How to count words?

Dmitri Piontkovski, Maxim Prasolov, Grigory Rybnikov

1 Main problem

Problem 1. *The language of Winnie-Pooh tribe has 100 words. All possible combinations of these words, in any order, are used as sentences of the language. The are two magic spells, "Earth stands on Great Crocodile" and "Every evening Crocodile swallows Sun", that cause tornado. That is why it is not allowed to pronounce sentences that contain the above sequences of words*[1]*. How many sentences of 20 words in this language are allowed?*

Problem 2. *A computer uses* 256 *commands. There is a sequence of four commands that breaks the computer. The programmers made all possible programs of* 7 *commands. Find the percentage of the programs that do not break the computer.*[2]

Problem 3 (Main Problem)**.** *The alphabet of a language L consists of N letters. Several words $v_1, \dots, v_k$ are called* forbidden *and are not used in the language. A word (that is, a finite sequence of letters) is called* admissible *if no part of it is a forbidden word. Find the number of admissible words of n letters in L.*

Problem 4. *Show that the Problems 1 and 2 are special cases of Problem 3.*

2 How to write down the answer?

Choose an alphabet A of N letters (for example, if $A = (a, b, c, \dots, z)$, then $N = 26$). By a *word* we will mean an arbitrary finite sequence of letters of the alphabet A. A part of a word is called its *subword.*

We assume that every language L has exactly one word of zero length, that is, an *empty* word.

We assume that distinct forbidden words are not subwords of each other. We also assume that each forbidden word has at least two letters, that is, the empty word and one-letter words are admissible. Recall that the set of forbidden words is finite.

Problem 5. *The* free language F_A *over the alphabet A is the language with no forbidden words. Prove that the number of the words of n letters in this language is equal to N^n.*

Problem 6. *Let B be the language whose forbidden words are all two-letter words with different letters. Prove that the number of admissible words of n letters in the language B is equal to N for any positive integer n.*

Let M be an arbitrary set of words. Let us denote by m_n the number of n-letter words in this set. The infinite sum

$$M(x) = m_0 + m_1 x + m_2 x^2 + m_3 x^3 + \dots$$

is called the *dimension series* of the set M. The infinite sums of such type (with arbitrary numbers as coefficients m_n) will be briefly referred to as *series* (their complete name, which will *not* be used here, is *formal power series*).

For any language L, by its dimension series $L(x)$ we will mean the dimension series of the set of admissible words. For example, for the free language F_A its dimension series is the geometric series $F_A(x) = 1 + Nx + N^2x^2 + N^3x^3 + \dots$, and for the language B above we have $B(x) = 1 + Nx + Nx^2 + Nx^3 + \dots$

Problem 7. *Write down the dimension series for the language over the alphabet $\{a, b\}$ with forbidden words aa and bb.*

3 The arithmetics of languages

If a set M contains finitely many words, then its dimension series is a polynomial in the variable x. For infinite sets, their dimension series are infinite as well, but they allow various arithmetic operations similar to the operations over the polynomials, that is, addition, subtraction, multiplication by each other and by numbers, and even sometimes division.

In the definitions and problems of this section, $S = s_0 + s_1 x + s_2 x^2 + \dots$ and $R = r_0 + r_1 x + r_2 x^2 + \dots$ are two series, and L_1 and L_2 are two languages over alphabets A_1 and A_2 without common letters. We will assume that the alphabet A_1 consists of upper-case letters while the alphabet A_2 consists of lower-case ones. Let the alphabet A be the union of the alphabets A_1 and A_2, that is, A contains both upper-case and lower-case letters.

[1] Even if the words are in other forms

[2] Similar story happened in 1990s with the first version of *Pentium* microprocessor.

Definition 1. a) The *sum* of two series R and S is the series

$$R + S = (s_0 + r_0) + (s_1 + r_1)x + (s_2 + r_2)x^2 + \dots$$

b) The *sum* of two languages L_1 and L_2 is the language $L_1 + L_2$ over A whose set of admissible words is the union of the sets of admissible words of the languages L_1 and L_2.

