Министерство образования Российской Федерации

Федеральное агентство по образованию

КАЗАНСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. А.Н. Туполева - КАИ

Кафедра АСОИУ

Лабораторная работа №1-2

по дисциплине

«ТЕОРИЯ ЯЗЫКОВ ПРОГРАММИРОВАНИЯ»

Выполнил:

Студент группы 4309

Ильичев Д.А.

Проверил: Бикмуллина И.АИ.

Казань 2024

**Лабораторная работа № 1. Разработка транслитератора**

**Литера** — буква, один из знаков азбуки; в типографии наборная буква.

**Транслитератор** — это инструмент, который помогает переводить текст с одного языка на другой, сохраняя при этом структуру и порядок слов.

Существует два основных подхода к созданию транслитераторов: машинное обучение (ML) и основанный на правилах (rule-based).

Подход с использованием машинного обучения строится на вероятностных моделях последовательностей, таких как марковские модели. В основе этого метода лежит алгоритм Витерби: сегментация латинских слов и поиск наиболее вероятной последовательности скрытых состояний — кириллических комбинаций.

Этот метод можно рассматривать как задачу машинного перевода: создаётся таблица соответствий букв, и на её основе модель обучается вероятностным преобразованиям. ML требуется качественный датасет, который необходимо собрать.

Теперь перейдём к подходу, основанному на правилах. Для этого используются словари, таблицы сочетаемости букв.

**Как работает транслитератор:**

1. Анализируем входную последовательность символов.
2. Ищет комбинацию в заданном классе, словаре или в таблице транслитерации.
3. Если комбинация найдена, то преобразуем её и корректируем несочетаемости символов с помощью регулярных выражений.

Некоторые последовательности могут иметь неоднозначные интерпретации на русском языке. Для решения таких ситуаций создаются правила:

1. Преобразование CH. В начале слова это может быть Ч, Х или К, в конце — Ч или Х. Простой способ: если перед CH стоит И или Ы, то это преобразуется в КР, иначе — в ХР.
2. Разделительный твёрдый знак. Проверяем приставки, и, если следующий символ — йотированная гласная (Е, Ё, Ю, Я), ставим твёрдый знак перед ней.
3. Йотированные гласные. Они идут после приставок с твёрдым знаком или в начале слова. Ошибки возможны (например, IONY → ЁНЫ вместо ИОНЫ), но это можно минимизировать через словари, где для слов с ИО часто следуют буквы ТАД.
4. ИЙ или ЫЙ? Если перед ними стоят звуки ГДЖКЦЧШЩ, то используется ИЙ, иначе — ЫЙ.

Как можно улучшить эту модель:

1. Распознавать мягкий знак.
2. Использовать актуальный орфографический словарь для корректировки правил и их веса, что может постепенно привести к ML-подходу.
3. Добавить спеллчекер (система проверки правописания) для улучшения результатов транслитерации.

**Текст задания:**

1. Спроектировать и отладить экранную форму для ввода исходных данных, вывода сообщений программы и управления программой.
2. Разработать и отладить транслитератор void GetSymbol(), пример имеется в модуле uLexicalAnalizer из папки «Программы».
3. Для отладки транслитератора временно включить в обработчик нажатия кнопки цикл чтения с помощью функции GetSymbol() символов исходного текста и вывода результатов анализа в поле диагностических сообщений.

**Код программы:**

**Form1.cs**

**using** System**;**

**using** System**.**Collections**.**Generic**;**

**using** System**.**ComponentModel**;**

**using** System**.**Data**;**

**using** System**.**Drawing**;**

**using** System**.**Text**;**

**using** System**.**Windows**.**Forms**;**

**namespace** Translator

**{**

**public** **partial** class Form1 **:** Form

**{**

**public** Form1**()**

**{**

InitializeComponent**();**

int n **=** tbFSource**.**Lines**.**Length**;** // Получение количества строк в tbFSource

**}**

// Обработчик события нажатия на кнопку btnFStart

**private** void btnFStart\_Click**(object** sender**,** EventArgs e**)**

**{**

// Создание экземпляра класса CLex для анализа текста

CLex Lex **=** **new** CLex**();**

Lex**.**strPSource **=** tbFSource**.**Lines**;** // Устанавливаем источник текста из текстового поля tbFSource

Lex**.**strPMessage **=** tbFMessage**.**Lines**;** // Устанавливаем сообщения из текстового поля tbFMessage

int x **=** tbFSource**.**TextLength**;** // Получаем количество символов в tbFSource

int y **=** tbFSource**.**Lines**.**Length**;** // Получаем количество строк в tbFSource

tbFMessage**.**Text **=** ""**;** // Очищаем tbFMessage для нового результата

**try**

**{**

// Цикл продолжается, пока состояние парсера не станет "Finish"

