'NE', 'héz', 'mert', 'HÁrom', 'ORszág',  'a', 'volt', 'PAUSE', ...
'mind', 'a', 'HÁrom', 'LEány', 'á',  'ra', 'JUT', 'ott', 'EGY', 'egy', 'ORszág',  'STOP' );
```

After any step of acquisition is completed, the subsequent input cannot be perceived on the same terms as the previous one. If some stable BONDS were recorded, the next input is perceived in terms of bonded doublets as generators. This seems to be the major difference between the statistical analysis of a corpus and the autonomic bootstrapping. **In the eyes and ears of robot-child, the world gradually takes meaning**

through discovering its regularity. This process can be visualized as the moving borderline between **the old** and **the new**, as in walking trough darkness with a flashlight.

Strictly speaking, all **BONDS** and **CATS** structures can be remembered, but for the purpose of illustration only the edited ones will be kept in memory and transferred to the next STEP.

Note that addresses in **G** may change from step to step, but addresses in BONDS and CATS always correspond to the current space **G**.

STEP 2

OUTPUT 1 changes the **perception** of the next **INPUT 2** so that the bonded syllables are combined into **words**, whether complete or incomplete. Haplology is eliminated.

INPUT 2

P2=char('HA', 'nem', 'a', 'HOGY', 'an', 'nincs', 'HÁ', 'rom', 'EGY', 'for', 'ma', 'AL','ma', 'PAUSE', 'úgy', 'a', ...
'HÁ', 'rom', 'OR', 'szág', 'sem', 'volt', 'EGY', 'for', 'ma', 'STOP', 'azt', 'MOND', 'ta', 'EGY', 'szer', 'a', 'KI', ...
'rály', 'a', 'LE', 'ány', 'a', 'i', 'nak','hogy', 'AN', 'nak', 'AD', 'ja', 'a', 'LEG', 'szebb', 'OR', 'szág', 'át', 'PAUSE', ...
'A', 'mely', 'ik', 'õt', 'PAUSE', 'LEG', 'job', 'ban', 'SZER', 'e', 'ti', 'STOP');

Compacting the input along BONDS 1:

P2=char('HA', 'nem', 'a', 'HOGY', 'an', 'nincs', 'HÁ', 'rom', 'EGY', 'for', 'ma', 'AL', 'ma', 'PAUSE', 'úgy', 'a', ...
'HÁrom', 'ORszág', 'sem', 'volt', 'EGY', 'for', 'ma', 'STOP', 'azt', 'MOND', 'ta', 'EGY', 'szer', 'a', 'KIrály', 'a', ...
'LEány', 'a', 'i', 'nak', 'hogy', 'AN', 'nak', 'AD', 'ja', 'a', 'LEG', 'szebb', 'ORszág', 'át', 'PAUSE', ...
'A', 'mely', 'ik', 'õt', 'PAUSE', 'LEG', 'job', 'ban', 'SZER', 'e', 'ti', 'STOP');

Next, PP1 and P2 are concatenated: P=strvcat(PP1, P2, 'END');

output:

The complete **G TABLE** is omitted.

1.Bonds 2

BONDS 2 add up to BONDS 1

BONDS 2 output	
1	volt_EGY
2	**EGY_for**
3	EGY_szer
4	Öreg_KIrály
5	HÁrom_LEány
6	HÁrom_ORszág
7	LEány_a
8	a_HÁrom
9	STOP_END
10	mind_a
11	**for_ma**

→

BONDS 2 edited; strong bonds in bold	
1	volt_EGY
2	**EGY_szer**
3	Öreg_KIrály
4	HÁrom_LEány
5	HÁrom_ORszág
6	LEány_a
7	a_HÁrom
8	mind_a
9	**EGY_for_ma**

BONDS 2 illustrate the fluid and provisional character of **BONDS** and the idea of equilibrium.

BONDS 1 include **Ö_reg** and **KI_rály**. There is a persistent tendency for their adjacency, so that until further solidification of bonds there is an equilibrium **soup**:

{Ö , reg , KI , rály , Ö_reg , KI_rály , Ö_reg_KI_rály }

The quantitative aspect is ignored here. The “weights” in the soup correspond to concentrations in chemistry and probabilities in Pattern Theory [2]. The term “weight,” an artifact of the first neural networks, seems very inappropriate because, unlike concentration, probability, and even energy, it is not normalized. The comparison with neural networks, however, is avoided here. The position of the **equilibrium** depends on the topic, context, and previous history. We tend to speak in larger blocks when the subject is familiar and frequent. We might stumble on an unfamiliar terrain.

The concept of equilibrium cannot be applied to the mind as a whole, where equilibrium is continuously shifting because of the aging of memory traces and the influx of new traces. In physical sense, is never achieved in any living system.

Gradually the representation becomes more and more coarse as the bonds and categories solidify and the atomic entries become tagged by their categories. This process of tagging is nothing mysterious: a generator is in equilibrium with all its neighbors and **a category is the neighbor of its entire domain.** Categories overlay **star topology** on **tree topology**.

Distinctively, the **perception also becomes coarser**: from sounds to phonemes to syllables to morphemes to words to expressions.

The idea of local equilibrium that I am trying to convey, not for the first time, but still with a great difficulty, is very simple in chemistry. The best way is just to look into the textbook of general chemistry, although the illustrative material is scattered all over the course. A chemical substance is always in equilibrium with all its possible fragments, down to the atoms, but the concentration of all or absolute majority of fragments or transposition at certain conditions (regarded "normal" in chemistry, physics, and human environment) is practically zero.

> **Example:** Vinegar is acetic acid CH_3COOH in water. It is in equilibrium with its two fragments: $\mathbf{CH_3COOH \rightleftarrows CH_3COO^- + H^+}$ (it is H^+ that tastes sour, whatever its origin). Theoretically, there could be equilibriums along all bonds, for example, $\mathbf{CH_3COOH \rightleftarrows CH_3CO^+ + OH^-}$, which at normal conditions is completely shifted to the left.

Nevertheless, the chemical transformations run through such rare, unstable, and improbable states. Otherwise, everything that could chemically happen with the atoms of our body would happen in an instant.

In terms of practical computation, equilibrium means that most of the memory of your personal computer is inaccessible **at the moment**, which only shows how unnatural computers are in their accessibility. It is hard not to remark that in the digital age the structure of society becomes unnatural if your most intimate social identity tags become accessible.

As a further illustration, recall how difficult it could be sometimes to retrieve a name of a person or location. But if we remember its fragment or any link, for example, that it is something related to horses, the name comes up: Mr. Rein? Mr. Spur? Mr. Bay? Mr. Hay? Mr. Oats, of course! (the idea is borrowed from a short story by Chekhov). This demonstrates the difference between the Hopfield network and human memories, although the former has the ability of a retrieval by fragment.

The memory I have in mind does not have addresses in the sense ROM and RAM have. The address of the natural memory cell is anything in equilibrium with the content of the cell. This can be either less (first letter of the name Oats) **or more** ("horses") than the cell content. I believe neurophysiology has its own view of the problem, but in psychology it has been known since long as association. Note that the behavior of the acetic acid in the above example is called dissociation from left to right and association from right to left.