Problem 8. *Define the language $L_1 + L_2$ by a finite set of forbidden words.*

Problem 9. *Prove that if $L = L_1 + L_2$, then*

$$L(x) = L_1(x) + L_2(x) - 1.$$

The product of two series is defined by the same way as the product of two polynomials.

Definition 2. *The product of a series R by a monomial ax^n* is the series

$$R \cdot ax^n = ar_0x^n + ar_1x^{n+1} + ar_2x^2x^{n+2} + \dots$$

The product of two series R and S is the sum

$$R \cdot S = R \cdot s_0 + R \cdot s_1x + R \cdot s_2x^2 + \dots$$

Note that this infinite sum of series is well-defined because the coefficient of every power of x is a finite sum of numbers.

Problem 10. *Prove that*

$$(1 - x) \cdot (1 + x + x^2 + \dots) = 1.$$

Definition 3. *The product of two sets of words M* and N is the set MN of all words of the form mn, where m is a word in M and n is a word in N.

The product of two languages L_1 and L_2 is the language $L_1 \cdot L_2$ over A whose set of admissible words is the product of the sets of admissible words of the languages L_1 and L_2.

Problem 11. *Define the language $L_1 \cdot L_2$ by a finite set of forbidden words.*

Problem 12. *Prove that*

$$L(x) = L_1(x) \cdot L_2(x).$$

The division of series has no version for languages, but it helps to write down their dimension series in a compact form. It is defined by a formula similar to the formula for the sum of an infinite geometric progression.

Definition 4. Suppose that a series S begins with the unit, that is, $s_0 = 1$, and $S = 1+\overline{S}$, where $\overline{S} = s_1x+s_2x^2+\dots$ Then its *inverse* is the series

$$\frac{1}{S} = 1 - \overline{S} + \overline{S}^2 - \overline{S}^3 + \dots$$

The *quotient* of two series R and S is the series

$$\frac{R}{S} = R - R \cdot \overline{S} + R \cdot \overline{S}^2 - R \cdot \overline{S}^3 + \dots$$

In general, the quotient of two dimension series can not be obtained as the dimension series for a language. For example, some of the coefficients of the quotient can be negative.

Problem 13. *a) Prove that*

$$S \cdot \frac{R}{S} = R.$$

b) Prove that if $S \cdot T = R$, where the series S begins with the unit, then $T = \frac{R}{S}$.

The use of the division of two series is that it helps to represent many infinite series by a finite formula, that is, a quotient of two polynomials.

Problem 14. *a) Prove that*

$$F_A(x) = \frac{1}{1 - Nx}.$$

b) Represent the dimension series from Problems 6 and 7 as a quotient of two polynomials.

Problem 15. *Prove that the dimension series of any language can be represented as a quotient of two polynomials.*

Thus, the answer to our Main Problem should be represented as a quotient of two polynomials.

4 Free word

Problem 16. *Let L be a language over the Latin alphabet with only one forbidden word "mouse". Find $L(x)$.*

Definition 5. Let a and b be words such that no one is a subword of each other. A nonempty word c is called an *overlap* of a and b if it is a beginning subword of a and, in the same time, the final subword of b (for example, the word "all" is an overlap of the words "ball" and "allow").

A word w is called *free* if it has no overlaps with itself except for the whole word w (e.g., the word "free" is free but the word "underground" is not).

Problem 17. *Suppose that in a language L over an alphabet A of N letters there is a single forbidden word, which is free and consists of m letters. Prove that*

$$L(x) = \frac{1}{1 - Nx + x^m}.$$

Problem 18. *Solve Problem 2 under the addition assumption that the sequence of commands breaking the computer is a free word.*

5 Transformations of words

Definition 6. Let M and M' be two sets of words. Let us divide the set M in two parts K and L. A function f mapping L to a subset I of M' is called a *transformation* of the set M to the set M' if f preserves lengths of words and is a one-to-one map of L onto I.

