A MATHEMATICAL LINGUISTIC APPROACH TO REBUS

Florentin Smarandache
University of New Mexico
200 College Road
Gallup, NM 87301, USA
E-mail: smarand@unm.edu

INTRODUCTION

The aim of this paper is the investigation of some combinatorial aspects of written language, within the framework determined by the well-known game of crossword puzzles. Various types of probabilistic regularities appearing in such puzzles reveal some hidden, not well-known restrictions operating in the field of natural languages. Most of the restrictions of this type are similar in each natural language. Our direct concern will be the Romanian language.

Our research may have some relevance for the phono-statistics of Romanian. The distribution of phonemes and letters is established for a corpus of a deviant morphological structure with respect to the standard language. Another aspect of our research may be related to the so-called tabular reading in poetry. The correlation horizontal-vertical considered in the first part of the paper offers some suggestions concerning a bi-dimensional investigation of the poetic sing.

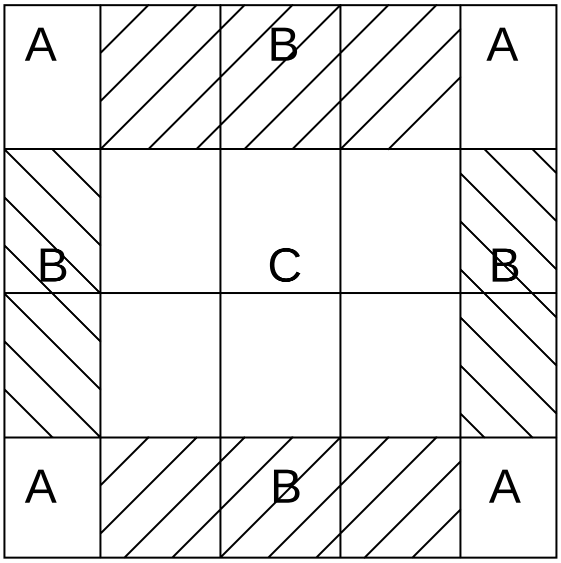
Our investigation is concerned with the Romanian crossword puzzles published in [4]. Various concepts concerning crossword puzzles are borrowed from N. Andrei [3]. Mathematical linguistic concepts are borrowed from S. Marcus [1], and S. Marcus, E. Nicolau, S. Stati [2].

SECTION 1. THE GRID

§1. MATHEMATICAL RESEARCHES ON GRIDS

It is known that a word in a grid is limited on the left and right side either by a black point or by a grid final border.

We will take into account the words consisting of one letter (though they are not clued in the Rebus), and those of two (even they have no sense (e.g. N T, RU,…)), three or more letters – even they represent that category of rare words (foreign localities, rivers, etc., abbreviations, etc., which are not found in the Romanian Language Dictionary (see [3], pp. 82-307 (“Rebus glossary”)).



The grids have both across and down words.

We divide the grid into 3 zones:

a) the four peaks of the grid (zone A)
b) grid border (without de four peaks) (zone B)
c) grid middle zone (zone C)

We assume that the grid has n lines, m columns, and p black points.

Then:

Proposition 1. The words overall number (across and down) of the grid is equal to $n + m + pNB + 2 \cdot pNC$, where

pNB = black points number in zone B,

pNC = black points number in zone C.

Proof: We consider initially the grid without any black points. Then it has $n + m$ words.

- If we put a black point in zone A, the words number is the same. (So it does not matter how many black points are found in zone A).

- If we put a black point in zone B, e.g. on line 1 and column j, $i < j < m$, words number increases with one unit (because on line 1, two words were formed (before there was only one), and on column j one word rests, too). The case is analog if we put a black point on column 1 and line i, $1 < i < n$ (the grid may be reversed: the horizontal line becomes the vertical line and vice versa). Then, for each point in zone B a word is added to the grid words overall number.

- If we put a black point in zone C, let us say i, $1 < i < n$, and column j, $1 < j < m$, then the words number increases by two: both on line i and column j two words appear now, different from the previous case, when only one word was there on each line. Thus, for each black point in zone C, two words are added at the grid words overall number. From this proof results:

Corollary 1. Minimum number of words of grid $n \times m$ is $n+m$. Actually, this statement is achieved when we do not have any black points in zones B and C.

Corollary 2. Maximum number of words of a grid $n \times m$ having p black points is $n+m+2p$ and it is achieved when all p black points are found in zone C.