Bonds **HÁrom—LEány** and **HÁrom—ORszág** dissociate and associate in the same manner as acetic acid. Suppose robot-child with *Salt* as its only life experience hears the word **HÁrom** ("three"). The words **LEány** ("girl") and **ORszág** ("land") will immediately activate in its memory.

2. CATS 2

LEFT CATS 2 (edited)		
1	a, sem	volt
2	ott, rom, ta, volt	**EGY**
3	EGY, szer	egy
4	az, egy	**Öreg**
5	a, Öreg	**KIrály**
6	a, mert, és	HÁrom
7	HÁrom, a, szép	**LEány**
8	KIrály, LEány, ORszág, ja, mind, nem, szer, úgy	a
9	KIrály, ban	SZER
10	lett, te	VOLna
11	LEány, ORszág	**át**
12	hez, nak	**AD**
13	HA, ez	nem
14	HÁrom, egy, szebb	**ORszág**
15	AL, for	ma

16	AN, i	**nak**

RIGH CATS 2 (edited)		
2	**EGY**	egy, for, szer
3	szer	a, egy
4	**egy**	ORszág, Öreg
5	KIrály	SZER, a
7	**HÁrom**	LEány, ORszág, szép
8	**LEány**	a, á, át
9	a	HOGY, HÁrom, KIrály, LEG LEány , i, volt
11	**SZER**	e, et
12	VOLna	NE, mind
14	**AD**	ja, ni
15	nem	a, is
16	ORszág	, a, sem, át
17	nak	AD, hogy
18	**LEG**	job, szebb

The new CATS 2 are in bold type. The previous CATS 1 may still be in memory.

OUTPUT 2 (compacted)

```
PP2=char('START', 'volt', 'EGYszer', 'egy', 'Öreg', 'KIrály', 'PAUSE', 'és', 'HÁrom', ...
 'szép', 'LEány', 'a', 'STOP', 'az', 'Öreg', 'KIrály', 'SZER', 'et', 'te', 'VOLna', 'mind', ...
'a', 'HÁrom', 'LEány', 'át', 'FÉRJ', 'hez', 'AD', 'ni', 'STOP', 'ez', 'nem', 'is', 'lett', ...
 'VOLna', 'NE', 'héz', 'mert', 'HÁrom', 'ORszág', 'a', 'volt', 'PAUSE', 'mind', 'a', ...
'HÁrom', 'LEány', 'á', 'ra', 'JUT', 'ott', 'EGY', 'egy', 'ORszág', 'STOP', 'END', 'HA', ...
'nem', 'a', 'HOGY', 'an', 'nincs', 'HÁ', 'rom', 'EGYforma', 'AL', 'ma', 'PAUSE', ...
'úgy', 'a', 'HÁrom', 'ORszág', 'sem', 'volt', 'EGYforma', 'STOP', 'azt', 'MOND', ...
'ta', 'EGYszer', 'a', 'KIrály', 'a', 'LEány', 'a', 'i', 'nak', 'hogy', 'AN', 'nak', 'AD', 'ja', ...
 'a', 'LEG', 'szebb', 'ORszág', 'át', 'PAUSE', 'A', 'mely', 'ik', 'õt', 'PAUSE', ...
 'LEG', 'job', 'ban', 'SZER', 'e', 'ti', 'STOP' );
```

GRAMMAR 2

Until now we were manipulating syllables in a formal manner, supposedly not knowing what they meant, although I felt a constant pressure of meaning. Now we can try to tentatively interpret what it all means. Translations are given for words and some morphemes/lexemes.

Left or right is indicated by letters L and R: EGY L means LEFT CAT ‘EGY’.

NOTE. This example reminds about the agglutination in Hungarian:
'LEány', 'a', 'i', 'nak' = “girl,” “his/her,” “Plural,” “Dative”;
'LEányainak' =   [give] “to his/hers girls”

Most probably, each of the countless blocks of morphemes in such languages as Hungarian, Finnish, Russian, and Turkish, are acquired by children as a whole. As a speaker of Russian, however, I must note that sometimes the trailing endings need some small but perceptible time to arrange in order before saying. I would not be able to use the following word up to, probably, the age of 7 or 8, and even now I would avoid it by all means:
проигрывающиеся
pro-igr-yva-yu-shchi-e-sya (7 morphemes)
[those that] can be played (playable)   [for example, on DVD player]
or: [those that] are being played now;   igr is the stem (“play”)

Table GRAMMAR 2 is compiled manually from **CATS 2.**

GRAMMAR 2					
	LEFT	**CAT**	**RIGHT**	INTERPRETATION	LABEL
1		**EGY R** “one”	**-egy , -szer** EGYegy, “one” EGYszer “once”	Semantics (“one”)	**EGY R**
2	**-ott, -ta, volt** Past , “was”	**EGY L** “one”		Verb (Past)	**volt**
3	**az, egy**	**Öreg** “old”		Article	**egy**
4	**a, Öreg**	**KIrály** “king”		Adjective	**Öreg**
5		**egy** “a”	**ORszág,** “land” **Öreg** “old”	Noun Single Indefinite	**ORszág**
6	**HÁrom, a, szép** “three”, “the”,	**LEány** “girl”		Noun group (Numeral, Adjective)	**szép**

	“beautiful”				
7		**HÁrom** “three”	**LEány, ORszág szép**	Numeral	**HÁrom**
8		LEány	**-a, -á, -át**	Possessive	**-a**
9	**LEány, ORszág**	**át**		Noun Possessive + Accusative.	**át**
10	**-hez, -nak** “to”, Dative	**AD** “give”		case endings Allative, Dative	**-hez**
11		**AD**	**-ja**, 3rd Person, **-ni** Infinitive	**Verb**	**AD R**
12	**HÁrom, egy, szebb** “three”, “a”, “best”	**ORszág** “land”		Noun group	**ORszág**
13	**Alma** “apple,” **forma** “form”	(ma)		(sound)	
14	**AN, -i-** ANnak “to that one” -i- Plural	**-nak**		Dative	**nak**
15		**STOP**	**az, azt, ez** “the”, “this”	Article, Pronoun, (Sentence start)	**STOP R**
16		**SZER**	**-e, -et** SZEReti “loves” SZERet “loved” 3rd Person	See comments to STEP 4	**SZER R**
17		**LEG**	**-job, -szebb** (LEGszebb “best”)	Superlative	**LEG**

IMPORTANT: Interpretation relates to the domain of CAT, not to the CAT itself.

I hope the **GRAMMAR 2** table speaks for itself, very much like an infant, as it is supposed to. What I see in it is the very beginning of crystallization of **individual** grammar in the **individual** mind of the robot-child who has never heard anything but the story of Salt. If *Salt* is its only experience, semantics is absent.

