Методы сжатия данных

Ватолин, Ратушняк, Смирнов, Юкин

Managing Gigabytes

Compressing and Indexing Documents and Images // 1-2

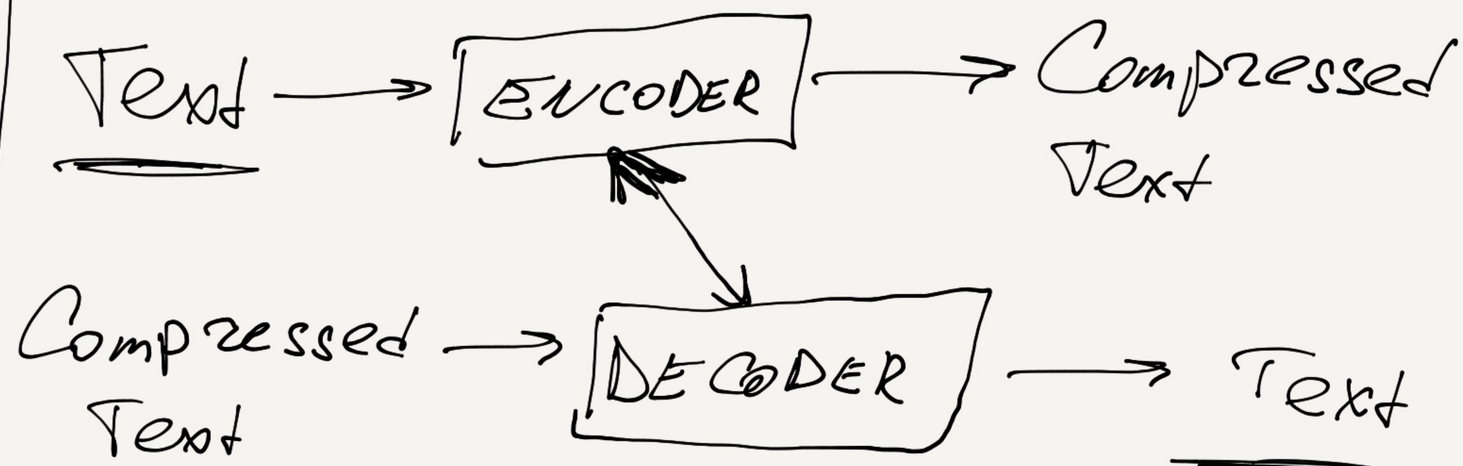
Witten, Moffat, Bell

Algorithms and Theory of Computation Handbook // ch. 12

Atallah

АБРАКАДАБРА — 1

000 001 010 011 100



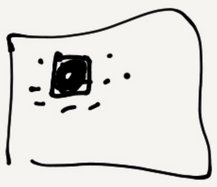
Text → EN → CT → DE → Text

lossless

Obj → EN → CO → DE → New Obj

≠

loss!



Obj → Temp Obj → ENCODER → ∞

New Obj ← DECODER ←

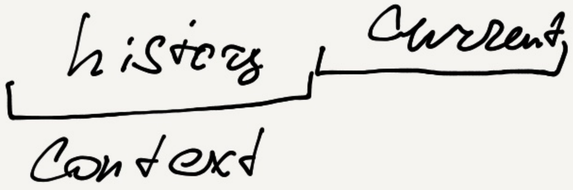
- Символьные — Хаффмана, арифметическое кодирование
- Словарные — LZ-77, LZ-78.

А роза упала

10111 111 01000110

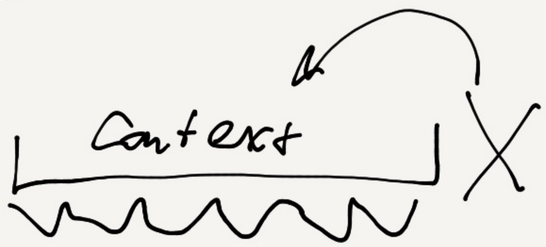
и — 2% 7 бит.

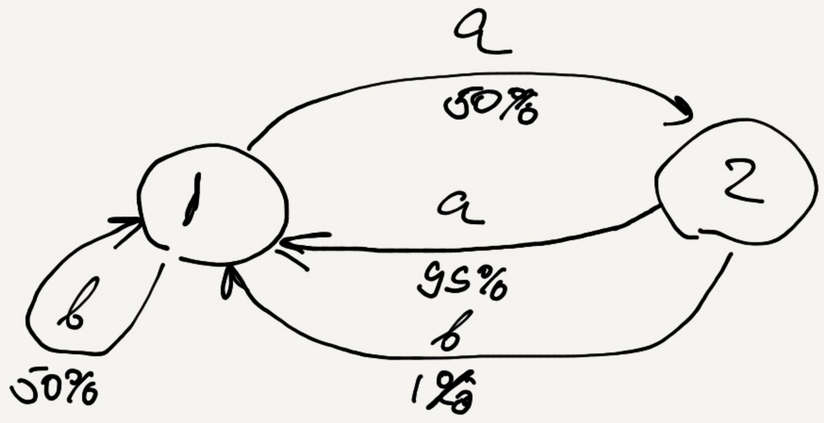
1 бит.



Finite context model.

- Размер
- Время





- статические
- полустатические / полуадаптивные
- адаптивные.

