Due to the presence of dependencies that the project pulls and outdated files, the proposed files are difficult to interpret, so the project carries a number that is indicative for its own portfolio. The main menu of the layout is shown in Figure 1.

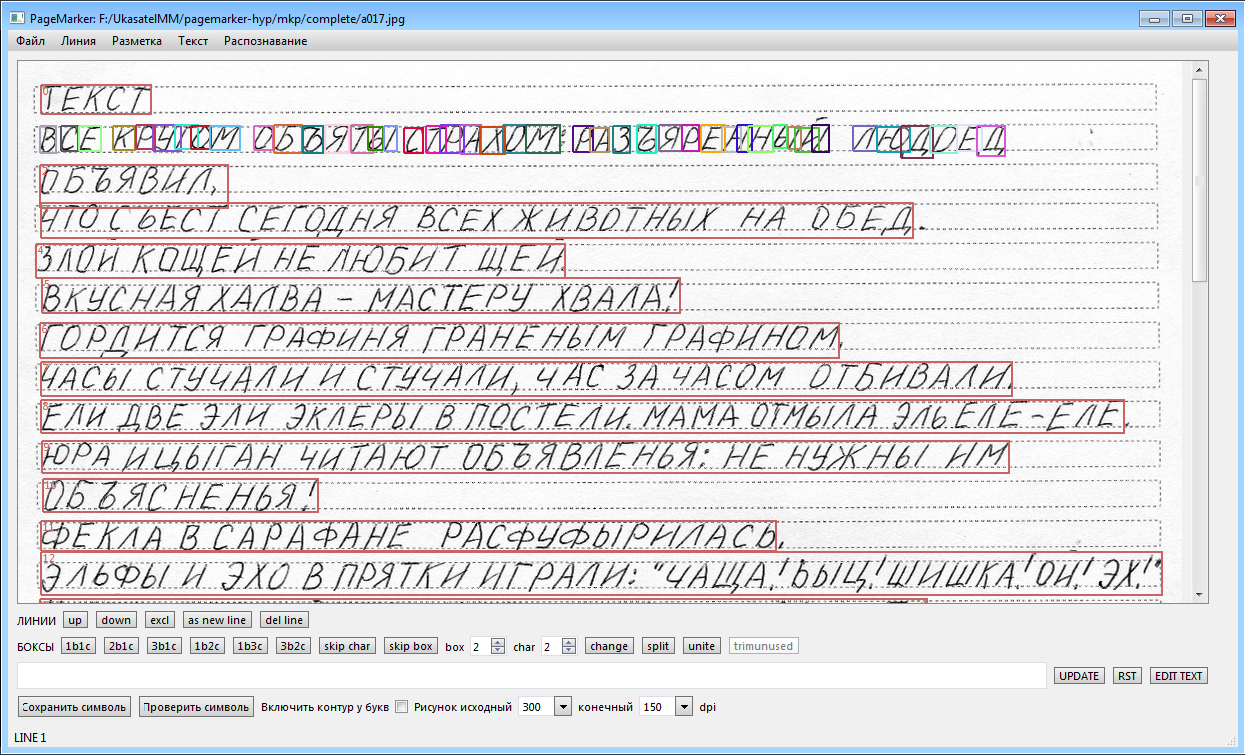


Figure 1 - The main menu of the project

The project provided for the loading of an image of printed text, automatic and manual marking, the calculation of hypotheses for each character, letter, and character sets. The final recognition accuracy based on the results of testing a database of 1000 pages has reached 94%.

Building a recognition model

The construction of a list of hypotheses of the most probable word (Russian or English word, a sequence of numbers (number or model name), syntactic symbols (-, signs> <=, links ([])) is considered. This stage of processing involves knowing the boundaries of the word, which is located within empty cells (pauses). Figure 2 shows an example of all possible speech segments that were written by hand. The purpose of the algorithm is to compose the correct sequence from these segments and find out which word the author means. It is necessary to make a list of the most likely hypotheses. In Figure 2, the word "EKO" should be recognized. The author wrote the letter "E" with 3 dashed lines (segments) - segments №1, №5, №9.