**while** **(**Lex**.**enumPState **!=** TState**.**Finish**)**

**{**

Lex**.**GetSymbol**();** // Получение текущего символа из исходного текста

Lex**.**NextToken**();** // Получение следующего токена

String s **=** ""**;** // Переменная для хранения литеры

String s1 **=** ""**;** // Переменная для хранения типа литеры

// Определение типа символа и присваивание значений переменным s и s1

**switch** **(**Lex**.**enumFSelectionCharType**)**

**{**

**case** TCharType**.**EngLetter**:** **{** s1 **=** "EngLetter"**;** **break;** **}**

**case** TCharType**.**RusLetter**:** **{** s1 **=** "RusLetter"**;** **break;** **}**

**case** TCharType**.**Digit**:** **{** s1 **=** "Digit"**;** **break;** **}**

**case** TCharType**.**Space**:** **{** s1 **=** "Space"**;** **break;** **}**

**case** TCharType**.**Star**:** **{** s1 **=** "Star"**;** **break;** **}**

**case** TCharType**.**Exclamation**:** **{** s1 **=** "Exclamation"**;** **break;** **}**

**case** TCharType**.**Equal**:** **{** s1 **=** "Equal"**;** **break;** **}**

**case** TCharType**.**Semicolon**:** **{** s1 **=** "Semicolon"**;** **break;** **}**

**case** TCharType**.**EndBracket**:** **{** s1 **=** "EndBracket"**;** **break;** **}**

**case** TCharType**.**OpenBracket**:** **{** s1 **=** "OpenBracket"**;** **break;** **}**

**case** TCharType**.**OpenSquadBracket**:** **{** s1 **=** "OpenSquadBracket"**;** **break;** **}**

**case** TCharType**.**EndSquadBracket**:** **{** s1 **=** "EndSquadBracket"**;** **break;** **}**

**case** TCharType**.**Colon**:** **{** s1 **=** "Colon"**;** **break;** **}**

**case** TCharType**.**Minus**:** **{** s1 **=** "Minus"**;** **break;** **}**

**case** TCharType**.**Plus**:** **{** s1 **=** "Plus"**;** **break;** **}**

**case** TCharType**.**Comma**:** **{** s1 **=** "Comma"**;** **break;**

**case** TCharType**.**Dot**:** **{** s1 **=** "Dot"**;** **break;** **}**

**case** TCharType**.**AnotherSymbol**:** **{** s1 **=** "ReservedSymbol"**;** **break;** **}**

**case** TCharType**.**NoInd**:** **{** s1 **=** "NoInd"**;** **break;** **}**

**case** TCharType**.**EndRow**:** **{** s **=** "KC"**;** s1 **=** "EndRow"**;** **break;** **}**

**case** TCharType**.**EndText**:** **{** s **=** "KT"**;** s1 **=** "EndText"**;** **break;** **}**

**}**

// Создание строки с литерой и ее типом

String m **=** "(" **+** s **+** "," **+** s1 **+** ")"**;**

// Добавление строки в tbFMessage

tbFMessage**.**Text **+=** m**;**

**}**

**}**

**catch** **(**Exception exc**)** // Обработка исключений

**{**

// Добавление сообщения об ошибке в tbFMessage

tbFMessage**.**Text **+=** exc**.**Message**;**

tbFSource**.**Select**();** // Устанавливаем фокус на tbFSource

tbFSource**.**SelectionStart **=** 0**;** // Устанавливаем начальную позицию выделения

int n **=** 0**;**

// Подсчет количества символов для выделения текста до текущей позиции

**for** **(**int i **=** 0**;** i **<** Lex**.**intPSourceRowSelection**;** i**++)**

n **+=** tbFSource**.**Lines**[**i**].**Length **+** 2**;** // +2 учитывает переход на новую строку

n **+=** Lex**.**intPSourceColSelection**;** // Добавляем текущую позицию в строке

tbFSource**.**SelectionLength **=** n**;** // Устанавливаем длину выделения

**}**

**}**

**}**

**}**

**uLex.cs**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Threading.Tasks;

namespace Translator

{

// Перечисление состояний анализа

public enum TState { Start, Continue, Finish }; // Тип состояния

// Перечисление возможных типов символов

public enum TCharType { EngLetter, RusLetter, Digit, EndRow, EndText, Space, Star, Slash, Exclamation, Equal, Semicolon, AnotherSymbol, OpenBracket, EndBracket, Colon, OpenSquadBracket, EndSquadBracket, Plus, Minus, Comma, Dot, NoInd }; // Тип символа

// Перечисление типов токенов

public enum TToken { lxmIdentifier, lxmNumber, lxmUnknown, lxmEmpty, lxmLeftParenth, lxmRightParenth, lxmIs, lxmDot, lxmComma };

// Класс лексического анализатора

public class CLex

{

// Поля класса

private String[] strFSource; // Массив строк, представляющий исходный текст

private String[] strFMessage; // Массив строк для сообщений (возможно, для вывода результатов)

public TCharType enumFSelectionCharType; // Тип текущего символа

public char chrFSelection; // Текущий символ

private TState enumFState; // Текущее состояние анализатора

private int intFSourceRowSelection; // Номер текущей строки

private int intFSourceColSelection; // Номер текущей колонки в строке

private String strFLexicalUnit; // Текущая лексическая единица

private TToken enumFToken; // Текущий токен

// Свойства для доступа к полям класса

public String[] strPSource { set { strFSource = value; } get { return strFSource; } }

public String[] strPMessage { set { strFMessage = value; } get { return strFMessage; } }

public TState enumPState { set { enumFState = value; } get { return enumFState; } }

public String strPLexicalUnit { set { strFLexicalUnit = value; } get { return strFLexicalUnit; } }

public TToken enumPToken { set { enumFToken = value; } get { return enumFToken; } }

public int intPSourceRowSelection { get { return intFSourceRowSelection; } set { intFSourceRowSelection = value; } }

public int intPSourceColSelection { get { return intFSourceColSelection; } set { intFSourceColSelection = value; } }

// Конструктор класса

public CLex()

{

}

// Метод получения текущего символа из источника

public void GetSymbol()

{

// Проверяем, не вышли ли за пределы строки

if (intFSourceColSelection > strFSource[intFSourceRowSelection].Length - 1)

{

intFSourceRowSelection++; // Переходим на следующую строку

if (intFSourceRowSelection <= strFSource.Length - 1)

{

// Если еще не конец текста, сбрасываем колонку и задаем символ конца строки

intFSourceColSelection = -1;

chrFSelection = '\0'; // Устанавливаем текущий символ в '\0'

enumFSelectionCharType = TCharType.EndRow; // Указываем, что достигнут конец строки

enumFState = TState.Continue; // Состояние анализа продолжается

}

else

{

// Если достигли конца всего текста

chrFSelection = '\0'; // Устанавливаем текущий символ в '\0'

enumFSelectionCharType = TCharType.EndText; // Указываем, что достигнут конец текста

enumFState = TState.Finish; // Меняем состояние на "Finish"