00 01 10 11

АА... Д ББ... Б ВВ... В ГГ... Г

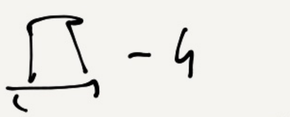
1000 1000 1000 1000

АБРА КАДАБРА

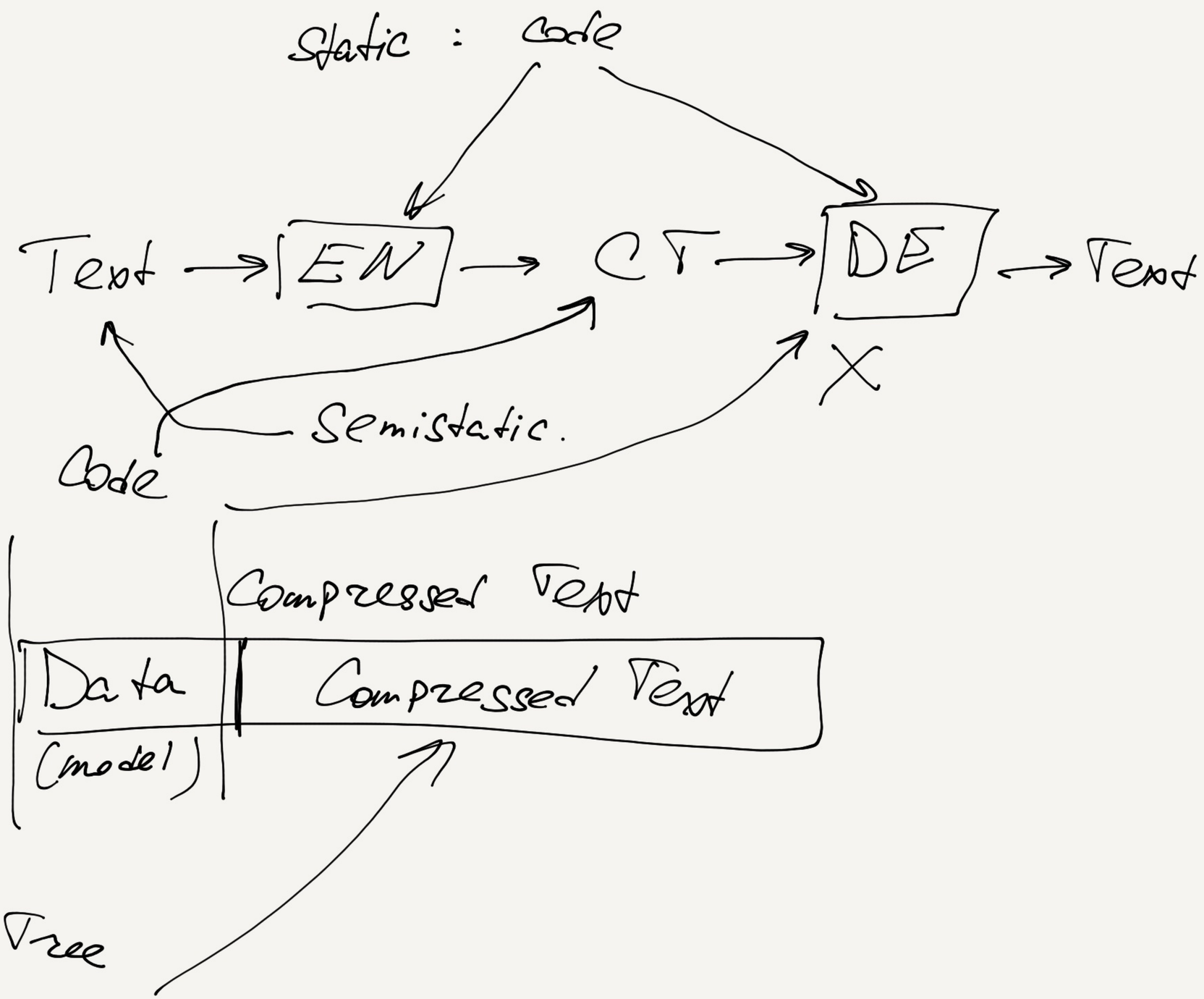
А - 2 11

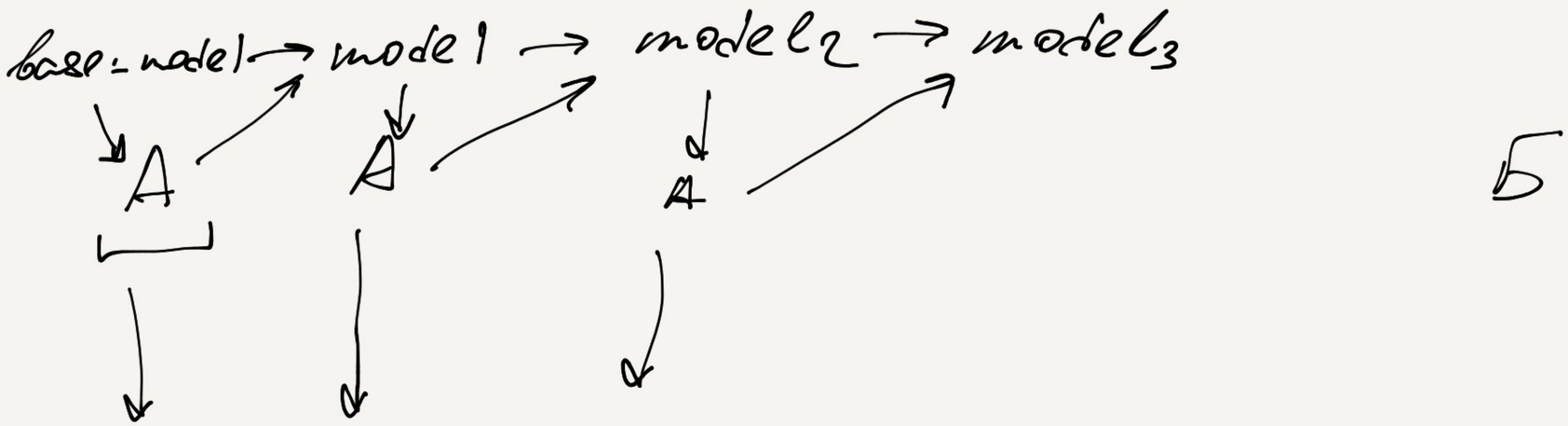
Б - 3 010

Р - 2



К - 5





5

10 10 ((\ |

$1000 \cdot 2 = 2000$

$998 \cdot 1 + 2 \cdot 2 = 1002.$

$a_1 \quad a_2 \quad \ldots \quad a_n$