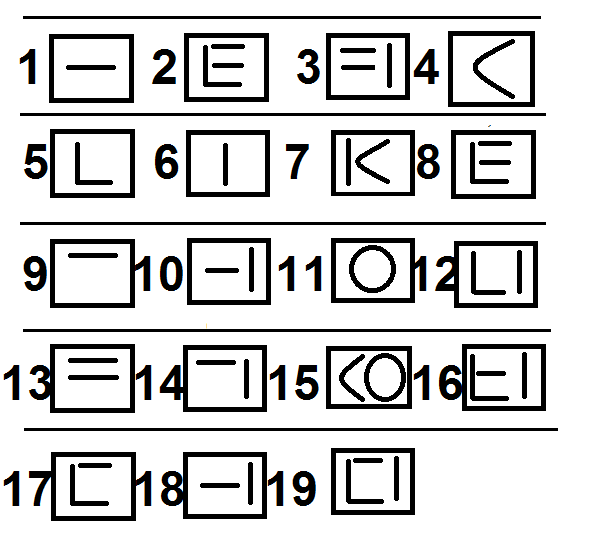


Figure 2 - Segments of the word "EKO"

The acceptable segment boundaries are known. That is, the segment  cannot merge with segment , but segment  combined with segment , I form the letter "E". The main problem is that the first segment can be recognized by the neural network as the letter "L", and the second as the sign "=". It is necessary for the algorithm to take into account the context when recognizing the whole word.

The first step is to make all kinds of combinations of character sequences. An example of possible sequences is shown in Figure 3.