}

}

else

{

// Получаем текущий символ в строке

chrFSelection = strFSource[intFSourceRowSelection][intFSourceColSelection];

// Классифицируем символ

if (chrFSelection == ' ') enumFSelectionCharType = TCharType.Space;

else if (chrFSelection >= 'a' && chrFSelection <= 'z') enumFSelectionCharType = TCharType.EngLetter;

else if (chrFSelection >= 'а' && chrFSelection <= 'я') enumFSelectionCharType = TCharType.RusLetter;

else if (chrFSelection >= '0' && chrFSelection <= '9') enumFSelectionCharType = TCharType.Digit;

else if (chrFSelection == '/') enumFSelectionCharType = TCharType.Slash;

else if (chrFSelection == '\*') enumFSelectionCharType = TCharType.Star;

else if (chrFSelection == '!') enumFSelectionCharType = TCharType.Exclamation;

else if (chrFSelection == '=') enumFSelectionCharType = TCharType.Equal;

else if (chrFSelection == ';') enumFSelectionCharType = TCharType.Semicolon;

else if (chrFSelection == '(') enumFSelectionCharType = TCharType.OpenBracket;

else if (chrFSelection == ')') enumFSelectionCharType = TCharType.EndBracket;

else if (chrFSelection == ':') enumFSelectionCharType = TCharType.Colon;

else if (chrFSelection == '[') enumFSelectionCharType = TCharType.OpenSquadBracket;

else if (chrFSelection == ']') enumFSelectionCharType = TCharType.EndSquadBracket;

else if (chrFSelection == '+') enumFSelectionCharType = TCharType.Plus;

else if (chrFSelection == '-') enumFSelectionCharType = TCharType.Minus;

else if (chrFSelection == ',') enumFSelectionCharType = TCharType.Comma;

else if (chrFSelection == '.') enumFSelectionCharType = TCharType.Dot;

else if (chrFSelection == '^' || chrFSelection == '%' || chrFSelection == '@' || chrFSelection == '<' || chrFSelection == '>' || chrFSelection == '?')

enumFSelectionCharType = TCharType.AnotherSymbol; // Считаем эти символы как другие (AnotherSymbol)

else enumFSelectionCharType = TCharType.NoInd; // Символ не распознан

enumFState = TState.Continue; // Продолжаем анализ

}

intFSourceColSelection++; // Переходим к следующему символу

}

// Метод добавления символа к текущей лексической единице

private void TakeSymbol()

{

char[] c = { chrFSelection }; // Создаем массив символов с текущим символом

String s = new string(c); // Преобразуем массив символов в строку

strFLexicalUnit += s; // Добавляем символ к текущей лексической единице

GetSymbol(); // Получаем следующий символ

}

// Метод перехода к следующему токену

public void NextToken()

{

strFLexicalUnit = ""; // Сбрасываем текущую лексическую единицу

// Начальная инициализация перед началом анализа

if (enumFState == TState.Start)

{

intFSourceRowSelection = 0; // Устанавливаем начальную строку

intFSourceColSelection = -1; // Устанавливаем начальный столбец

GetSymbol(); // Получаем первый символ

}

// Пропуск комментариев (если встречается '//')

if (chrFSelection == '/')

{

GetSymbol(); // Получаем следующий символ

if (chrFSelection == '/')

{

// Игнорируем все символы до конца строки

while (enumFSelectionCharType != TCharType.EndRow)

GetSymbol();

GetSymbol(); // Переходим к следующему символу после конца строки

}

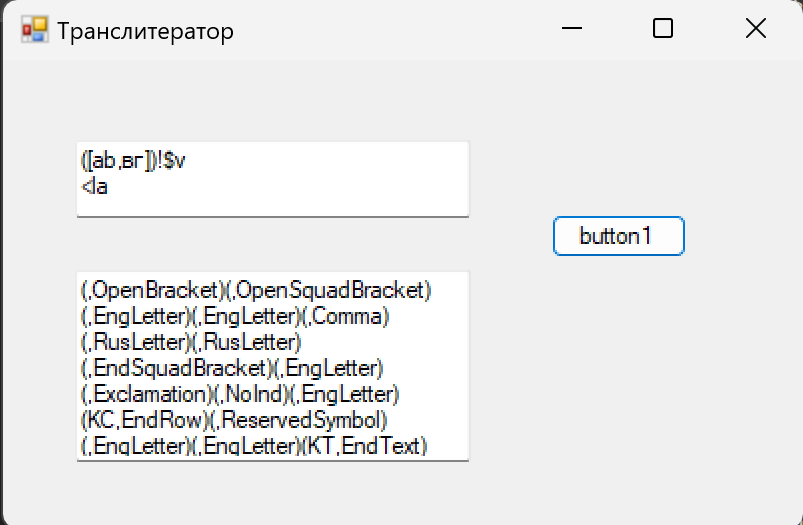
}

}

}

}

**Результаты тестирования:**



**Лабораторная работа № 2. Разработка лексического анализатора**

**Текст задания:**

1. Спроектировать и отладить экранную форму для ввода исходных данных, вывода сообщений программы и управления программой.
2. Включить из лабораторной работы № 1 транслитератор **void GetSymbol().**
3. Составить регулярную грамматику для каждого вида слов.
4. Построить конечные автоматы для каждого вида слов, как правило, они будут недетерминированными.
5. Построить детерминированные конечные автоматы для каждого вида слов.
6. Составить объединенный конечный автомат.
7. Написать и отладить модуль лексического анализатора по алгоритму объединенного конечного автомата. Для чтения исходного текста использовать транслитератор. Предусмотреть обработчик лексических ошибок исходного текста, используется конструкция **try … catch**.
8. Для отладки лексического анализатора временно включить в обработчик нажатия кнопки цикл чтения слов исходного текста и вывода результатов лексического анализа.