But how are categories labeled in the mind of an infant? Certainly not by terms of grammar. Of course, the category is just a generator and it is labeled just by its individuality of a set member. But some first words that enter CATS may contribute themselves as internal labels for **patterns of grammar**. This is reasonable in case of LEFT CATS, but for RIGHT CATS the name of the CAT can be taken as label. I wonder if this is because Hungarian is left-branching. Will that be different in Spanish? To speculate, **the first impressions of robot-child imprint large subsequent categories of whether syntax or semantics** (I begin to suspect that their opposition may be as useful but as artificial as the concept of syllable). Could this work for a real child? It is quite probable that the first impressions of the child imprint internal labels for large **natural** categories of light, darkness, comfort, discomfort, hunger, satisfaction, fear, and joy, from which the tree of knowledge grows. Ulf Grenander describes in *Patterns of Thought* [2] the outer branches of the tree.

STEP 3

In subsequent STEPS only compacted new inputs will be shown.

INPUT 3

```
P3=char('FEL', 'elj', 'NEK', 'em', 'PAUSE', 'ÉD', 'es', 'LE', 'ány', 'om', 'PAUSE', 'hogy', 'SZER', 'etsz', ...
'EN', 'gem', 'PAUSE', 'KÉR', 'dez', 'te', 'a', 'LEG', 'i', 'dõ', 'sebb', 'ik', 'et', 'STOP', 'mint', 'a', 'GA', 'lamb', ...
'a', 'TISZ', 'ta', 'BÚZ', 'át', 'PAUSE', 'MOND', 'ta', 'a', 'LE', 'ány', 'STOP', 'hát', 'te', 'PAUSE', 'ÉD', 'es', ...
'LE', 'ány', 'om', 'PAUSE', 'KÉR', 'dez', 'te', 'a', 'KÖZ', 'ép', 'sõt', 'STOP', 'én', 'úgy', 'ÉD', ...
'es', 'AP', 'ám', 'PAUSE', 'mint', 'FOR', 'ró', 'NYÁR', 'ban', 'a', 'SZEL', 'lõt', 'STOP');
```

No compacting is needed

P =183 G =96

BONDS 3 and CATS 3 (Strong bonds are in bold type)

BONDS 3 edited
Öreg-KIrály
HÁrom-LEány
HÁrom-ORszág
LEány-a
LEány-om
a-HÁrom
a-LEG
a-LEány
te-a
mind-a
MOND-ta
ÉD-es
es-LEány
KÉR-dez
dez-te

RIGHT CATS 3	
volt	EGYforma, EGYszer PAUSE
EGYszer	a, egy
egy	ORszág, Öreg
KIrály	PAUSE, SZER, a
PAUSE	A, KÉR, LEG, MOND hogy, mind, mint, ÉD, és, úgy
HÁrom	EGYforma, LEány, ORszág, szép
LEány	STOP, a, om, á, át
a	GA, HOGY, HÁrom, KIrály KÖZ, LEG, LEány, SZEL, TISZ
a	i, volt
STOP	END, FEL, HA, az, azt, ez, hát, mint, én
SZER	e, et, etsz
et	STOP, te
te	PAUSE, VOLna, a
VOLna	NE, mind
át	FÉRJ, PAUSE
AD	ja, ni
nem	a, is
ORszág	STOP, a, sem, át
EGYforma	AL, STOP
úgy	a, ÉD
ta	BÚZ, EGYszer, a
i	dõ, nak
nak	AD, hogy
hogy	AN, SZER
LEG	i, job, szebb
ik	et, õt
ban	SZER, a
es	AP, LEány
mint	FOR, a

LEFT CATS 3	
START, a, sem	volt
ta, volt	EGYszer
EGY, EGYszer	egy
az, egy	Öreg
a, Öreg	KIrály
KIrály, em, gem, ma, om te, volt, ám, át, õt	PAUSE
a, mert, nincs, és	HÁrom
HÁrom, a, es, szép	LEány
EGYszer, KIrály, LEány ORszág,	a
ban, ja, lamb mind, mint, nem, ta, te, úgy	a
EGYforma, LEány, ORszág, a, et, lõt, ni, sõt, ti	STOP
KIrály, ban, hogy	SZER
SZER, ik	et
dez, et, hát	te
lett, te	VOLna
PAUSE, VOLna	mind
BÚZ, LEány, ORszág	át
hez, nak	AD
HA, ez	nem
HÁrom, egy, szebb	ORszág
HÁrom, volt	EGYforma
PAUSE, én	úgy
PAUSE, azt	MOND
MOND, TISZ	ta
LEG, a	i
AN, i	nak
PAUSE, nak	hogy
PAUSE, a	LEG
mely, sebb	ik
NYÁR, job	ban
PAUSE, úgy	ÉD
PAUSE, STOP	mint

Some Hungarian morphemes, such as a and t, are used as markers in several roles across various parts of speech. The red border in the CATS 3 tables encloses an example of how this multifunction can be dealt with. The lines of CAT a, both LEFT and RIGHT, are split depending on the **stress** of the syllables in the domain.

Note that in Hungarian the possessor is unmarked, while the possession is:

A KIrály ORszága , “The King land-his” , “The King’s land,” In LEFT CATS, if a precedes a stressed syllable, a strong hypotheses of robot-child is that a is the definite article. If the next after a syllable is unstressed or the word is monosyllabic, it is a possession mark: ORszága volt , “his/her-land was.”

In RIGHT CATS, if a **noun** is followed by a, a somewhat weak hypothesis can be formed that a marks a possession: LEánya , “his/her-girl.” Otherwise, it can be the definite article.

BONDS 1-3, edited
BONDS 2
MOND_ta
KÉR_dez
ÉD_es
BONDS 2
EGY_for_ma
EGY_szer
volt_EGY
Öreg_KIrály
HÁrom_LEány
HÁrom_ORszág
LEány_a
a_HÁrom
mind_a
BONDS 1
Ö_reg
KI_rály
HÁ_rom
LE_ány
VOL_na
OR_szág

Blocks of words are not strong bonds. I left them in the BONDS 1-3 table to illustrate the semantic and **contextual** significane of longer blocks: **volt_EGYszer** “there was once,” **Öreg—KIrály** “old king,” **HÁrom—LEány** “three girls,” **HÁrom—ORszág** “three lands.” The plural of the noun after a numeral is unmarked in Hungarian.