Corollary 3. A grid $n \times m$ having p black points will have a minimum number of words when we fix first the black points in zone A, then in zone B (alternatively – because it is not allowed to have two or more black points juxtaposed), and the rest in zone C.

Proposition 2. The difference between the number of words on the horizontal and on the vertical of a grid $n \times m$ is $n - m + pNBO - pNBV$, where

$pNBO$ = black points number in zone BO,

$pNBV$ = black points number in zone BV.

We divide zone B into two parts:

- zone BO = B zone horizontal part (line 1 and n)
- zone BV = B zone vertical part (line 1 and m).

The proof of this proposition follows the previous one and uses its results.

If we do not have any black points in the grid, the difference between the words on the horizontal and those on the vertical line is $n-m$.

- If we have a black point in zone A, the difference does not change. The same for zone C.

If we have a black point in zone BO, then the difference will be $n-m-1$. From this proposition 2 results:

Proposition 3. A grid $n \times m$ has $n + pNBO + pNC$ words on the horizontal and $m + pNBV + pNC$ words on the vertical.

The first solving method uses the results of propositions 1 and 2.

The second method straightly calculates from propositions 1 and 2 the across and down words number (their sum (proposition 1) and difference (proposition 2) are known).

Proposition 4. Words mean length (=letters number) of a grid $n \times m$ with p black points is $\geq \dfrac{2(nm-p)}{n+m+2p}$.

Actually, the maximum words number is $n+m+2p$, the letter number is $nm-p$, and each letter is included in two words: one across and another down. One grid is the more crossed, the smaller the number of the words consisting of one or two letters and of black points (assuming that it meets the other known restrictions). Because in the Romanian grids the black points percentage is max.

15% out of the total (rounding off the value at the closer integer – e.g. 15% with a grid 13x13 equals 25.35 ≈ 25; with a grid 12x12 is 21.6 ≈ 22), so for the previous properties, for grids $n \times m$ with p black points we replace p by $\left[\dfrac{3}{20}\right]nm$, where $[x] = \max\{\alpha \in \mathrm{N},\ |\alpha - x| \leq 0.5\}$.

§2. STATISTIC RESEARCHES ON GRIDS

In [1] we find the notion "écart of a sound x", denoted by $\alpha(x)$, which equals the difference between the rank of x in Romanian and the rank of x in the analyzed text.

We will extend this notion to the notion of *a text écart* which will be denoted by: $\alpha(t)$, and

$$\alpha(t) = \frac{1}{n}\sum_{i=1}^{n}\left|\alpha(A_i)\right|$$

where $\alpha(A_i)$ is A_i sound écart (in [1]) and n represents distinct sounds number in text t. (If there are letters in the alphabet, which are not found in the analyzed text, these will be written in the frequency table giving them the biggest order.)

Proposition 1. We have a double inequality:

$0 \le \alpha(t) \le \frac{n-1}{2} + \frac{1}{n}\left[\frac{n}{2}\right]$ where $[y]$ represents the whole part of real number y.

Actually, the first inequality is evident.

Let $\Phi = \begin{pmatrix} 1 & 2 & \dots & n \\ \\ j_1 & j_2 & \cdots & j_n \end{pmatrix}$. Then $\sum_{i=1}^{n}\left|\alpha(A_i)\right| = \sum_{i=1}^{n}\left|i - j_i\right|$

This permutation constitutes a mathematical pattern of the two frequency tables of sounds; in Romanian (the first line), in text t (the second line).

For permutation $\psi = \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ n & n-1 & \dots & 2 & 1 \end{pmatrix}$ we have

$$\sum_{i=1}^{n}\left|i - j_i\right| = 2[(n-1)+(n-3)+(n-5)+\dots] = 2\sum_{k=1}^{\left[\frac{n}{2}\right]}(n-2k+1) =$$

$$= 2\left[\frac{n}{2}\right]\left(n - \left[\frac{n}{2}\right]\right) = \frac{n(n-1)}{2} + \left[\frac{n}{2}\right],$$

where $\alpha(t) = \frac{n-1}{2} + \frac{1}{n}\cdot\left[\frac{n}{2}\right]$.

By induction with respect to $n \ge 2$, we prove now the sum $S = \sum_{i=1}^{n}\left|i - j_i\right|$ has max. value for permutation ψ.