|  |  |  |
| --- | --- | --- |
| (001)\*101(110)\* | (a|b|c|d)+ | Длина не более 4 |

**Первое слово:**

(001)\*101(110)\*

A → 0B | 1C

B → 0D

C → 0Е

D → 1А

E → 1 | 1F

F → 1G

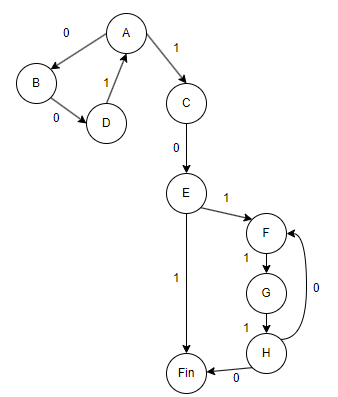
G → 1H

H → 0 | 0F

**Недетерминированная матрица:**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| A | B | C |
| B | D |  |
| C | E |  |
| D |  | A |
| E |  | F,Fin |
| F |  | G |
| G |  | H |
| H | F,Fin |  |
| Fin |  |  |

**Граф:**

****

**Детерминированная матрица:**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| A | B | C |
| B | D |  |
| C | E |  |
| D |  | A |
| E |  | FFin |
| FFin |  | G |
| G |  | H |
| H | FFin |  |

**Второе слово:**

(a|b|c|d)+

Длина не более 4

A → a | b | c | d | aB | bB | cB | dB

B → a | b | c | d | aC | bC | cC | dC

C → a | b | c | d | aD | bD | cD | dD

D → a | b | c | d

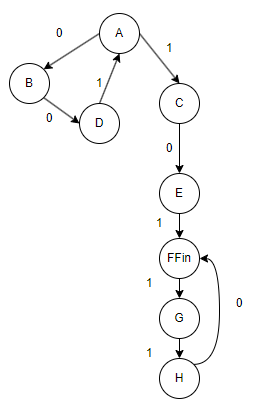
**Недетерминированная матрица:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | a | b | c | d |
| A | B,Fin | B,Fin | B,Fin | B,Fin |
| B | C,Fin | C,Fin | C,Fin | C,Fin |
| C | D,Fin | D,Fin | D,Fin | D,Fin |
| D | Fin | Fin | Fin | Fin |
| Fin |  |  |  |  |

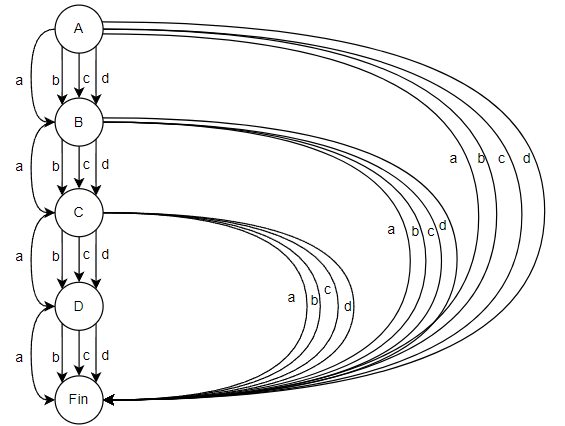
**Детерминированная матрица:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | a | b | c | d |
| A | BFin | BFin | BFin | BFin |
| BFin | CFin | CFin | CFin | CFin |
| CFin | D | D | D | D |
| D | Fin | Fin | Fin | Fin |
| Fin |  |  |  |  |

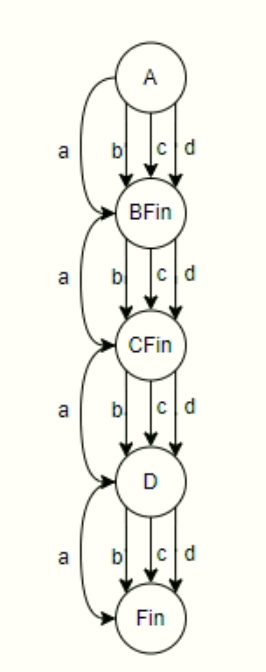
**Граф:**



**Граф:**



**Граф:**



**Код программы:**

**Form1.cs**

using System;

using System.Collections.Generic;

using System.ComponentModel;

using System.Data;

using System.Drawing;

using System.Text;

using System.Windows.Forms;

namespace Translator

{

public partial class Form1 : Form

{

public Form1()

{

InitializeComponent();

int n = tbFSource.Lines.Length; // Получение количества строк в tbFSource

}

// Обработчик события нажатия на кнопку btnFStart

private void btnFStart\_Click(object sender, EventArgs e)

{

// Создание экземпляра класса CLex для анализа текста

CLex Lex = new CLex();

Lex.strPSource = tbFSource.Lines; // Устанавливаем источник текста из текстового поля tbFSource

Lex.strPMessage = tbFMessage.Lines; // Устанавливаем сообщения из текстового поля tbFMessage

int x = tbFSource.TextLength; // Получаем количество символов в tbFSource

int y = tbFSource.Lines.Length; // Получаем количество строк в tbFSource

tbFMessage.Text = ""; // Очищаем tbFMessage для нового результата

try

{

// Цикл продолжается, пока состояние парсера не станет "Finish"

while (Lex.enumPState != TState.Finish)

{

Lex.GetSymbol(); // Получение текущего символа из исходного текста

Lex.NextToken(); // Получение следующего токена

String s = ""; // Переменная для хранения литеры

String s1 = ""; // Переменная для хранения типа литеры

// Определение типа символа и присваивание значений переменным s и s1

//switch (Lex.enumFSelectionCharType)