$p_1 \quad p_2 \quad \ldots \quad p_n$

$$H = -p_i \log_2 p_i$$

$\sum_i p_i = 1.$

Словарные

LZ-77 / LZ-78

Lempel, Ziv

LZ-1 / LZ-2

ZL-1 / ZL-2

LZ-77

a b a c a b a b a c a b b b b b b b a

<0, 0, a>	
<0, 0, b>	
<2, 1, a>	
<3, 2, b>	
<6, 4, b>	
<2, 2, b>	
<5, 5, a>	

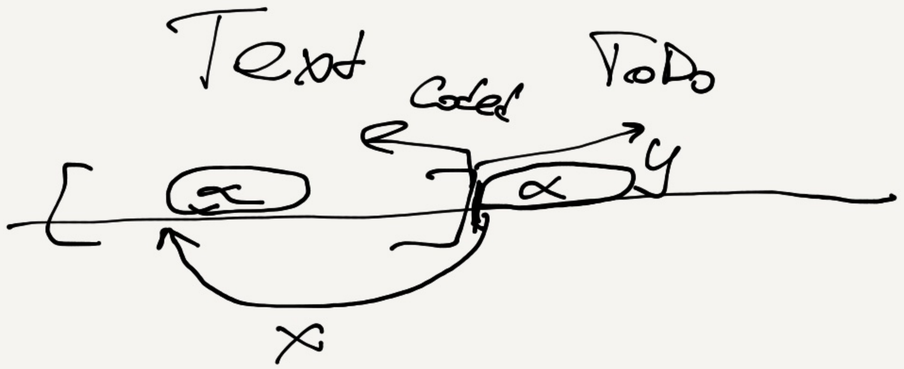
a b a c a b a b a a c b b b b b b b a

<Shift, size, is_new>

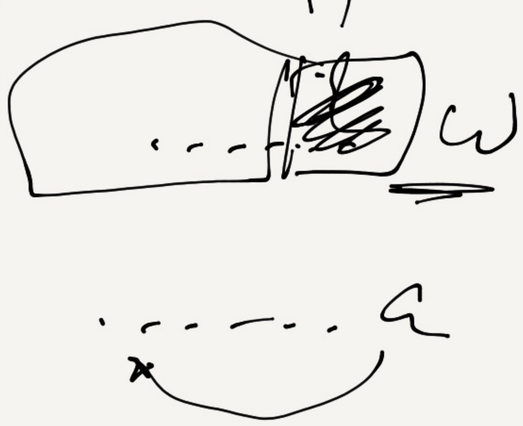