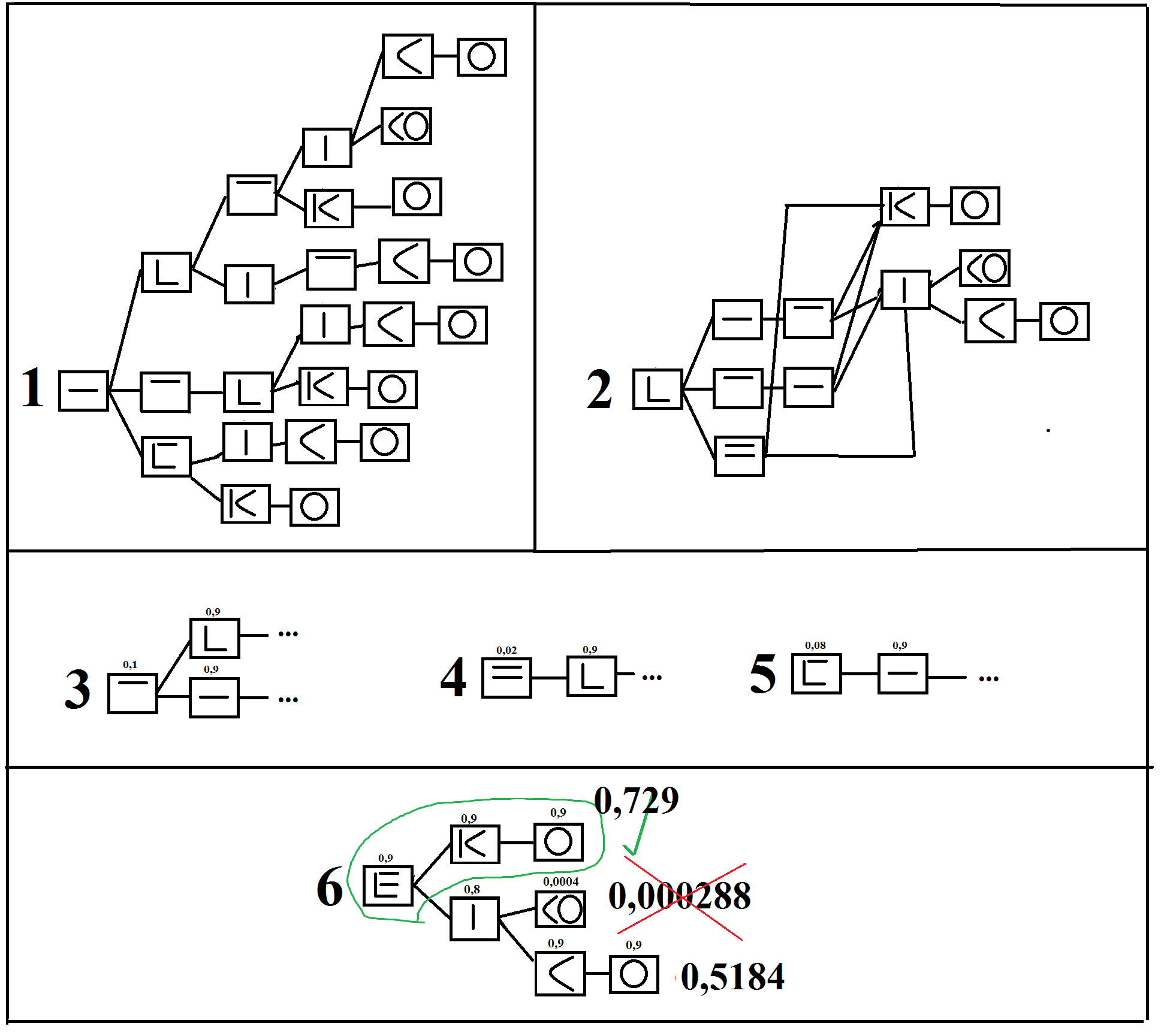


Figure 3 - Example of possible sequences

As mentioned above, for each segment there is a vector of probability of belonging to the class - PW, consisting of 83 elements (example - Figure 4).

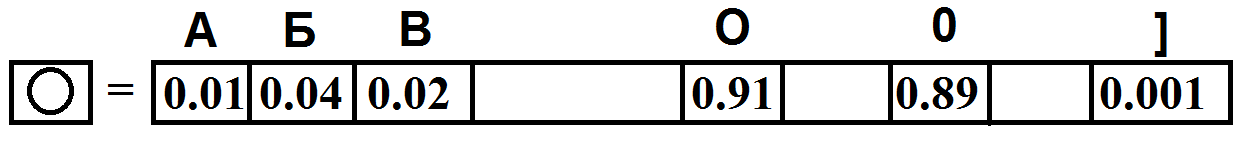


Figure 4 - Example of the output vector

Figure 4 shows that the segment has close probabilities to the letter "O" and to the number "0". Let's move on to building a list of hypotheses Hi

For each segment, a recognized character (letter / number / syntactic sign) is assigned with a probability according to the formula 1.

 (1)

where r – is the number of symbols, the length of the output vector of the neural network,

n – is the number of possible hits in the confidence interval.

The confidence interval is taken as a constant value equal to 0.1. There may be a case in which the number of characters falling into the confidence interval turns out to be a certain set. This procedure is necessary to further compile the list and take into account all possible hypotheses. Figure 5 shows an example in which the hypothesis tree No. 6 in Figure 3 is expanded when taking into account new branches. The extension itself is necessary for the case when the probabilities are close, as in the example in Figure 4.