In this case, the set K is called the *kernel* of the transformation f and the set I is called the *image* of f.

A transformation will be denoted by an arrow: $M \Longrightarrow M'$.

Definition 7. A sequence of transformations

$$M_1 \Longrightarrow M_2 \Longrightarrow \ldots \Longrightarrow M_n$$

is called *exact* if the kernel of each subsequent transformation coincides with the image of the previous one.

Problem 19. *Let L be a language over an alphabet A, let G be the set of admissible words, and let N be the set of all non-admissible words. Construct an exact sequence of transformations*

$$\emptyset \Longrightarrow N \Longrightarrow F_A \Longrightarrow G \Longrightarrow \emptyset,$$

where F_A is the set of admissible words of the free language, that is, the set of all words over the alphabet A, and $\emptyset$ denotes the empty set.

Problem 20. *10 boys and 10 girls are sitting in a line so that boys' neighbors are girls and vice versa; their teacher is sitting next to them. Each of the children has some bonbons, and the total number of the boys' bonbons is equal to the total number of the girls' ones. The first boy gives all his bonbons to the girl sitting next to him. The girl eats all these bonbons, then she eats the same number of her own bonbons, and then she gives the rest of her bonbons to the next boy. He does the same (eats and gives the rest of bonbons to the next girl), then the next girl does the same, and so on. The last girl gives the rest of her bonbons to the teacher. How many bonbons does the teacher get?*

Problem 21. *Let*

$$\emptyset \Longrightarrow M_1 \Longrightarrow M_2 \Longrightarrow \ldots \Longrightarrow M_n \Longrightarrow \emptyset$$

be an exact sequence of transformations.

) Prove that if each set M_i consists of a finite number m_i of words, then

$$m_1 + m_3 + m_5 + \cdots = m_2 + m_4 + \ldots$$

b) Prove the following formula for the dimension series:

$$M_1(x) + M_3(x) + M_5(x) + \cdots = M_2(x) + M_4(x) + \ldots$$

Definition 8. A set M of words is called *free* if no word in M is a subword of another word in M, all words in M are free, and the words in M have no overlaps with each other.

Problem 22. *Let L be a language over an alphabet A, and let the set B of forbidden words of L be free. Denote the set of all admissible words by G and the set of all nonempty admissible words by $\overline{G}$. Construct an exact sequence of transformations*

$$\emptyset \Longrightarrow B \cdot G \Longrightarrow A \cdot G \Longrightarrow \overline{G} \Longrightarrow \emptyset.$$

Problem 23. *Let L be a language over an alphabet A of N letters, and let the set B of forbidden words of L be free. Prove the formula*

$$L(x) = \frac{1}{1 - Nx + B(x)}.$$

Problem 24. *Prove that the set of magic spells in Problem 1 is free, and solve the problem.*

Problem 25. *Find $L(x)$ provided that the alphabet of the language L is Latin and the forbidden words are* veni, vidi, vici.

Definition 9. Let L be a language. A *simple linkage* is a word $v = str$, where s, t, r are nonempty words such that the words $g = st$ and $f = tr$ are forbidden and there are no other forbidden subwords in v. The end r of the simple linkage (which is produced by cutting off the first forbidden subword g) is called the *tail* of v.