//{

// case TCharType.EngLetter: { s1 = "EngLetter"; break; }

// case TCharType.RusLetter: { s1 = "RusLetter"; break; }

// case TCharType.Digit: { s1 = "Digit"; break; }

// case TCharType.Space: { s1 = "Space"; break; }

// case TCharType.Star: { s1 = "Star"; break; }

// case TCharType.Exclamation: { s1 = "Exclamation"; break; }

// case TCharType.Equal: { s1 = "Equal"; break; }

// case TCharType.Semicolon: { s1 = "Semicolon"; break; }

// case TCharType.EndBracket: { s1 = "EndBracket"; break; }

// case TCharType.OpenBracket: { s1 = "OpenBracket"; break; }

// case TCharType.OpenSquadBracket: { s1 = "OpenSquadBracket"; break; }

// case TCharType.EndSquadBracket: { s1 = "EndSquadBracket"; break; }

// case TCharType.Colon: { s1 = "Colon"; break; }

// case TCharType.Minus: { s1 = "Mius"; break; }

// case TCharType.Plus: { s1 = "Plus"; break; }

// case TCharType.Comma: { s1 = "Comma"; break; }

// case TCharType.Dot: { s1 = "Dot"; break; }

// case TCharType.AnotherSymbol: { s1 = "ReservedSymbol"; break; }

// case TCharType.NoInd: { s1 = "NoInd"; break; }

// case TCharType.EndRow: { s = "KC"; s1 = "EndRow"; break; }

// case TCharType.EndText: { s = "KT"; s1 = "EndText"; break; }

//}

switch (Lex.enumPToken)

{

case TToken.lxmNumber: { s = "LxmNumber"; s1 = Lex.strPLexicalUnit; break; }

case TToken.lxmIdentifier: { s = "lxmId"; s1 = Lex.strPLexicalUnit; break; }

}

// Создание строки с литерой и ее типом

String m = "(" + s + "," + s1 + ")";

// Добавление строки в tbFMessage

tbFMessage.Text += m;

}

}

catch (Exception exc) // Обработка исключений

{

// Добавление сообщения об ошибке в tbFMessage

tbFMessage.Text += exc.Message;

tbFSource.Select(); // Устанавливаем фокус на tbFSource

tbFSource.SelectionStart = 0; // Устанавливаем начальную позицию выделения

int n = 0;

// Подсчет количества символов для выделения текста до текущей позиции

for (int i = 0; i < Lex.intPSourceRowSelection; i++)

n += tbFSource.Lines[i].Length + 2; // +2 учитывает переход на новую строку

n += Lex.intPSourceColSelection; // Добавляем текущую позицию в строке

tbFSource.SelectionLength = n; // Устанавливаем длину выделения

}

}

}

}

**uLex.cs**

using System;

using System.Collections.Generic;

using System.Linq;

using System.Text;

using System.Threading.Tasks;

namespace Translator

{

// Перечисление состояний анализа

public enum TState { Start, Continue, Finish }; // Тип состояния

// Перечисление возможных типов символов

public enum TCharType { EngLetter, RusLetter, Digit, EndRow, EndText, Space, Star, Slash, Exclamation, Equal, Semicolon, AnotherSymbol, OpenBracket, EndBracket, Colon, OpenSquadBracket, EndSquadBracket, Plus, Minus, Comma, Dot, NoInd }; // Тип символа

// Перечисление типов токенов

public enum TToken { lxmIdentifier, lxmNumber, lxmUnknown, lxmEmpty, lxmLeftParenth, lxmRightParenth, lxmIs, lxmDot, lxmComma };

// Класс лексического анализатора

public class CLex

{

// Поля класса

private String[] strFSource; // Массив строк, представляющий исходный текст

private String[] strFMessage; // Массив строк для сообщений (возможно, для вывода результатов)

public TCharType enumFSelectionCharType; // Тип текущего символа

public char chrFSelection; // Текущий символ

private TState enumFState; // Текущее состояние анализатора

private int intFSourceRowSelection; // Номер текущей строки

private int intFSourceColSelection; // Номер текущей колонки в строке

private String strFLexicalUnit; // Текущая лексическая единица

private TToken enumFToken; // Текущий токен

// Свойства для доступа к полям класса

public String[] strPSource { set { strFSource = value; } get { return strFSource; } }

public String[] strPMessage { set { strFMessage = value; } get { return strFMessage; } }

public TState enumPState { set { enumFState = value; } get { return enumFState; } }

public String strPLexicalUnit { set { strFLexicalUnit = value; } get { return strFLexicalUnit; } }

public TToken enumPToken { set { enumFToken = value; } get { return enumFToken; } }

public int intPSourceRowSelection { get { return intFSourceRowSelection; } set { intFSourceRowSelection = value; } }

public int intPSourceColSelection { get { return intFSourceColSelection; } set { intFSourceColSelection = value; } }

// Конструктор класса

public CLex()

{

}

// Метод получения текущего символа из источника

public void GetSymbol()

{

// Проверяем, не вышли ли за пределы строки

if (intFSourceColSelection > strFSource[intFSourceRowSelection].Length - 1)

{

intFSourceRowSelection++; // Переходим на следующую строку

if (intFSourceRowSelection <= strFSource.Length - 1)

{

// Если еще не конец текста, сбрасываем колонку и задаем символ конца строки

intFSourceColSelection = -1;

chrFSelection = '\0'; // Устанавливаем текущий символ в '\0'

enumFSelectionCharType = TCharType.EndRow; // Указываем, что достигнут конец строки

enumFState = TState.Continue; // Состояние анализа продолжается

}

else

{

// Если достигли конца всего текста

chrFSelection = '\0'; // Устанавливаем текущий символ в '\0'

enumFSelectionCharType = TCharType.EndText; // Указываем, что достигнут конец текста

enumFState = TState.Finish; // Меняем состояние на "Finish"

}

}

else

{

// Получаем текущий символ в строке

chrFSelection = strFSource[intFSourceRowSelection][intFSourceColSelection];