<5, 2, true, w>

<5, 3, false>

Text



<X, |α|, y>



<0,0,a>

<0,0,b>

<2,1,a>

<3,2,b>

<2,2,b>

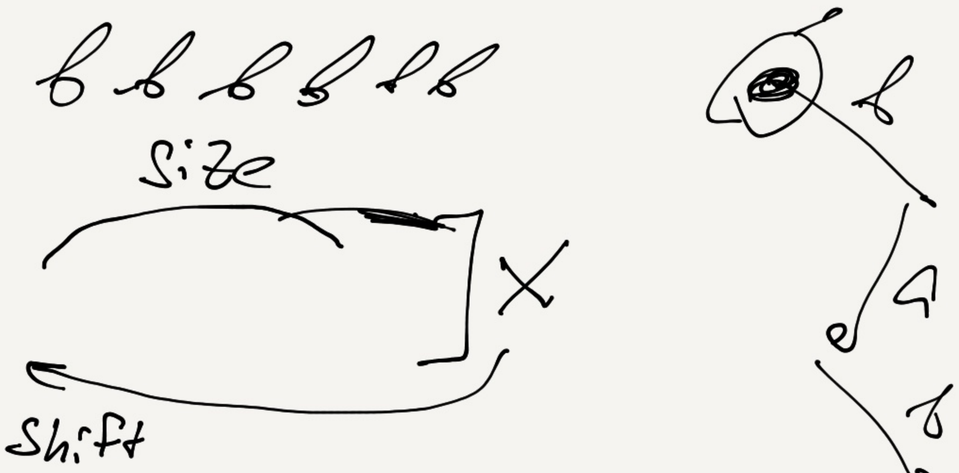
<2,2,b>

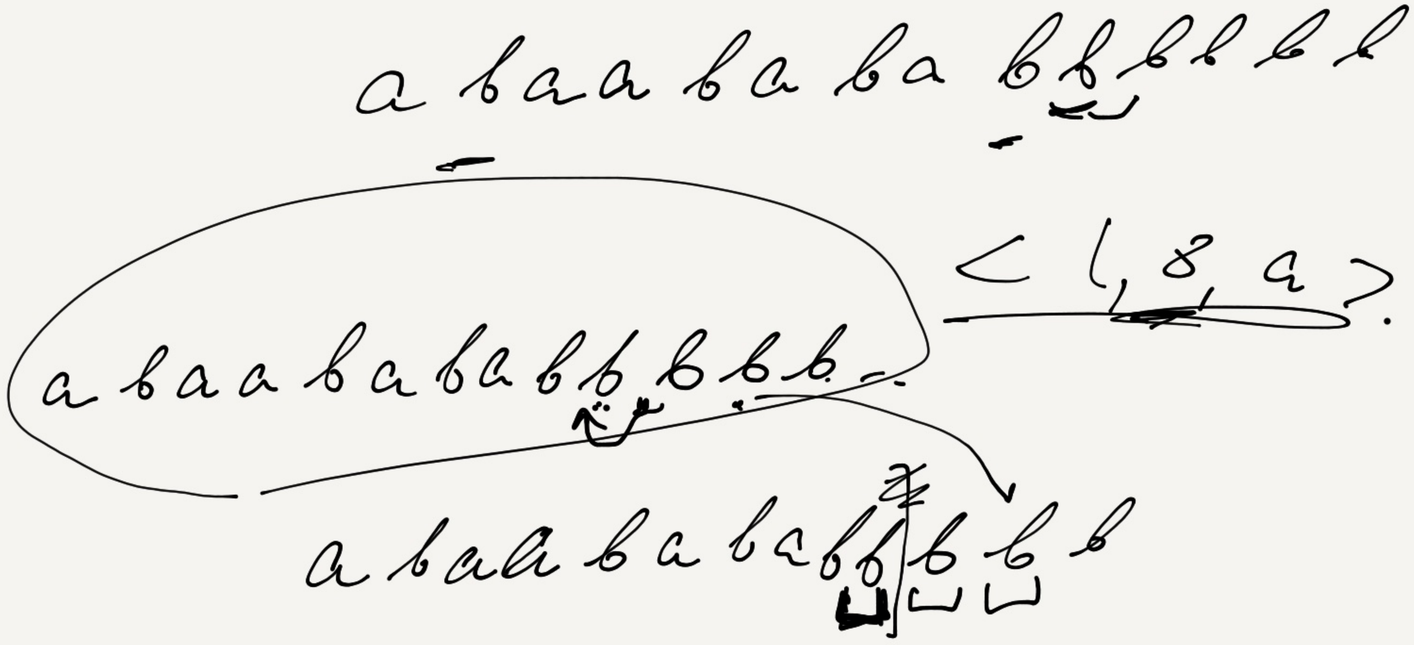
<5,5,a>

<1,8,a>

<shift, size, symbol>

Shift ≥ size

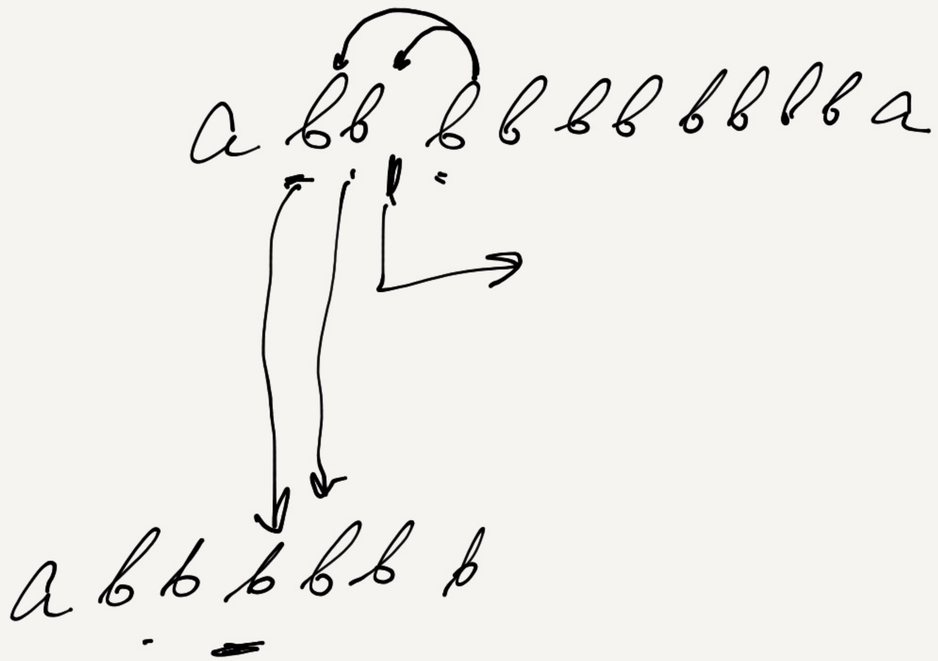


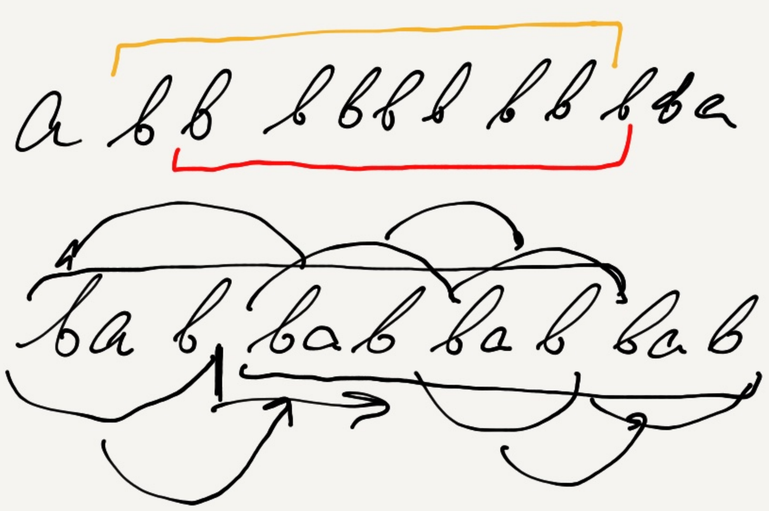


<1,8,a>

<3,9,EOF>

EOM.