Problem 26. *Prove that the set of forbidden words of a language is free if and only if there are no simple linkages in it.*

Problem 27. *Let L be a language over an alphabet A, let B be its set of forbidden words, and let S be the set of all simple linkages. Denote the set of all admissible words by G and the set of all nonempty admissible words by $\overline{G}$. Construct an exact sequence of transformations*

$$S \cdot G \Longrightarrow B \cdot G \Longrightarrow A \cdot G \Longrightarrow \overline{G} \Longrightarrow \emptyset.$$

Problem 28. *Find the conditions on the set of forbidden words of a language L under which the exact sequence from Problem 27 could be extended to an exact sequence*

$$\emptyset \Longrightarrow S \cdot G \Longrightarrow B \cdot G \Longrightarrow A \cdot G \Longrightarrow \overline{G} \Longrightarrow \emptyset$$

(such languages are called non-tangled*). Give a formula to express the dimension series $L(z)$ of a non-tangled language in terms of the number N of letters and the dimension series of the sets B and S.*

Problem 29. *Find the dimension series of the language over the alphabet $\{a, b, c\}$ with forbidden words abb, bbc, bac.*

Problem 30. *Find the dimension series of the language over the alphabet $A = \{x_1, \dots, x_n, y_1, \dots, y_n, z_1, \dots, z_n\}$, if the forbidden words are the words of the form $x_i y_j$ and $y_j z_k$, where $1 \le i, j, k \le n$.*

Problem 31. *Prove that if the set of forbidden words of a non-tangled language consists of a single word, then this set is free.*

6 Free sets revisited

Problem 32. *Construct an infinite free set over an alphabet of two letters.*

Problem 33. *Suppose that the set of forbidden words B of a language L is free and the alphabet has more than one letter. Prove that the set of admissible words of the language is infinite.*

Definition 10. Let $S = s_0 + s_1 x + s_2 x^2 + \dots$ and $R = r_0 + r_1 x + r_2 x^2 + \dots$ be two series. If the inequality $s_k \ge r_k$ holds for any k, then we say that the following inequality for the series holds:

$$S \ge R.$$

Problem 34. *Prove that if series P, Q, and R satisfy the inequalities*

$$P \ge Q \text{ and } R \ge 0,$$

then

$$PR \ge QR.$$

Problem 35. *Suppose that for every $d > 0$ the sets B and B' of forbidden words of the languages L and L' over the same alphabet A contain the same number of words of length d, so that $B(z) = B'(z)$. Prove that if the set B is free, then the inequality*

$$L'(z) \ge L(z)$$

holds; in addition, we have $L'(z) = L(z)$ if and only if the set B' is also free.

Problem 36. *Suppose that the alphabet consists of two letters and the set B contains at least two words, including a word w of length 2.*

a) Prove that the set B is not free.

b) Is it possible that B is free if w is of length 3?

Problem 37. *Suppose that an alphabet consists of n letters and B consists of g two-letter words. Prove that if $g \leq n^2/4$, then the set B may be chosen to be free.*

Problem 38. *Prove that if $n = kd$ and $m \leq k^d(d-1)^{d-1}$, where the numbers d, k, m, n are positive integers, then, over an alphabet of n letters, one can choose a free set consisting of m words of length d.*

Problem 39. *) Prove that, if B is a free set over an alphabet of n letters, then there is the following inequality*

$$\frac{1}{1-nx+B(x)} \geq 1.$$

b) Is the converse true, that is, is it true that if the inequality

$$\frac{1}{1-nx+p(x)} \geq 1$$

holds for a positive integer n and a polynomial $p(x)$ whose coefficients are positive integers and whose constant term is zero, then there exists a free set B over an alphabet of n letters, with $B(x) = p(x)$?

Problem 40. *Let n be a positive integer and let $p(x)$ be a polynomial with positive integer coefficients and zero constant term. Prove that there exists a free set B with dimension series $B(x) = p(x)$ if and only if there exist two polynomials f and g with nonnegative integer coefficients with $f(0) = g(0) = 0$ such that*

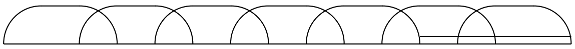
$$(1-f)(1-g) \geq 1 - nx + p(x).$$

Problem 41. [3] *Find a condition describing possible dimension series of the sets of forbidden words for non-tangled languages (like we described dimension series of free sets in problem 40).*

7 Words and chains

Definition 11. Let L be a language. *Chains of length one* are the forbidden words, and *chains of length 2* are the simple linkages. Next, one can define the chains of length 3, 4 etc. Namely, a word $v = str$ (where all the words s, t, r are nonempty) is called a *chain of length n* if its initial subword $g = st$ is a chain of length $n-1$, the final subword $f = tr$ is a forbidden word, where t is a subword of the tail p of the chain g, and there are no other forbidden subwords but f in the final subword pr. The *tail* of the chain v is the word r.