// Классифицируем символ

if (chrFSelection == ' ') enumFSelectionCharType = TCharType.Space;

else if (chrFSelection >= 'a' && chrFSelection <= 'z') enumFSelectionCharType = TCharType.EngLetter;

else if (chrFSelection >= 'а' && chrFSelection <= 'я') enumFSelectionCharType = TCharType.RusLetter;

else if (chrFSelection >= '0' && chrFSelection <= '9') enumFSelectionCharType = TCharType.Digit;

else if (chrFSelection == '/') enumFSelectionCharType = TCharType.Slash;

else if (chrFSelection == '\*') enumFSelectionCharType = TCharType.Star;

else if (chrFSelection == '!') enumFSelectionCharType = TCharType.Exclamation;

else if (chrFSelection == '=') enumFSelectionCharType = TCharType.Equal;

else if (chrFSelection == ';') enumFSelectionCharType = TCharType.Semicolon;

else if (chrFSelection == '(') enumFSelectionCharType = TCharType.OpenBracket;

else if (chrFSelection == ')') enumFSelectionCharType = TCharType.EndBracket;

else if (chrFSelection == ':') enumFSelectionCharType = TCharType.Colon;

else if (chrFSelection == '[') enumFSelectionCharType = TCharType.OpenSquadBracket;

else if (chrFSelection == ']') enumFSelectionCharType = TCharType.EndSquadBracket;

else if (chrFSelection == '+') enumFSelectionCharType = TCharType.Plus;

else if (chrFSelection == '-') enumFSelectionCharType = TCharType.Minus;

else if (chrFSelection == ',') enumFSelectionCharType = TCharType.Comma;

else if (chrFSelection == '.') enumFSelectionCharType = TCharType.Dot;

else if (chrFSelection == '^' || chrFSelection == '%' || chrFSelection == '@' || chrFSelection == '<' || chrFSelection == '>' || chrFSelection == '?')

enumFSelectionCharType = TCharType.AnotherSymbol; // Считаем эти символы как другие (AnotherSymbol)

else enumFSelectionCharType = TCharType.NoInd; // Символ не распознан

enumFState = TState.Continue; // Продолжаем анализ

}

intFSourceColSelection++; // Переходим к следующему символу

}

// Метод добавления символа к текущей лексической единице

private void TakeSymbol()

{

char[] c = { chrFSelection }; // Создаем массив символов с текущим символом

String s = new string(c); // Преобразуем массив символов в строку

strFLexicalUnit += s; // Добавляем символ к текущей лексической единице

GetSymbol(); // Получаем следующий символ

}

// Метод перехода к следующему токену

public void NextToken()

{

strFLexicalUnit = ""; // Сбрасываем текущую лексическую единицу

char[] allowedChars = { 'a', 'b', 'c', 'd' };

// Начальная инициализация перед началом анализа

if (enumFState == TState.Start)

{

intFSourceRowSelection = 0; // Устанавливаем начальную строку

intFSourceColSelection = -1; // Устанавливаем начальный столбец

GetSymbol(); // Получаем первый символ

}

// Пропуск комментариев (если встречается '//')

if (chrFSelection == '/')

{

GetSymbol(); // Получаем следующий символ

if (chrFSelection == '/')

{

// Игнорируем все символы до конца строки

while (enumFSelectionCharType != TCharType.EndRow)

GetSymbol();

GetSymbol(); // Переходим к следующему символу после конца строки

}

}

// Variant 13

switch (enumFSelectionCharType)

{

case TCharType.EngLetter:

{

// a b c d

// A |BFin|BFin|BFin|BFin|

// BFin |CFin|CFin|cFin|CFin|

// CFin | D | D | D | D |

// D |Fin | Fin| Fin| Fin|

// Fin | | | | |

A:

{

if (allowedChars.Contains(chrFSelection))

{

TakeSymbol();

goto BFin;

}

else

{

enumFToken = TToken.lxmIdentifier;

return;

}

}

BFin:

{

if (allowedChars.Contains(chrFSelection))

{

TakeSymbol();

goto CFin;

}

else

{

enumFToken = TToken.lxmIdentifier;

return;

}

}

CFin:

{

if (allowedChars.Contains(chrFSelection))

{

TakeSymbol();

goto D;

}

else

{

enumFToken = TToken.lxmIdentifier;

return;

}

}

D:

{

if (allowedChars.Contains(chrFSelection))

{

TakeSymbol();

goto Fin;

}

else

{

enumFToken = TToken.lxmIdentifier;

return;

}

}

Fin:

{

enumFToken = TToken.lxmIdentifier;

return;

}

}

if (chrFSelection == '/')

{

GetSymbol();

if (chrFSelection == '/')

while (enumFSelectionCharType != TCharType.EndRow)

{

GetSymbol();

}

GetSymbol();

}

case TCharType.Digit:

{

// 0 1

// A | B | C |

// B | D | |

// C | E | |

// D | | A |

// E | |FFin |

// FFin | | G |

// G | H | |

// H | FFin| |

A:

if (chrFSelection == '0')

{

TakeSymbol();

goto B;

}

else if (chrFSelection == '1')

{

TakeSymbol();

goto C;

}

else throw new Exception("Ожидался 0 или 1");

B:

if (chrFSelection == '0')

{

TakeSymbol();

goto D;

}

else throw new Exception("Ожидался 0");

C:

if (chrFSelection == '0')

{

TakeSymbol();

goto E;

}

else throw new Exception("Ожидался 0");

D:

if (chrFSelection == '1')

{

TakeSymbol();

goto A;

}

else throw new Exception("Ожидалась 1");

E:

if (chrFSelection == '1')

{

TakeSymbol();

goto FFin;

}

else throw new Exception("Ожидалась 1");

FFin:

if (chrFSelection == '1')

{

TakeSymbol();

goto G;