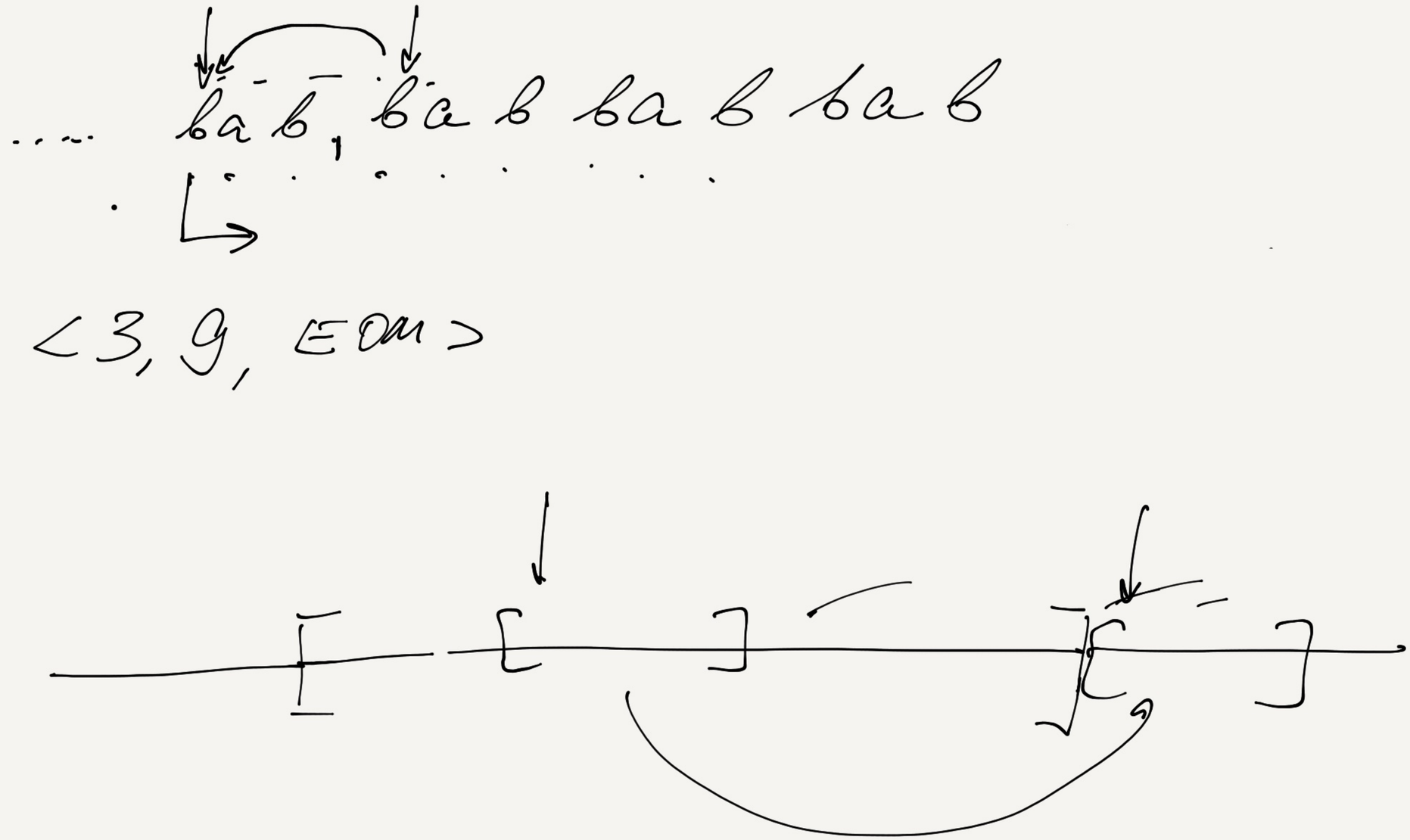




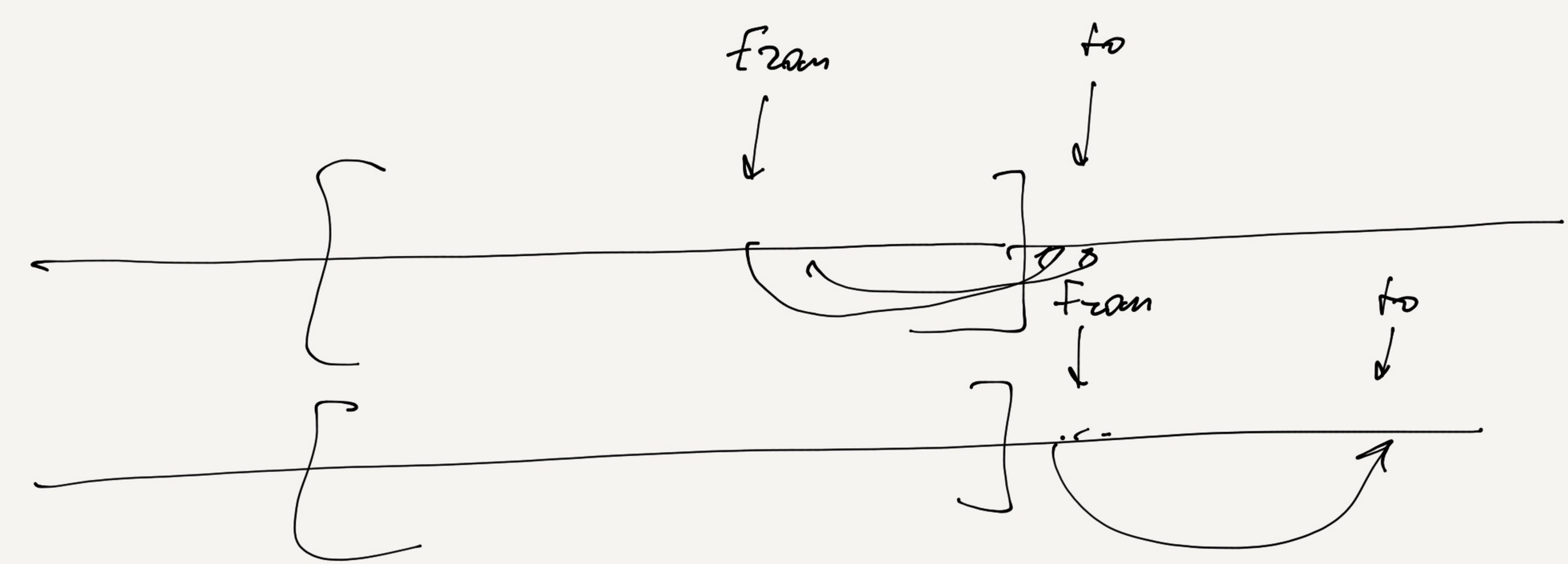
<Shift, size, Symbol>

Shift

<3, 9, EOM>



<3, 9, EOM>



LZ-78

ACAGAATAGAGA

<0, A> A
<0, C> C
<1, G> AG
<1, A> AA
<0, T> T
<3, A> AGA
<0, G> G
<1, EOM>

A-1
C-2
AG-3.
AA-4
T-5
AGA-6
G-7

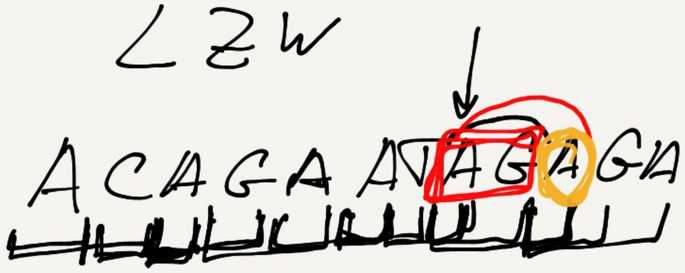
Flush
<0, A>.

<0, A>
<0, C>
<1, G>
<1, A>
<0, T>
<3, A>
<0, G>
<1, EOM>

A-1
C-2
AG-3
AA-4
T-5
AGA-6
G-7

ACAGAATAGAGA