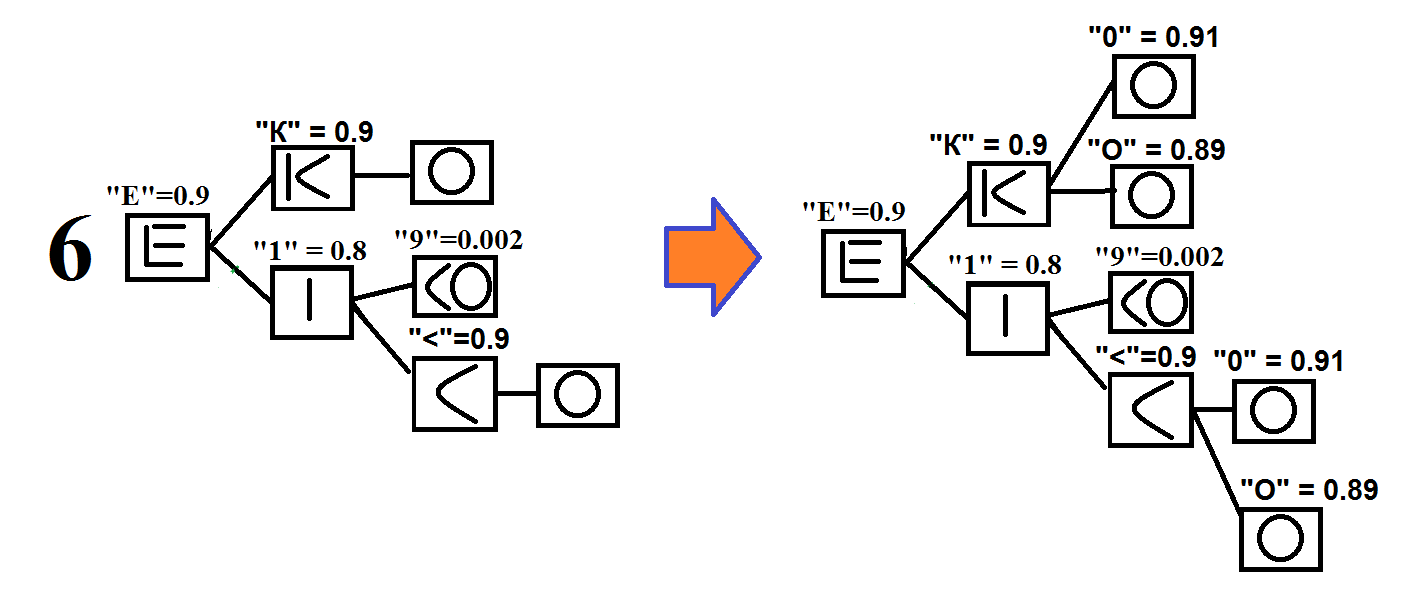


Figure 5 - An example of expanding tree No. 6 from Figure 3, when linking possible sequential segments with the result of recognition by a neural network.

The list of hypotheses will be called the possible paths along the constructed all constructed trees. Figure 6 shows an example of a list of hypotheses for tree 6 in Figure 7

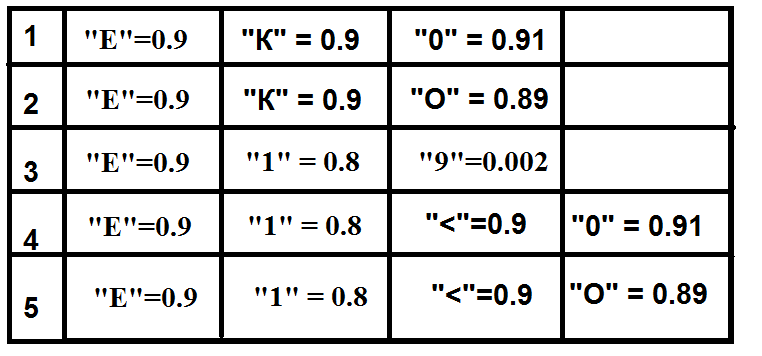


Figure 7 - An example of a list of hypotheses Hi

I – is the total number of hypotheses made up of tree branches for one word

Having compiled a list of all possible recognition options, we proceed to search for the most likely combination.

To compile the following model, you need a dictionary of Russian and English words, combinations with syntactic symbols (Figure 8), as well as calculated constant values of combinations from the dictionary:

1) W1 (symbol) - the number of occurrence of a symbol in the dictionary (W1 (a) = 399834, W1 (b) = 72922, ..., W1 (i) = 150,000)

2) W2 (symbol) - the number of occurrence of the studied pair in the dictionary (W2 (ab) = 6853, W2 (for) = 25378, W2 (br) = 9001)

3) W3 (symbol) - the number of occurrence of the investigated triplet in the dictionary

4) W4 (symbol) - the number of occurrence of the investigated sequence of 4 letters.

5) N1 - the total sum of all W1 (characters), where the number of characters = 83

6) N2 - the total sum of all W2 (characters), where the number of characters =?

7) N3 - the total sum of all W3 (characters), where the number of characters =?

8) N4 - the total sum of all W4 (characters), where the number of characters =?

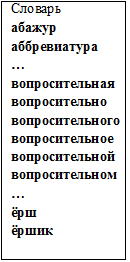


Figure 8 - Example vocabulary

General formula:

Where m is the number of single recognized beech

j - letter pairs

g - triples

d - fours



– probability of occurrence of a combination of symbols

This probability will build the priority on the fully recognized characters.

 – discharge

 – the total number of investigated combinations found in the dictionary.

 – of all combinations of the i-th digit in the dictionary (for example:

* In the case of the first digit - all characters (а, б, в,…, A,B,С,…,]=,! etc.)
* In the case of the second digit - combinations of double sequences of characters (ал, ма, па, ->, о., ve, r1 и т.д.)
* In the case of the third digit, combinations of triple character sequences (mom, dad, 123, lon, [54, etc.)
* In the case of the fourth digit - combinations of quadruple sequences of characters