LEFT CATS 3 (edited)	
ta, volt	EGYszer
KIrály, em, gem, ma, om te, volt, ám, át, õt	**PAUSE**
a, mert, nincs, és	**HÁrom**
HÁrom, a, es, szép	**LEány**
EGYszer, KIrály, LEány ORszág,	**a**
ban, ja, lamb mind, mint, nem, ta, te, úgy	**a**
KIrály, ban, hogy	SZER
SZER, ik	et
dez, et, hát	te ??? TE
lett, te	VOLna
BÚZ, LEány, ORszág	**át**
hez, nak	**AD**
HA, ez	**nem**
HÁrom, egy, szebb	**ORszág**
HÁrom, volt	EGYforma
MOND, TISZ	ta
LEG, a	i
AN, i	**nak**
mely, sebb	**ik**
NYÁR, job	**ban**

RIGHT CATS 3 (edited)	
volt	EGYforma, EGYszer
EGYszer	**a, egy**
egy	**ORszág, Öreg**
HÁrom	**EGYforma, LEány, ORszág, szép**
LEány	-a, -om, -á, -át
a	GA, HOGY, HÁrom, KIrály KÖZ, LEG, LEány, SZEL, TISZ
a	i, volt
SZER	**e, et, etsz**
AD	**ja, ni**
nem	**a, is**
ORszág	a, sem, át
i	dõ, nak
nak	AD, hogy
LEG	i, job, szebb
ik	et, õt
ban	SZER, a
es	AP, LEány
mint	FOR, a

GRAMMAR 3

STEP 3 does not add much to the grammar. The following are some new or expanded old categories:

	LEFT	**CAT**	RIGHT	INTERPRETATION	**Label**
GRAMMAR 2 is here					
18	**-mely-, -sebb-** (Amelyik “which,” LEGsebbik “which is the most beautiful”)	**ik**		Suffix of “adjectivity,” working as an object pronoun	**ik**
19	**HA, ez (HAnem** “however” **ez nem** “this is not”)	**nem** “not, no”		Semantics: negativity	**nem**
20		**SZER**	**e, et, etsz**	See comments to STEP 4	

COMMENTS to STEP 3:

1.Category **{ NYÁR, job } ban** is a wrong hypotheses. **NYÁRban** means ”in the summer” , but in the phonological **LEG-job-ban** “best of all” , the adverb morpheme is an , not ban, “in”. Morphology requires **LEG-jobb-an**. The double b is a stem variation.

> **NOTE**. The struggle of morphology and phonetics raises a lot of dust over the notion of syllable. This problem has been repeatedly addressed in the literature, sometimes in strong words against phonology, but is too technical to discuss here. Regarding Hungarian, I wish to refer to the ingenious solution found, as I suspect, by people at least partly outside linguistics. Instead of syllables, “half-syllables” were used as atoms of speech [3]. Examples: **ta-, a-, te-, le-, ke-** (first half-syllable), **-a, -e, -i, …. -ol, -el, -in, -ek** (second half), etc. 326 half-syllables

describe 95% of general Hungarian texts. It is, essentially, uses haplology as the means of “mutual recognition” of atoms of speech.

2. **PAUSE** in the left CAT 3 follows an unstressed syllable.
3. **MONDta** “said,” and **TISZta** “clean” are together for a wrong reason.

STEP 4

INPUT 4

```
P4=char('no', 'most', 'TÉ', 'ged', 'KÉRdez', 'lek', 'PAUSE', 'FOR', 'dult', 'a', 'LEG', 'kis', 'ebb', 'ik', 'hez', ...
'STOP',  'MOND', 'jad', 'hogy', 'SZER', 'etsz', 'STOP', 'úgy', 'ÉD', 'es', 'AP', 'ám', 'PAUSE', 'a', 'HOGY', ...
'az', 'EM', 'ber', 'ek', 'a', 'sót', 'PAUSE', 'FEL', 'el', 'te', 'a', 'KI', 'csi', 'KIrály', 'kis', 'asz', ...
'szony', 'STOP', 'mit', 'BE', 'szélsz', 'te', 'PAUSE', 'FÖR', 'medt', 'rá', 'a', 'KIrály', 'STOP', 'ki', 'az', ...
'UD', 'var', 'om', 'ból', 'de', 'még', 'az', 'ORszág', 'om', 'ból', 'is', 'PAUSE', 'STOP', 'ne', 'IS', 'lás', 'sa', 'lak', ...
'PAUSE', 'ha', 'csak', 'EN', 'nyi', 're', 'SZER', 'etsz', 'STOP');
```

P =264 G =136

BONDS 4 (unedited)		
5	6	Öreg_KIrály
9	11	HÁrom_LEány
9	32	HÁrom_ORszág
11	12	LEány_a “his girl”
11	69	LEány_om, “my girl”
12	39	a_HOGY, “as”
12	9	a_HÁrom
12	6	a_KIrály
12	54	a_LEG
12	11	a_LEány
15	70	SZER_etsz “you (s.) love”
17	12	te_a
19	12	mind_a
51	15	hogy_SZER
68	11	ÉDes_LEány “dear girl”
69	7	om_PAUSE
69	124	om_ból “out of my”
70	13	etsz_STOP
73	17	KÉRdez_te “asked”
87	88	**AP_ám** “my father”
88	7	ám_PAUSE

The CATS are omitted at this step.

COMMENTS to STEP 4:

1. In STEP 4 BONDS and CATS add new words: APám , “my father,” suffix –bol “out of,” and the personal suffix category of possession -om, “my.” Due to the harmony of vowels, the marking morphemes belong to one of two phonetic series. The other “my” morpheme is -am as in APám . This is where **Rule 2** comes into play: –am and –om form a **phonetic** category -m, “my,” which is **not syllabic**. Harmony of vowels requires a separate non-morphemic and non-syllabic category of vowel type.

2. Line 20 in **GRAMMAR 3** is an evidence of confusion that comes from the fuzziness of syllable. I created the confusion by resisting the temptation to split **SZEReti** into **SZER-et-i**, as the morphology required, because **SZERet** is the stem of verb “to love”. There is no such word or stem as **szer**. There is not enough data for robot-child at this step to form the bond **SZER—et** , for which the rest of the tale will give enough evidence. The total count for the entire tale is 11:

1	**SZER-**et-te	(he) loved	1	**SZER-**et-ik	(he) loves
2	**SZER-**e-ti	(he) loves	1	**SZER-**et-ték	(they) loved
3	**SZER-**etsz	(you) love	3	**SZER-**et-em	(I) love

The syllabic division of **SZER**-e-ti and **SZER-**etsz conflict with morphology. There is a good chance, however, that the initial hypotheses in **STEPS 1 to 4** will be replaced by better founded ones. The morpheme **sz** , which is not exactly syllabic, will be differentiated phonetically.

Language acquisition certainly starts at phonemic level. I have already noted elsewhere that optimality principle of Prince and Smolensky, first developed in phonetics, is very chemically-friendly. Paradoxically, it might be easier to translate speech than text. This is a very intriguing problem, which will be left until better times, however.

3. Note the crystallization (highlighted yellow) of the definite article a and nouns a_HÁrom, “the three,” a_KIrály, “the king,” a_LEány, “the girl,” as well as the standard block mind_a , “each/all” + article, which could well be written as one word. Further, the verb forms become more solid: KÉRdez_te , “(he/she) asked.” This poses a question: how much strength should we attribute to the bonds? The answer is: I have no idea and this is exactly the point of the project. We should make a model and trust robot-child to tackle the problem on its own. Until that, my attribution of bold type to strong bonds is intuitive, which is not much better than arbitrary. Since I am familiar with the meaning of the words, I must be excluded from decision making. I realize that this is a truly heretic idea, but definitely not the only heretic idea in the realm of ideas, some of them later accepted. To reformulate this idea: I am forbidden to be the homunculus for robot-child because I have a mind of my own.