LZW



wordy ∉ Dict

word y

→ wordy → Dict:

1, 2, 1, 3, 1, 1, 4, 7, 12.

A-1
C-2
G-3
T-4.
AC-5
CA-6
AG-7
GA-8
AA-9
AT-10
TA-11
AGA-12

1, 2, 1, 3, 1, 1, 4, 7, 12

ACAGAATAGAGA

A-1 AT-10

C-2 TA-11

G-3 AGA-12

T-4

AC-✓

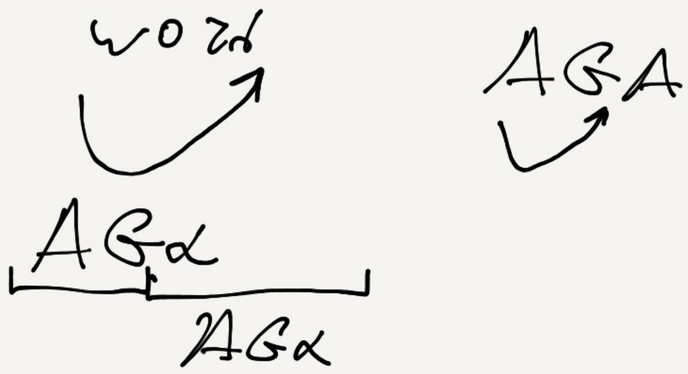
CA-6

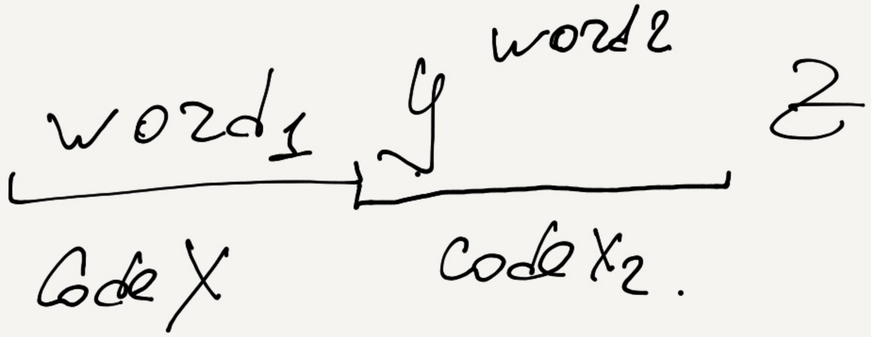
AG-7

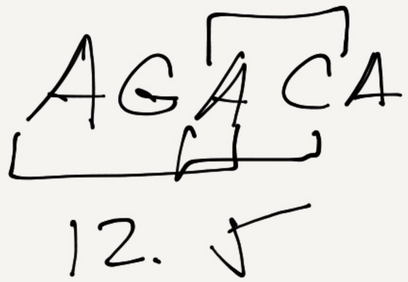
GA-8

AA-9

wordX — Code







ACA - 13.

word1 — Code X

word2 — Code X2.