A chain looks as follows (each arc denotes a forbidden subword in the chain):



The length of the chain is the number of arcs. The only overlaps are of neighboring arcs (and the overlaps of neighboring arcs are non-empty). The emphasized two final tails do not contain any forbidden subword but the last arc.

For example, if the only forbidden word is *aba*, then the only chain of length one is *aba*, the only chain of length two is *ababa*, the only one of length 3 is *abababa*, etc.

Problem 42. *Suppose that the forbidden words in a language L are the words "tournament", "of", "towns". Write up all chains of length n.*

Problem 43. Antichains *of length n are defined in the same way as chains of length n, with the only difference that we read words of L in Definition 11 "from right to left", i.e., the tail of an antichain is to the left, and the initial antichain of length $n-1$ is to the right. Prove that the sets of length n chains and length n antichains coincide.*

Problem 44. *Prove that a chain of length n contains no other chain of length n as a subword.*

Problem 45. *Prove that if a word is decomposed as $w = gc$, where g is an admissible word and c is a chain, then, if in addition the length of c is greater than 1, w has exactly two decompositions of this form, and the lengths of the chains in these decompositions differ by 1.*

The next problem gives a way to solve the Main Problem.

Problem 46. *Let L be a language over an alphabet A. Let G be the set of its admissible words and $\overline{G}$ the set of all nonempty admissible words. Let C_1 be the set of chains of length one, C_2 the set of chains of length 2, and so on.*

Prove that

$$L(x) = \frac{1}{1 - Nx + C_1(x) - C_2(x) + C_3(x) - \dots}$$

[3] Neither a solution nor even an answer to this problem are known to the Jury

Problem 47. *Find the dimension series for the language in Problem 42.*

Problem 48. *Find all possible answers to Problem 2 depending on the form of the breaking sequence.*

Problem 49. *We say that a subword c of a word w is its* maximal subchain *if w can be decomposed as $w = gcu$, where g is an admissible word and c is a chain, and for any other decomposition $w = gc'u'$ with another chain c' the word c' is always a subword of c. Prove that any non-admissible word has a single maximal subchain of odd length.*

Problem 50. *Let L be a language over an alphabet A, and let A' be a new alphabet which extends A by one additional letter. Let L' be a language over A' with the same list of forbidden words as L. Prove that*

$$L'(x) = \frac{1}{\dfrac{1}{L(x)} - x}.$$

Problem 51. *A language W is called the* free product *of languages L and L' over disjoint alphabets A and A' if the alphabet of W is the union of the alphabets A and A' and the set of forbidden words is the union of the sets of forbidden words of L and L'. Express the dimension series of the free product W in terms of the dimension series of L and L'.*

Problem 52. *Suppose that all forbidden words of a language L are of two letters. Over the same alphabet, consider another language M whose forbidden words are all two-letter admissible words of L. Prove that*

$$L(x)M(-x) = 1.$$

8 Additional problems

Problem 53. *Prove that there exists a free set of m words of length d over an alphabet of $n = kd$ letters if and only if $m \leq k^d(d-1)^{d-1}$ (cf. Problem 38)*
a) for $d = 2$; *b) for $d = 3$;* *c) for $d > 3$.*

Definition 12. A language is said to be *d-defined* if the maximal length of its forbidden words is d. A 2-defined language is said to be *quadratic.*

Problem 54. *Quadratic languages L and M in Problem 52 are said to be* dual *to each other (notation: $M = L^!$).*
a) Prove that $(L^!)^! = L$.
b) Find $(L_1 + L_2)^!$.
c) Describe $(L_1 \cdot L_2)^!$.