The last squeak of homunculus: “Should we cut **robot-child** into **Rotchild**? “

STEP 5

```
P5=char('HI', 'á', 'ba', 'sírt', 'a', 'KIrály', 'kis', 'asz', 'szony', 'PAUSE', 'HI', 'á', 'ba', 'MA', 'gyar', 'áz', ...
'ta', 'hogy', 'az', 'EM', 'ber', 'ek', 'SZER', 'et', 'ik', 'a', 'sót', 'PAUSE', 'VI', 'lág', 'gá', 'KEL', 'lett', ...
'hogy', 'MENJ', 'en', 'STOP', 'EL', 'in', 'dult', 'a', 'KI', 'csi', 'KIrály', 'kis', 'asz', 'szony', 'SÍR', ...
'va', 'PAUSE', 'és', 'BE', 'ért', 'egy', 'nagy', 'ER', 'dõ', 'be', 'STOP', 'ON', 'nan', 'nem', 'is', 'ment', ...
'TO', 'vább', 'PAUSE', 'ott', 'élt', 'egy', 'DA', 'rab', 'ig', 'EGY', 'ma', 'gá', 'ban', 'STOP');
```

BONDS 5 (unedited)	
5 6	Öreg_KIrály
6 99	**KIrály_kis**
7 73	PAUSE_KÉRdez
7 68	PAUSE_ÉDes
7 8	PAUSE_és
9 11	HÁrom_LEány
9 32	HÁrom_ORszág
11 12	LEány_a
11 69	LEány_om
12 39	a_HOGY
12 9	a_HÁrom
12 110	a_KI
12 6	a_KIrály
12 54	a_LEG
12 11	a_LEány
12 108	a_sót
14 105	**az_EM**
15 16	**SZER_et**
15 70	SZER_etsz
17 7	te_PAUSE
17 12	te_a
19 12	mind_a
20 7	át_PAUSE
26 27	nem_is
33 136	**á_ba**
51 15	hogy_SZER
68 11	ÉDes_LEány

69 7	om_PAUSE
69 123	om_ból
70 13	etsz_STOP
73 17	**KÉRdez_te**
87 7	APám_PAUSE
98 12	dult_a
99 112	**kis_asz**

105 106	**EM_ber**
106 107	**ber_ek**
108 7	sót_PAUSE
110 111	**KI_csi**
111 6	csi_KIrály
112 113	**asz_szony**
135 33	**HI_á**

BONDS 5 edited
KIrály_kis
az_EM
SZER_et
á_ba

KÉRdez_te
kis_asz
EM_ber
ber_ek
KI_csi
asz_szony

The reason why I show the numbers of generators in BONDS tables is to illustrate the detection of haplology. The numbers follow without interruption:

14 **105**	**az_EM**
105 106	**EM_ber**
106 107	**ber_ek**

Next CATS will be shown selectively because their interpretation will involve too much reference to their meaning. This will be difficult to follow without the knowledge of the language.

RIGHT CATS 5 (illustrative selections)		
RIGHT CAT	**Domain**	**Interpretation**
HÁrom “three”	**EGYforma_LEány_ORszág_szép** “equal,” “girl,” “land,” “beautiful”	Noun or Adjective
LEány	_a_om_á_át	Noun suffixes
a “the”	**HÁrom, KIrály, LEány,** “three,” “king,” “girl” (Nominative) **sót** ,”salt (Accusative)”	Noun or numeral
et	**ik, te**	Verb suffixes
AD	**ja, ni**	Verb suffixes
LEG- “the most”	(i), **job,kis, szebb** “good,” “little,” “beautiful”	Superlative
ik	**a, et, hez, õt**	A very confused CAT

ÉDes "dear, sweet"	**_APám_LEány** "my father," "girl"	Noun
KÉRdez	**_lek_te**	Verb suffixes

LEFT CATS 5 (selectively)		
Domain	**LEFT CAT**	**Interpretation**
MONDta, volt "said" "was"	**EGYszer** "once"	Standard block
HÁrom, a, szép, ÉDes "three", "the," "beautiful," "sweet"	**LEány** "girl"	Noun group
LEány, ORszág, -var "girl," "land," "court" (second syllable)	**-om** "my"	Personal Possession

Note the following CAT:

a	**HÁrom, KIrály, LEány, sót**

in which **sót** is the Object Case (Accusative) of **só** "salt." Should I have capitalized monosyllabic nouns?

Anticipating the next shaky steps of the model, I hope that if **só** is going to be mentioned in various cases, the problem will be solved somehow. These are some of the later entries of **só** in the text: 'SÓtlan' , "saltless" ', **'SÓval' ,** "with salt," as well as just **só**. But how exactly it is going to be solved, I don't know. Listening to the sound track, I cannot decide whether it is really stressed, although it seems to be so in **a sót** , "the salt (object)." The article **a** places it in the category of nouns.

My Russian ear is not used to distinguish between a long and a stressed vowel, which again brings us to the obvious idea that linguistics as exact natural science must start with phonemes. Chemistry starts with elements.

BONDS 5 edited		
6 99		**KIrály_kis**
14 105		**az_EM**
15 16		**SZER_et**
33 136		**á_ba**
73 17		**KÉRdez_te**

99 112		**kis_asz**
105 106		**EM_ber**
106 107		**ber_ek**
110 111		**KI_csi**
112 113		**asz_szony**

BONDS 5 add words: **EM_ber_ek**, "people," **KIrály_kis_ as_szony**, "princess" **KÉRdez_te**, "(he/she) asked."

STEP 6

```
P6=char('EGYszer', 'MI', 'kor', 'már', 'egy', 'ESZT', 'en', 'dõ', 'is', 'el', 'telt', 'PAUSE', 'ar', 'ra', 'JÁR', 'ta', ...
'SZOM', 'széd', 'KIrály', 'fi', 'PAUSE', 'és', 'MEG', 'lát', 'ta', 'a', 'KIrálykisasszonyt','STOP', 'MEG', ...
'tet', 'szett', 'a', 'KIrály', 'fi', 'nak', 'a', 'KIrálykisasszony', 'PAUSE', 'mert', 'A', 'kár', 'mi', 'lyen', 'PISZ', ...
'kos', 'volt', 'a', 'RU', 'há', 'ja', 'PAUSE', 'szép', 'volt', 'KÜ', 'lön', 'ös', 'en', 'az', 'AR', 'ca', 'STOP', 'SZÉP', ...
'en', 'MEG', 'fog', 'ta', 'a', 'KEZ', 'ét', 'PAUSE', 'HA', 'za', 'vez', 'et', 'te', 'a', 'PA', 'lot', 'á',  'já',  'ba', 'PAUSE', ...
'és', 'két', 'HET', 'et', 'sem', 'várt', 'PAUSE', 'de', 'még', 'EGY', 'et', 'sem', 'de', 'TAL', 'án', 'még', 'egy', 'ÓR', ...