word1 y ∉ Dict

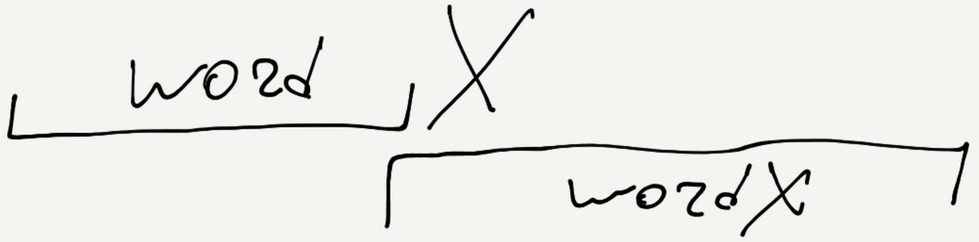
word1 y — Code y

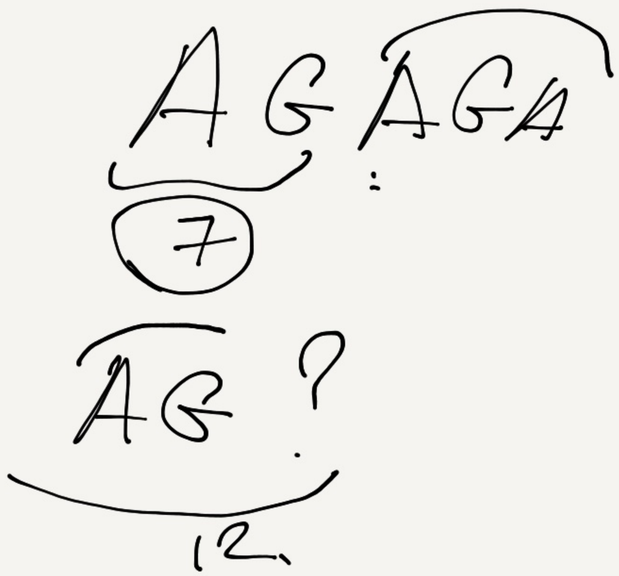
word2 z ∉ Dict

word2 z — Code y2

7, 12

AGA - 12





ENCODE

ACAGAATAGAGA

1, 2, 1, 3, 1, 1, 4, 7, 12

A-1
C-2
G-3
T-4
AC-5
CA-6
AG-7
GA-8
AA-9
AT-10
TA-11
AGA-12

decode

1, 2, 1, 3, 1, 1, 4, 7 | 12

ACAGAATAGAGA

A-1
C-2
G-3
T-4
AC-5
CA-6
AG-7
GA-8
AA-9
AT-10
TA-11
AGA-12

AGA

C

AG AGA

7, 12.

AG - 7

TA - 11

AGA - 12.

, 7, 12

TA - 11

12 ?

AG.

AG - 7.

AG A C A

AG - 7

AGA - 12.

7.

ACAGAATAGACA

1, 2, 1, 3, 1, 1, 4, 7, 5, 1

A - 1

C - 2

G - 3

T - 4

AC - 5

CA - 6

AG - 7.

GA - 8

AA - 9

AT - 10

TA - 11

AGA - 12

ACA - 13

1, 2, 1, 3, 1, 1, 4, 7, 5, 1

ACAGAATAGACA

A - 1

C - 2

G - 3

T - 4

AC - 5

CA - 6

AG - 7

GA - 8

AA - 9

AT - 10

TA - 11

AGA - 12

ACA - 13

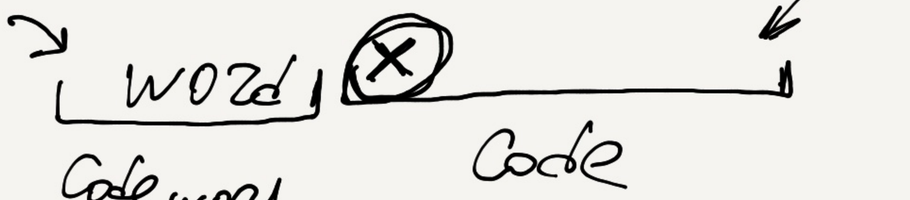
AGAGC

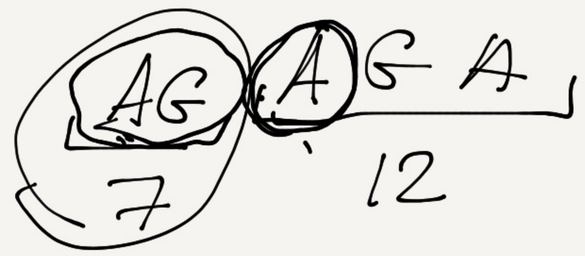
7, 7

AGA - 12.