For $n = 2$ and 3 it is easily checked directly. Let us suppose the assertion true for values $< n+2$. Let us show for $n+2$:

$$\psi = \begin{pmatrix} 1 & 2 & \dots & n+1 & n+2 \\ n+2 & n+1 & \dots & 2 & 1 \end{pmatrix}$$

Removing the first and last column, we obtain:

$$\psi' = \begin{pmatrix} 2 & \dots & n+1 \\ n+1 & \dots & 2 \end{pmatrix},$$

which is a permutation of n elements and for which S will have the same value as for permutation

$$\psi'' = \begin{pmatrix} 1 & ... & n \\ n & ... & 1 \end{pmatrix},$$

i.e. max. value (ψ'' was obtained from ψ' by diminishing each element by one).

The permutation of 2 elements $\eta = \begin{pmatrix} 1 & n+2 \\ n+2 & 1 \end{pmatrix}$ gives maximum value for S. But ψ is obtained from ψ' and η;

$$\psi(i) = \begin{cases} \psi'(i), & \text{if } i \notin \{1, n+2\} \\ \eta(i), & \text{otherwise} \end{cases}$$

Remark : The bigger one text écart, the bigger the “angle of deviation” from the usual language.

It would be interesting to calculate, for example, the écart of a poem.

Then the notion of écart could be extended even more:

a) *the écart of a word* being equal to the difference between word order in language and word order in the text;
b) *the écart of a text (ref. words):*

$$\alpha_c(t) = \frac{1}{n}\sum_{i=1}^{n} |\alpha_c(a_i)|,$$

where $\alpha_c(a_i)$ is word a_i écart, and n - distinct words number in the text t.

*

We give below some rebus statistic data. By examining 150 grids [4] we obtain the following results:

Occurrence frequency of words in the grid, depending on their length (in letters)

Letter order	Letter	Letter occurrence mean percentage	Vowels mean percentage	Consonants mean percentage
1	A	15.741%	47.462%	52.538%
2	I	12.849%		
3	T	9.731%		
4	R	9.411%		
5	E	8.981%		
6	O	5.537%		
7	N	5.053%		
8	U	4.354%		
9	S	4.352%		
10	C	4.249%		
11	L	4.248%		
12	M	4.010%		
13	P	3.689%		
14	D	1.723%		
15	B	1.344%		
16	G	1.290%		
17	F	0.860%		
18	V	0.806%		
19	Z	0.752%		
20	H	0.537%		
21	X	0.430%		
22	J	0.053%		
23	K	0.000%		

It is easy to see that a percentage of 49,035% consists of the words formed only of 1, 2 or 3 letters; - of course, there are lots of incomplete words.

*

The study of 50 grids resulted in:

Occurrence frequency of words in a grid (see next page).

It is noticed that vowels percentage in the grid (47.462%) exceeds the vowels percentage in language (42.7%).

So, we can generalize the following:

Statistical proposition (1): In a grid, the vowels number tends to be almost equal to 47.5% of the total number of the letters.

Here is some evidence: one word with n syllables has at least n vowels (in Romanian there is no syllable without vowel (see [2]).

The vowels percentage in Romanian is 42.7%; because a grid is assumed to form words across and down, the vowels number will increase. Also, the last two lines and columns are endings of other words in the grid; thus they will usually have more vowels. When black points number decreases, vowels number will increase (in order to have an easier crossing, you need either more black points or more vowels) (A vowel has a bigger probability to enter in the contents of a word than a consonant.)

Especially in “record grids” (see [3], pp. 33-48) the vowels and consonants alternation is noticed. Another criterion for estimating the grid value is the bigger deviation from this “statistical law” (the exception confirms the rule!): i.e. the smaller the vowel percentage in a grid, the bigger its value.

Statistical proposition (2): Generally, the horizontal words number 73 equals the vertical one.

Here is the following evidence: 100 classical grids were experimentally analyzed, in [4], getting the percentage of 49.932% horizontal words. Usually, the classical grids are square clues, the difference between the horizontal and vertical words being (see Proposition 2):

$$n - m + pNBO - pNBV = pNBO - pNBV\ .$$

The difference between the black points number in zone BO and zone BV can not be too big ($\pm 1,\ \pm 2$ and rarely ± 3). (Usually, there are not many black points in zone B, because it is not economical in crossing (see proof of Proposition 1)).