'át', 'sem', 'PAUSE', 'és' , 'MEG', 'es', 'küd', 'tek', 'STOP', 'END');
```

P =416 G =200

BONDS 6

BONDS 6 edited	
5 6	Öreg, KIrály
6 164	**KIrály-fi (prince)**
8 165	és, MEG
9 11	HÁrom, LEány
9 31	HÁrom, ORszág
11 12	**LEány-a (his girl)**
11 69	**LEány-om (my girl)**
12 38	a, HOGY
12 9	a, HÁrom
12 112	a, KIcsi
12 6	a, KIrály
12 113	a, KIrálykisasszony
12 53	a, LEG
12 11	a, LEány

12 110	a, sót
14 107	az, EM
16 12	te, a
18 12	mind, a
25 26	nem, is
50 61	hogy, SZER
61 70	**SZER, etsz**
68 11	ÉDes, LEány
69 123	om, ból
73 12	KÉRdezte, a
76 45	et, sem
107 108	**EM-ber (man)**
107 108 109	**EM-ber-ek (people)**
112 113	KIcsi, KIrálykisasszony
124 125	de, még

We can see the development of an important generalization (highlighted yellow): the patterns of the definite article **a** and the subsequent noun group. The bond, therefore, can be completely described at the level of the interpreted CATS:

[Definite Article]—[Numeral, Adjective, or Noun]
or, to emphasize the higher level: **Article—Noun**

There are also case and possessive suffixes of a noun (highlighted green) and some stable expressions **mind a** ("all of the…, each of the…") and **nem is** ("not so [bad]")

RIGHT CATS 6 edited; selectively		**Interpretation**
volt	**EGYforma, EGYszer, KÜ, PAUSE, a**	
EGYszer	**MI, a, egy**	
egy	**DA, ESZT, ORszág, ÓR, Öreg**	

KIrály	**PAUSE, STOP, SZERet, a, fi**	
és	**HÁrom, MEG, két**	
HÁrom	**EGYforma, LEány, ORszág, szép**	
szép	**LEány, volt**	
LEány	**a, om, á, át**	**Case and possession markers of nouns**
az (“the”)	**AR, EM, ORszág, UD, Öreg**	**Noun starting with vowel**
SZERet	**ik, te**	**Verb suffixes**
AD	**ja, ni**	**Verb suffixes**
nem	**a, is**	**Noun suffixes**
ORszág	**a, om, (sem), át**	
sem	**(de), volt, várt**	**Verb after negation**
MONDta	**EGYszer, a**	
LEG	**(I), job, kis, szebb**	**Superlative**
SZER	**e, etsz**	**Verb person**
FEL	**el, elj**	
ÉDes	**APám, LEány**	**Noun in addressing**
el	**te, telt**	
MEG (multi-functional prefix)	**es, fog, lát, tet**	**Verb stem after prefix**

LEFT CATS 6 edited; selectively		**Interpretation**
MONDta, volt	**EGYszer**	**Verb (Past)**
az, egy	**Öreg**	**Articles**
a, széd, Öreg	**KIrály**	
a, mert, nincs, és	**HÁrom**	**Predecessors of noun group**
HÁrom, a, szép, ÉDes	**LEány**	**Noun group**
EGYszer, KIrály, KÉRdezte, LEány, MONDta, ORszág, ban, dult, ek, ik, ja, lamb, mind, mint, nak, nem, rá, szett, sírt, ta, te, volt, úgy	**a**	**Various words and endings requiring definite article; nouns among them**
HOGY, en, hogy, ki, még	**az**	**Same as previous**
KIrály, -ek	**SZERet**	**Noun**
SZERet, el, et, hát, szélsz	**te**	**Past tense**
lett, te	**VOLna** “would”	**Verb** **Conditional**
BÚZ, LEány, ORszág, ÓR	**át**	**Object Case/Accus**ative
hez, nak	**AD**	**Dative/Allative before a verb**
HÁrom, az, egy, szebb	**ORszág**	**Noun group**
HÁrom, volt	**EGYforma**	

SZERet, ebb, mely, sebb	**ik**	A mixed-up cat
LEány, ORszág, var	**om**	
JÁR, TISZ, fog, lát, áz	**ta**	Past Tense verb (predominately)
KIcsi, a	**KIrálykisasszony**	

The first of the following two lines from **LEFT CATS 6** contains **HÁrom** ("three") as a CAT, but the second has **HÁrom** in the domain

a, mert, nincs, és	**HÁrom**	**Predecessors of noun group**
HÁrom, a, szép, ÉDes	**LEány**	**Noun group**

Since **LEány** ("girl") is a noun, another high level pattern solidifies:

Predecessor of noun group—Noun group—Noun—CASE/POSSESSION

As far as the extracted GRAMMAR is concerned, it has no use until it comes to speech generation, which could be the subject of the next part. Obviously, to be applied to speech generation, **each generator must be tagged by all CATS.** This seems like an extraordinary requirement to natural memory (computers will take anything). But in chemistry, remarkably, it is not only natural but absolutely necessary for the chemists in order to **talk** to each other about chemical matters.

Each chemical structure can be described as a list of all its "tags," meaningful fragments, individual atoms, and their connections in such a way that the entire structure could be reconstructed from its **linear** description. The grammar of such linear descriptions of non-linear 3D-structures is called chemical nomenclature, and it is indeed a grammar of an artificial language used every day by chemists. More about it in any textbook of organic chemistry and in [4].

EXAMPLE. The chemical name of common aspirin is acetylsalicylic acid, which is a kind of old chemical slang, not quite grammatical. Nevertheless, it tags aspirin as containing benzene ring, and tags it time and time again as an acid, as an ester of acetic acid and a phenol, and as something containing two adjacent appendages to the benzene ring. List only the tags—aspirin's CATS—to a chemist and the reply will be "aspirin."

I do not believe that a simulation of robot-child on simplistic "chemical" principles, using regular consecutive computers, is a gratifying occupation, although it is possible. Nature is inherently parallel, but not in the sense of parallel computing as simultaneous execution of multiple tasks within a single problem. Neither individuals nor governments are good at that. I understand parallelism as translation of **random or partially ordered** collisions into communication. I regard the elementary acts of analysis and extraction as binary encounters of quasi-molecules: small linear sequences of sounds, syllables, morphemes, words, and blocks recognize each other and negotiate the outcome of the "collision:" deal or no deal. The outcome is recorded.

Figure 1 illustrates the formation of two-level CATS by copy eliminations. Segments of different color correspond letters in the **Rules** of input processing. One of two black segments is always eliminated. The other becomes a domain of a CAT. The circles do not need to be either correct or even closed.