}

else if (enumFSelectionCharType != TCharType.Digit) { enumFToken = TToken.lxmNumber; return; }

else throw new Exception("Ожидалась 1");

G:

if (chrFSelection == '1')

{

TakeSymbol();

goto H;

}

else throw new Exception("Ожидался 0");

H:

if (chrFSelection == '0')

{

TakeSymbol();

goto FFin;

}

else throw new Exception("Ожидался 0");

}

case TCharType.AnotherSymbol:

{

if (chrFSelection == '/')

{

GetSymbol();

if (chrFSelection == '/')

{

while (enumFSelectionCharType != TCharType.EndRow)

GetSymbol();

}

GetSymbol();

}

if (chrFSelection == '(')

{

enumFToken = TToken.lxmLeftParenth;

GetSymbol();

return;

}

if (chrFSelection == ')')

{

enumFToken = TToken.lxmRightParenth;

GetSymbol();

return;

}

break;

}

case TCharType.EndText:

{

enumFToken = TToken.lxmEmpty;

break;

}

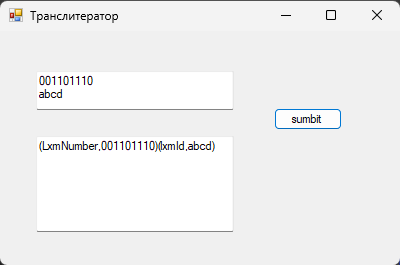
}

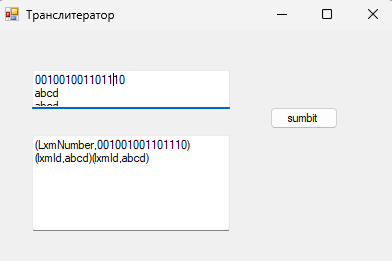
}

}

}

**Результаты тестирования:**

****

****

**1. Метод рекурсивного спуска**

1. **Инициализация**: установите входную строку и позицию чтения.
2. **Определение токенов**: создайте функции для распознавания разных токенов (числа, идентификаторы и т.д.).
3. **Сканирование**: вызывайте функции последовательно для нахождения токенов.
4. **Обработка ошибок**: если токен не найден, выбросите ошибку.

**2. Метод конечных автоматов**

1. **Определение состояний**: определите состояния для различных токенов.
2. **Определение переходов**: установите правила переходов между состояниями.
3. **Сканирование**: изменяйте состояния в зависимости от входных символов.
4. **Обработка ошибок**: если перехода нет, выбросите ошибку.

**3. Метод регулярных выражений**

1. **Определение шаблонов**: опишите регулярные выражения для токенов.
2. **Сканирование**: используйте регулярные выражения для нахождения токенов в строке.
3. **Обработка ошибок**: если совпадений нет, выбросите ошибку.

**Детерминированный конечный автомат (ДКА)** — это модель вычисления, которая в любой момент времени может находиться только в одном состоянии. Переходы между состояниями определяются детерминированно: для каждого состояния и входного символа есть ровно один переход.

**Недетерминированный конечный автомат (НДКА)** — это модель вычисления, которая может находиться в нескольких состояниях одновременно. Переходы между состояниями могут быть неопределёнными, то есть для одного состояния и входного символа может быть несколько возможных переходов, включая переходы по ε-переходам (переходам без чтения символа).

**Терминальный символ** — это символ, который принадлежит алфавиту языка и не может быть далее разложен на более простые символы. Эти символы являются конечными единицами входного текста, которые обрабатываются в процессе лексического анализа или транслитерации.

**Нетерминальный символ** — это символ, который используется в грамматике для обозначения групп символов или конструкций и может быть разложен на терминальные символы или другие нетерминальные символы. Нетерминальные символы помогают задавать структуру языка и используются для построения правил грамматики.

**Лабораторная работа № 3. Разработка контекстно-свободного (КС) синтаксического анализатора**

**Текст задания:**

Для предложенного преподавателем варианта КС-грамматики разработать методом рекурсивного спуска синтаксический анализатор

**Грамматика:**

S → A | S ; A

A → <2> : = [ B ]

B → <1> | <1> , B

**Освобождение от левой рекурсии:**

S → A | AC

C → ; A | ; C

A → <2> : = [ B ]

B → <1> | <1> , B

**Код программы:**

**Form1.cs**

using System;

using System.Collections.Generic;

using System.ComponentModel;

using System.Data;

using System.Drawing;

using System.Text;

using System.Windows.Forms;

using nsSynt;

using nsLex;

namespace Lab3

{

public partial class Form1 : Form

{

public Form1()

{

InitializeComponent();

tbFSource.AppendText("adc:=[000101110]" + "\r\n");

int n = tbFSource.Lines.Length;

}

private void btnFStart\_Click(object sender, EventArgs e)

{

tbFMessage.Clear();

uSyntAnalyzer Synt = new uSyntAnalyzer();

Synt.Lex.strPSource = tbFSource.Lines;

Synt.Lex.strPMessage = tbFMessage.Lines;

Synt.Lex.enumPState = TState.Start;

try

{

Synt.Lex.NextToken();

Synt.S();

throw new Exception("Текст верный");

}

catch (Exception exc)

{

tbFMessage.Text += exc.Message;

tbFSource.Select();

tbFSource.SelectionStart = 0;

int n = 0;

for (int i = 0; i < Synt.Lex.intPSourceRowSelection; i++) n += tbFSource.Lines[i].Length + 2;

n += Synt.Lex.intPSourceColSelection;

tbFSource.SelectionLength = n;