Problem 55. *Let L be a d-defined language. Let us define a new language $L^{(n)}$ over the alphabet consisting of all length n admissible words of L as the language whose admissible words are all admissible words of L whose length is a multiple of n (rewritten in the new alphabet).*
a) Prove that $L^{(n)}$ is defined by finitely many forbidden words.
b) Is $L^{(n)}$ always d-defined?
c) For what minimal n, the language $L^{(n)}$ is necessarily either quadratic or free (for all d-defined languages L)?

Problem 56. *For any quadratic language L over the alphabet $x_1, \dots, x_n$, let us define an oriented graph Γ_L as follows: it has n vertices labelled with $x_1, \dots, x_n$, and there is an edge (an arrow) $x_i \to x_j$ if and only if the word x_ix_j is admissible. Denote the number of admissible words of length k by a_k. Prove that*
a) the language L is finite if and only if Γ_L has no cycles;
b) the language L has polynomial growth (i. e., there exist two nonzero polynomials p, q of the same degree d with positive leading coefficient such that $p(k) \geq a_k \geq q(k)$ for each $k \geq 0$) if and only if Γ_L has a cycle but has no intersecting cycles;
c) the language L has exponential growth (i. e., for some $c_1 > c_2 > 1$ and for all k, we have $c_1^k \geq a_k \geq c_2^k$) if and only if Γ_L has at least two intersecting cycles.

Problem 57. *Let L and $L^!$ be a pair of dual quadratic languages. Is it possible that both have exponential growth?*

Problem 58. *For any d-defined language L over the alphabet $x_1, \dots, x_n$, we define the oriented graph Γ_L as follows: its vertices are labelled with all admissible words of length $d-1$, and there is an edge (arrow) $v \to w$ if and only if there is a letter x_i such that the word vx_i is admissible and the last $d-1$ letters of it constitute the word w. Prove all properties a), b), c) in Problem 56 for Γ_L.*

Definition 13. Let M be a set over an alphabet A. Words u and v (over the same alphabet) are said to be *M-equivalent* if, for any word w, the words uw and vw either both belong to M or neither of them belongs to M. The set M is said to be *regular* if there is a natural number n such that any set of n contains two M-equivalent words.

Problem 59. *Prove that the set of admissible words of any language is regular.*

Definition 14. A *finite automaton* over an alphabet A is an oriented graph Γ with a finite set of vertices V such that
a) the arrows are marked by the letters of the alphabet A, and for every vertex $v \in V$ and each letter $a \in A$, there is a unique arrow marked by a whose tail is v;
b) an *initial vertex* $v_0 \in V$ and a *set of approving vertices* $W \subseteq V$ are given.

Let us consider each word over the alphabet A as an instruction for a trip by arrows over the finite automaton (Γ, v_0, W), that is, we begin with the initial vertex, then go by the arrow marked by the first letter of the word, then follow the arrow marked by the second letter of the word, and so on. We say that the automaton *approves* a word if the path corresponding to the word ends with an approving vertex.

Problem 60. *a) Prove that for every regular set M there exists a finite automaton approving the words of M and no other words.*
b) Prove that for every finite automaton the set of approving words is regular.

Problem 61. *Prove that for every regular set M its dimension series can we represented as a quotient of two polynomials.*

Problem 62. *Let L be a language and M_w the set of all admissible words of L which have a final subword equal to a given word w. Prove that the dimension series of the set M_w can we represented as a quotient of two polynomials.*

Note. Parts 1–5 were suggested before the intermediate consideration of the problems. Parts 6–8 were added after the intermediate consideration of the problems.