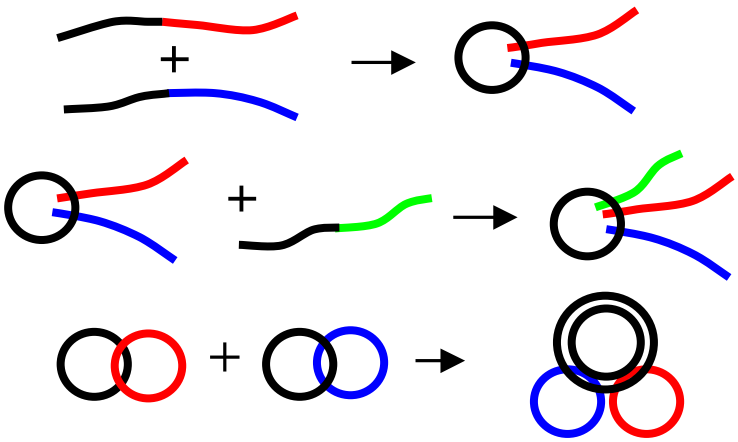

Figure 1: CATS formation in terms of linear segments

As a preview of the next phase of exploration, I will give here only one, rather weak, illustration of a possible principle of speech generation. At this point it makes little sense because utterance starts with thought, but we do not have any settled unified approach to semantics.

Suppose we have the following content with atomic ideas:

{Király Öreg szeret só} = {king old love salt}

We do not know how the generators are connected in the thought. Regardless of the probabilities of the generators and bonds between them (which we may never know), let us retrieve all relevant lines from BONDS and CATS:

2-a;2-Öreg;	**KIrály** "king"	PAUSE; STOP; SZERet; a;
az; egy;	**Öreg** "old"	2-KIrály;
EMberek; KIrály;	**SZERet** **"love"**	ik; (3rd person plural definite) te; (2nd person singular definite)
2-a;	**sót** "salt" (object)	2-PAUSE;
	Öreg_KIrály "old king"	
	a_sót "salt" "definite object"	

Although we do not know what the thought is, we could develop a set of rules for semantics. For example, **old** refers only to **king** and nothing else, but **king**, **salt**, and **love** form a **triangle**: salt is the object of king's desire.

Intuitively, I feel that the semantic relations can be represented by a kind of triangulation in a way similar to the way speech is represented by the squashed triangles of triplets, but I cannot substantiate this idea at this point. I hope to explore this central problem elsewhere.

Figure 2 decorates the black semantic relations of the thought by red linear "comments" of the grammar taken from the above lines.

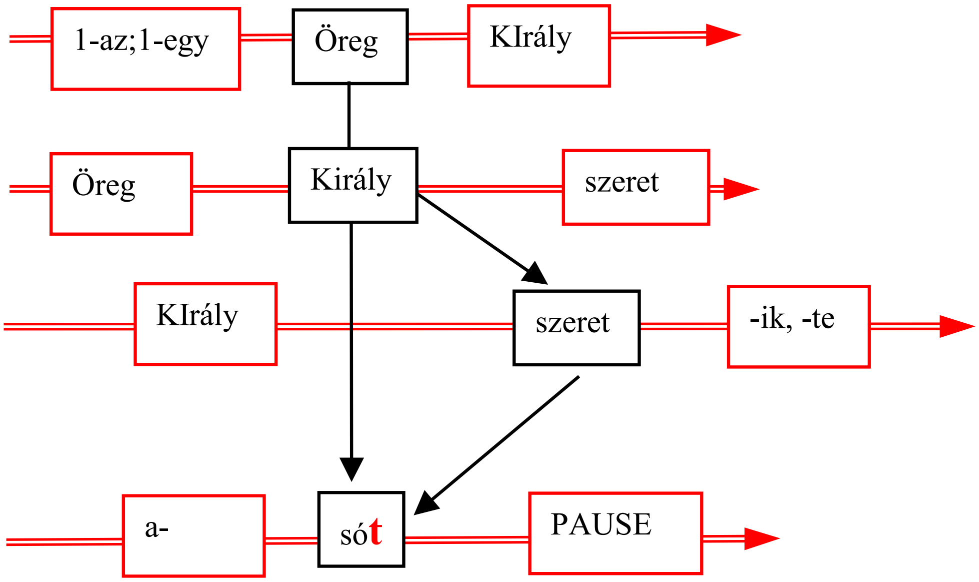

Figure 2. Comments of grammar (red) to semantic relations (black)

What follows from the comments ("catalysts," as a chemist would say), is that the degree of conformity with the grammar is the highest when **Öreg (old), Király (king), szeret (love),** and –**ik** (Present Tense) or **te** (Past Tense) somehow line up in this order. But **a sót** (**salt**, Object Case) has no definite place and can dangle anywhere. And why not if Hungarian has no fixed word order! Of course, **a sót** wedging in between **Öreg** and **Király** would create a tension in the rather strong bond. Otherwise, any position is fine, but PAUSE makes the end position more probable **in the context of *Salt***.

The choice between –**te** and –**ik** is not decisive because of the lack of data. With a considerable stretch of rigor we can attribute the choice of –**te** to the absence of the end **-k** in **KIrály** because the end **-k** is a practically universal marker of plural in both nouns and—except 1st person—verbs.

EMbere**k**; KIrály; "people", "king"	**SZERet** **"love"**	i**k**; (3rd person plural present definite) te; (2nd person singular past definite)

But the real solution must come from a unified approach to grammar and semantics. With all the liberties taken, the final output is:

Az Öreg Király szerette a sót , which seems to me grammatical enough**.**
The old king loved salt.

Note the locality of the mechanism by which I came to the above output. There were never more than three generators in the focus of my attention.

Neither the last example, however, nor all the preceding tables and examples prove anything but the need of something more convincing. They point to the role of multiple repetitions of basic speech patterns during the acquisition. What can be more convincing in our times than a solid computer simulation? The following sideshow discussion addresses this problem.

DISCUSSION

It is already clear, in the very beginning of testing the concept of robot-child, that the work is going to be very cumbersome. No wonder we have big brains. Formally, it may involve the following computational steps:

1. Each syllable or word in the input is compared with the entire generator space **G** and labeled as either old or new. If it is **new**, it is filed into **G .**

2. The memory content is re-analyzed in terms of BONDS and CATS (in advanced stages of acquisition it concerns only a very small part of memory).

3. Generator space **G, BONDS** and **CATS** are updated, so that each generator is re-tagged by all **current** BONDS and CATS it belongs to.

4. The entire memory is updated regarding the age of entries.

5. The next input is rewritten (compacted from phonemes to morphemes) basing on the entire stored grammar. This is the same as to say that the input is **recognized**.

Even this partial description looks like a description of a social system.

Next, how are we going to use that knowledge? Suppose, we want to express a thought, which is a configuration represented by the list of generators (content) and connections (connector graph) between them. Each entry in the content and connector (comprised by a single matrix with a non-trivial diagonal), has a numerical or quasi-numerical (in terms of partially ordered set) measure.

The thought can be a configuration of generators, sometimes incompatible, as in "Is that person over there in a dark overcoat a man or a woman?" Or: "Now you see it; now you don't."

Each recognized generator retrieves its entire **equilibrium cloud of associations** (I am calling psychology for help) from BONDS and CATS. In other words, the configuration of the thought should be "decorated," as in **Figure 2**, by multiple triplet quotations from memory. In the mind of a child the quotations are certainly not in the vocabulary of the English grammar.