Taking from [1] the following letters frequency in language:

1. E
2. I
3.A
4.R
5.N
6.T
7.U
8.C
9. L
10. S
11. O
12. D
13. P
14. M
15. B
16. V
17. G
18. F
19. Z
20. H
21. J
22. X
23. K

(because in the grid Ă, Â, Î, Ş, Ţ: are replaced by A: I: S: T, respectively, in the above order they were cancelled) the écart of the 150 grids becomes

$$\alpha(g) = \frac{1}{23}\sum_{i=1}^{23}\left|\alpha(A_i)\right| \approx 1.391\,;$$

the entropy is:

$$H_1 = -\frac{1}{log_{10}2}\sum_{i=1}^{23} p_i\ log_{10}p_i \approx 3.865$$

and the informational energy (after O. Onicescu) is:

$$E(g) = \sum_{i=1}^{23} p_i^2 \approx 0.084$$

Examining 50 grids we obtain:

Words frequency in a grid with respect to the syllables number

Mean percentage of occurrence of a word in a grid								Mean length of a word in syllables
1 syllable	2	3	4	5	6	7	8	
35.588%	26.920%	21.765%	9.551%	5.294%	0.882%	0.000%	0.000%	2.246

(in the category of the one syllable-words, the word of one, two or, three letters, without any sense – rare words – were also considered.) One can see that the percentage of words consisting of one and two syllables is 65.508% (high enough).

Another statistics (of 50 grids), concerning the predominant parts of speech in a grid has established the following first three places:

1. nouns 45.441%
2. verbs 6.029%
3. adjectives 2.352%

Notice the large number of nouns.

*

SECTION II. REBUS CLUES

§1. STATISTICAL RESEARCHES ON REBUS CLUES

Studying the clues of 100 "clues grids", the following statistical data resulted:

Rebus clues frequency according to their length (words number)

(see the next page)

It is noticed that the predominant clues are formed of 2, 3, or 4 words. For results obtained by investigating 100 "clues grids", see the next page.

It is worth mentioning that vowels percentage (46.467%) from rebus clues exceeds vowels percentage in the language (42.7%).

By calculating the clues écart (in accordance with the previous formula) it results:

$$\alpha(dr) = \frac{1}{27}\sum_{i=1}^{27}\left|\alpha(A_i)\right| \approx 1.185$$

(sound frequency used by Solomon Marcus in [1] was used here), the entropy (Shannon) is:

$$H_1 = -\frac{1}{log_{10} 2} \sum_{i=1}^{27} p_i \ log_{10} p_i \approx 4.226$$

and informational energy (O. Onicescu) is:

$$E(dr) = \sum_{i=1}^{27} p_i^2 \approx 0.062 .$$

(The calculations were done by means of a pocket calculator).

Letters occurrence frequency in the rebus clues

Letter order	Letter	Mean percentage of letter occurrence in clues	Vowels percentage	Consonants mean percentage	Letters no. (mean) necessary to clue a grid	Mean length of a word (in letters) used in clues
1	E	10.996%	46.679%	53.321%	657.342	4.374
2	I	9.778%				
3	A	9.266%				
4	R	7.818%				
5	U	6.267%				
6	N	6.067%				
7	T	5.611%				
8	C	5.374%				
9	L	4.920%				
10	O	4.579%				
11	P	4.027%				
12	Ă	3.992%				
13	S	3.831%				
14	Î	3.309%				
15	D	3.079%				
16	Â	1.801%				
17	V	1.527%				
18	F	1.449%				
19	Ş	1.360%				
20	Ţ	1.338%				
21	G	1.330%				
22	B	1.238%				
23	H	0.532%				
24	J	0.358%				
25	Z	0.092%				
26	X	0.037%				
27	K	0.024%				

REFERENCES

[1] Marcus, Solomon – "Poetica matematica" – Ed. Academiei, Bucureşti, 1970 (German translation, Athenäum, Frankfurt am Mein, 1973).

[2] Marcus, Solomon, Edmond Nicolau, S. Stati – Introducere in lingvistica matematica, Bucureşti, 1966 (Italian translation, Patron, Bologna, 1971; Spanish translation, Teide, Barcelona, 1978).

[3] Andrei, Dr. N. – Indreptar rebusist, Ed. Sport-Turism, Bucureşti, 1981.

[4] "Rebus" magazine collection, Bucureşti, 1979-1982.

The Craiova University
Natural Sciences Department

[Published in "Review Roumaine de linguistique", Tome XXVIII, 1983, "Cahiers de linguistique théorique et appliquée", Bucharest, Tome XX, 1983, No. 1, pp. 67-76.]