AGC - 13

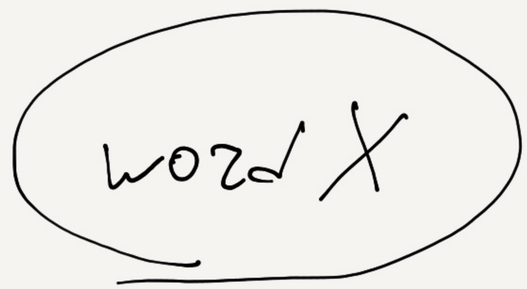
word X



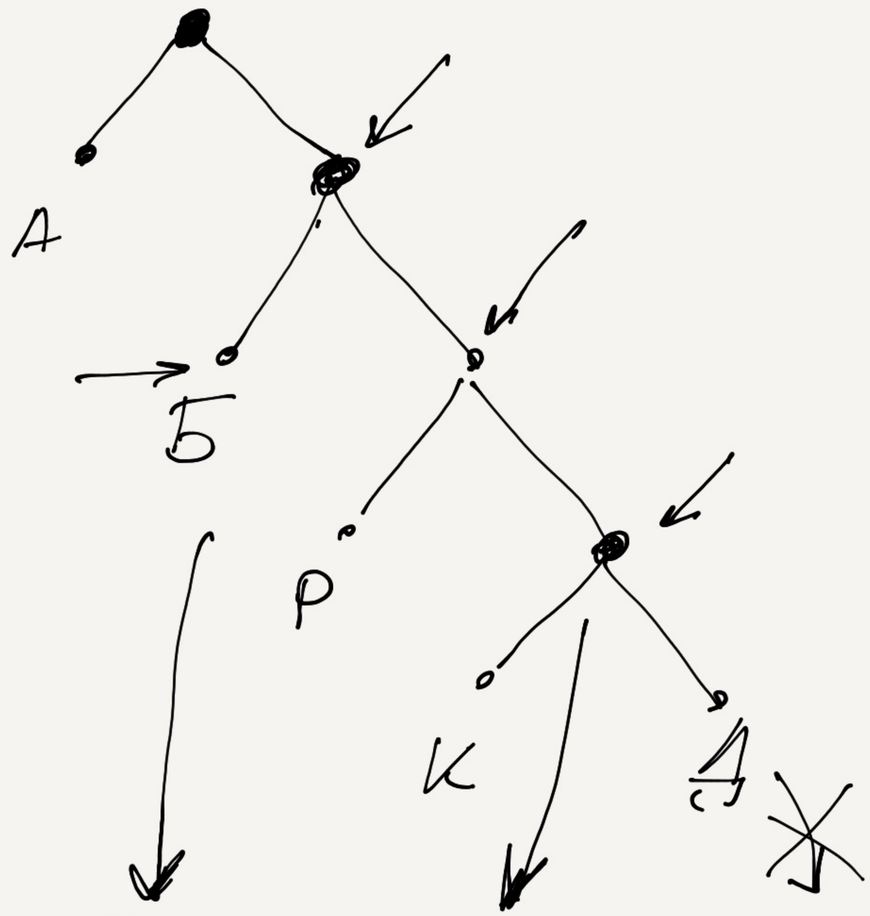


AGA - 12

Code & Dict.



Codeword , Code



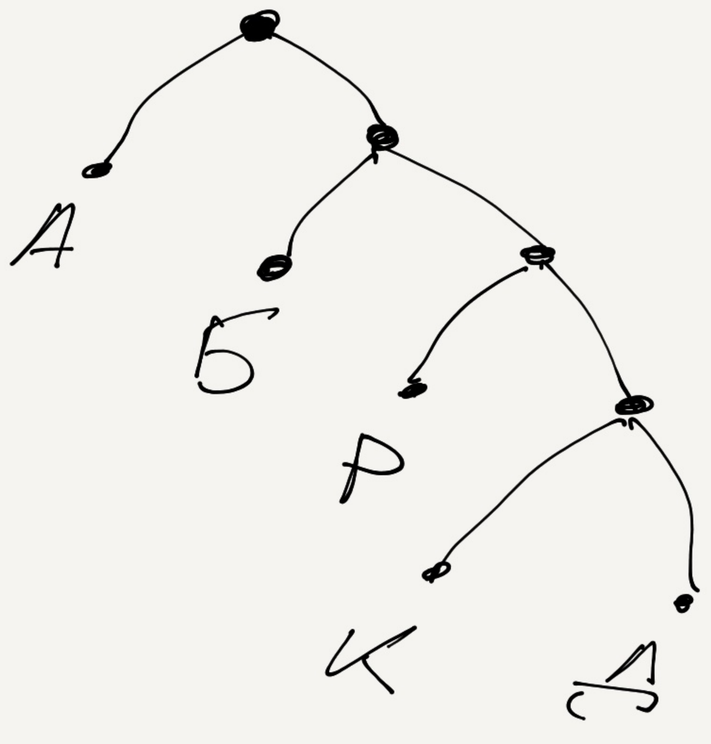
A - 0

Б - 10

Р - 110

К - 1110

Д - 1111



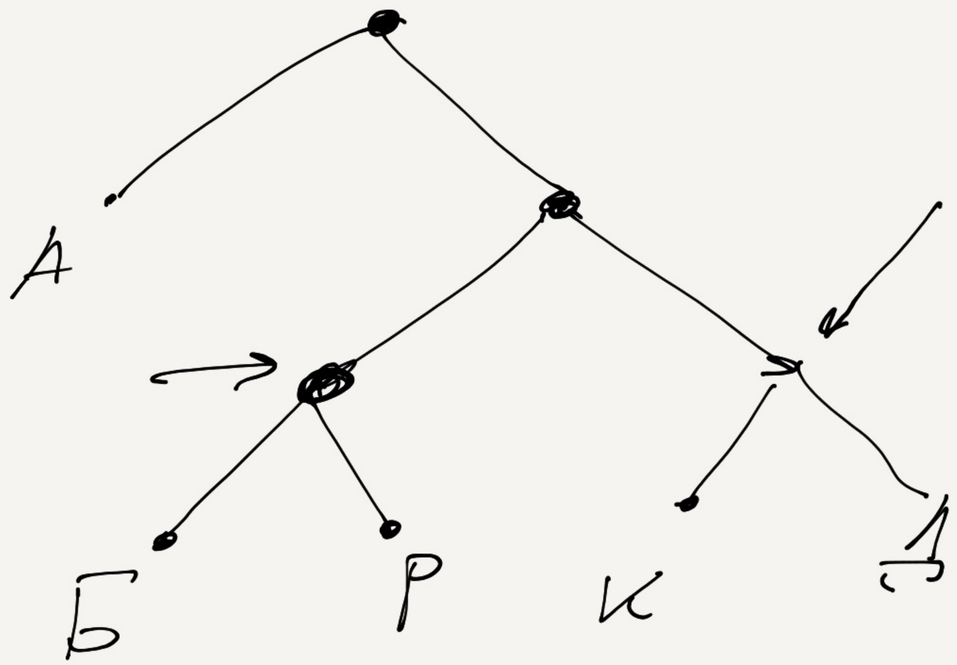
Text → EN → CT → DE → ?

Compressed Text

0 1 A 0 1 Б 0 1 Р 0 1 К 1 Д	Data

Decoder Data

0 1 A 0 1 Б 0 1 Р 0 1 К 1 Д [Data]



0 1 А 0 0 1 Б 1 Р 0 1 К 1 Д