NOTE: My insistence on the size of linear neighborhood equal to a triplet is just for the sake of simplicity. In fact, the interactions between generators can be felt at a larger distance, which, by the way, is also a fact of chemistry.

The crucial computation stage is to find the linear arrangement of generators that which has the highest probability, i.e. lowest overall energy/lowest stress/lowest deviation from the grammar. By grammar I mean nothing but the state of the evolving mind of the robot-child, with all its CATS and BONDS (I almost said DOGS: an illustration to the concept of equilibrium).

Omitting the subtleties, all this promises a large volume of boring programming and computation. As a champion of simplicity, I am the least suitable man for this hard labor of bug squashing, number crunching, and computer whipping. If molecules and our brains do it well enough on their own, let us better do it their way.

It is not clear about the brains, but how do the molecules do it?

There is an extremely hard (i.e., time-consuming) computational task known as the protein folding problem [5], in which, theoretically, each bead on a string must be tested in a certain way regarding its interaction with all the other beads. The solution is the configuration in 3-D with minimal overall energy. This task is so unpleasant (the so called NP-problem) because the computation is consecutive while real folding is to a significant degree parallel (sounds almost like Malthus). In other words, folding is natural and fast and our computation is unnatural and takes enormous time.

We are not as fast as molecules, so how can we speak and think at all? My answer is that this is possible because our thoughts are small, our attention span is short, our memory is a far cry from a hard drive, and our knowledge is limited. We can do it because we are imperfect. In fact, protein folding problem becomes solvable if the proteins are short. But the language is so big! Right, but the protolanguage was not. The children do not speak as layers write and even presidents can speak as children.

Let us take a break, anyway.

As a non-linguist, I can afford some unwarranted leaps of imagination. Thus, knowing nothing about the topic, I can derive the peculiar Hungarian possession marker from some ancient form with two articles (or, more probably, none at all), a király a ország, "the king the land." The second article moves to the end: a király országa .

Further, I can derive the Hebrew possessive form from the same ancient pattern:

סוס האיש , **sus ha-ish , **horse the man, man's horse. The first article drops off.

I can also throw in the English variation on the theme: **the king cobra** and all the other noun modifiers, although it is not a possessive form.

A more corpulent form remains in German: **Das Pferd des Mannes,** The horse of-the man-his, or the horse the man-his, the horse of the man.

But Hungarian catches up and even overtakes in another, reversed-German style, formal possessive construct:

Az embernek a lova , the man-him the horse-his , the man's horse.

This is, probably, too much and modern Hungarian requires only whichever one article: Peter **Peter lova** (Peter's horse) or **az ember lova** (the man's horse)

There is no article in Russian, but the marker is in place and you can have it both ways:

Лошадь Пети , horse Pete-his.

Or: **Петина лошадь**, Pete-his horse, Pete's horse (note different word order)

Furthermore, I can fantasize that the fixed stress in Hungarian compensates for the wide use of the scarce possession and tense marking morphemes. The trade-off is the free word order.

In English, the marker morphemes are very scarce and the word order is not free. In Russian, with a wide variety of markers, both the word order and the stress are free.

But the simple fact that Polish, very similar to Russian and with an equally rich choice of markers, has a fixed word order, wakes me up from my sweet dreams.

This awakening gives me an opportunity to explain once more my position. I am definitely not a linguist. I present neither a theory nor a working model. It is just an abstract idea, not yet completely clear to myself, a concept that should be tested by professionals, although it originates outside linguistics from a higher abstract ground of Pattern Theory. As a chemist, however, I feel a certain hopelessness about the current algorithmic and numerical methods of computational linguistics. They take a lot of work but prove anything but the intelligence and inventiveness of the authors. The computer models do not converge to a consensus **unless you can test them** as if they were weather forecasts or at least neural networks.

I suspect that the failure or, let us hope, a long delay in developing automatic translation follows from the very idea of algorithm, The bootstrapping mind of an infant, unlike the mind of an adult, does not use anything like either statistical inference or algorithms, **although results could be the same**.

I acknowledge that I am not familiar with the theories of computability and complexity. But I am aware that the tacit prerequisite of computer science is that **almost** anything of practical importance can be computed within the current symbolic-consecutive paradigm, the success of which is tremendous and proven. But is the success absolute? And is it a success when it takes a lot of time and comes too late? It seems to me that the problems of automatic translation or robotic communication could be the true test for some new concepts of intelligence, all the more, they are pretty closely related to the Turing test of intelligence based on **verbal communication.** This cannot be said about playing chess. To pass the strong Turing test, the computer must express itself without help.

Trying to formulate my idea in the most succinct way, I present it like this: we could possibly create realistic models of language evolution and language acquisition if we were daring enough to change horses in the middle of the stream and switch to a new type of **natural** computers working on ordered chaos, homeostasis, and competition for energy. I presented some vague and possibly not new ideas about such pattern computers in [5]. **Automatic translation with acquired grammar and lexicon** could be a possible application and a stimulus for acquiring (what's the heck), chemical and pattern-theoretical idea by young linguists.

I can reformulate the concept in the form of an answer to the question "what is natural?" **Natural is what has infancy.**

Being unable to calculate 2 x 2, such computers could be capable of computing the behavior and communication between children of pre-school age, who would be ready to start learning math and physics (not sure about chemistry), foreign languages, use algorithms, and operate PC.

I am glad I will not live in such a world, but evolution does not ask for anybody's consent. Probably we already have no choice.

MINIMAL REFERENCES

spirospero.net

1 Yuri Tarnopolsky. 2005. Salt: The Incremental Chemistry of Language Acquisition

http://spirospero.neti/Salt.pdf

See also: http://spirospero.net/complexity.htm

TIKKI TIKKI TEMBO: The Chemistry of Protolanguage.

http://spirospero.neti/Nean.pdf

Pattern Theory and "Poverty of Stimulus" Argument in Linguistics.

http://spirospero.net/Poverty of stimulus.pdf

The Three Little Pigs : Chemistry of Language Acquisition.

http://spirospero.net/3LP.pdf

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

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

———. *Patterns of Thought*. www.dam.brown.edu/ptg/REPORTS/mind.pdf

(watch for updates; see also: www.dam.brown.edu/ptg/**publications**.shtml)

3. Vicsi, Klara**,** *et al.* 2004. Hungarian Speech Databases, BABEL — Multi–Language Database, Project No. 1304, http://alpha.ttt.bme.hu/speech/hdbbabel.php

4. Yuri Tarnopolsky. 2003. *Molecules and Thoughts: Pattern Complexity and Evolution in Chemical Systems and the Mind* , 2003.

http://www.dam.brown.edu/ptg/REPORTS/MINDSCALE.pdf

or: http://spirospero.net/MINDSCALE.pdf

5. ———. 2005. Molecular computation: a chemist's view.

http://users.ids.net/~yuri/PTutor.pdf

EMAIL: http://spirospero.net/email.html Last updated March 22, 2009

What if the words were atoms?