}

}

}

}

**uLex.cs**

using System;

using System.Collections.Generic;

using System.Text;

namespace nsLex

{

public enum TState { Start, Continue, Finish }; //тип состояния

public enum TCharType { Letter, Digit, EndRow, EndText, Space, ReservedSymbol }; // тип символа

public enum TToken { lxmIdentifier, lxmNumber, lxmUnknown, lxmEmpty, lxmLeftParenth, lxmRightParenth, lxmIs, lxmDot, lxmComma, lxmText, lxmtz, lxmdt, lxmr, lxmrs, lxmls};

public class CLex //класс лексический анализатор

{

private String[] strFSource; // указатель на массив строк

private String[] strFMessage; // указатель на массив строк

public TCharType enumFSelectionCharType;

public char chrFSelection;

private TState enumFState;

private int intFSourceRowSelection;

private int intFSourceColSelection;

private String strFLexicalUnit;

private TToken enumFToken;

public String[] strPSource { set { strFSource = value; } get { return strFSource; } }

public String[] strPMessage { set { strFMessage = value; } get { return strFMessage; } }

public TState enumPState { set { enumFState = value; } get { return enumFState; } }

public String strPLexicalUnit { set { strFLexicalUnit = value; } get { return strFLexicalUnit; } }

public TToken enumPToken { set { enumFToken = value; } get { return enumFToken; } }

public int intPSourceRowSelection { get { return intFSourceRowSelection; } set { intFSourceRowSelection = value; } }

public int intPSourceColSelection { get { return intFSourceColSelection; } set { intFSourceColSelection = value; } }

public CLex()

{

}

public void GetSymbol() //метод класса лексический анализатор

{

intFSourceColSelection++; // продвигаем номер колонки

if (intFSourceColSelection > strFSource[intFSourceRowSelection].Length - 1)

{

intFSourceRowSelection++;

if (intFSourceRowSelection <= strFSource.Length - 1)

{

intFSourceColSelection = -1;

chrFSelection = '\0';

enumFSelectionCharType = TCharType.EndRow;

enumFState = TState.Continue;

}

else

{

chrFSelection = '\0';

enumFSelectionCharType = TCharType.EndText;

enumFState = TState.Finish;

}

}

else

{

chrFSelection = strFSource[intFSourceRowSelection][intFSourceColSelection]; //классификация прочитанной литеры

if (chrFSelection == ' ') enumFSelectionCharType = TCharType.Space;

else if (chrFSelection >= 'a' && chrFSelection <= 'd') enumFSelectionCharType = TCharType.Letter;

else if (chrFSelection == '0' || chrFSelection == '1') enumFSelectionCharType = TCharType.Digit;

else if (chrFSelection == '/') enumFSelectionCharType = TCharType.ReservedSymbol;

else if (chrFSelection == '\*') enumFSelectionCharType = TCharType.ReservedSymbol;

else if (chrFSelection == ';' || chrFSelection == ',' || chrFSelection == '[' || chrFSelection == ']' || chrFSelection == '=' || chrFSelection == ':') enumFSelectionCharType = TCharType.ReservedSymbol;

else throw new System.Exception("Cимвол вне алфавита");

enumFState = TState.Continue;

}

}

private void TakeSymbol()

{

char[] c = { chrFSelection };

String s = new string(c);

strFLexicalUnit += s;

GetSymbol();

}

public void NextToken()

{

strFLexicalUnit = "";

if (enumFState == TState.Start)

{

intFSourceRowSelection = 0;

intFSourceColSelection = -1;

GetSymbol();

}

while (enumFSelectionCharType == TCharType.Space || enumFSelectionCharType == TCharType.EndRow)

{

GetSymbol();

}

if (chrFSelection == '/')

{

GetSymbol();

if (chrFSelection == '/')

while (enumFSelectionCharType != TCharType.EndRow)

{

GetSymbol();

}

GetSymbol();

}

// Вариант 6

switch (enumFSelectionCharType)

{

case TCharType.Letter:

{

// a b c d

// A | B | | | |

// B | | | |CFin|

// CFin |CFin|CFin|CFin|CFin|

A:

{

if (chrFSelection == 'a')

{

TakeSymbol();

goto B;

}

else throw new Exception("Слово должно начинаться с 'ad'");

}

B:

{

if (chrFSelection == 'd')

{

TakeSymbol();

goto CFin;

}

else throw new Exception("Слово должно начинаться с 'ad'");

}

CFin:

{

if (chrFSelection == 'a' || chrFSelection == 'b' || chrFSelection == 'c' || chrFSelection == 'd')

{

TakeSymbol();

goto CFin;

}

else

{

enumFToken = TToken.lxmIdentifier;

return;

}

}

}

if (chrFSelection == '/')

{

GetSymbol();

if (chrFSelection == '/')

while (enumFSelectionCharType != TCharType.EndRow)

{

GetSymbol();

}

GetSymbol();

}

case TCharType.Digit:

{

// 0 1

// A | B | C |

// B | D | |

// C | E | |

// D | A | |

// E | |FFin |

// FFin | | G |

// G | H | |

A:

if (chrFSelection == '0')

{

TakeSymbol();

goto B;

}

else if (chrFSelection == '1')

{

TakeSymbol();

goto C;

}

else throw new Exception("Ожидался 0 или 1");

B:

if (chrFSelection == '0')

{

TakeSymbol();

goto D;

}

else throw new Exception("Ожидался 0");

C:

if (chrFSelection == '0')

{

TakeSymbol();

goto E;

}

else throw new Exception("Ожидался 0");

D:

if (chrFSelection == '0')

{

TakeSymbol();

goto A;

}

else throw new Exception("Ожидался 0");

E:

if (chrFSelection == '1')

{

TakeSymbol();

goto FFin;

}

else throw new Exception("Ожидалась 1");

FFin:

if (chrFSelection == '1')

{

TakeSymbol();

goto G;

}

else if (enumFSelectionCharType != TCharType.Digit) { enumFToken = TToken.lxmNumber; return; }

else throw new Exception("Ожидалась 1");

G:

if (chrFSelection == '1')

{

TakeSymbol();

goto H;

}

else throw new Exception("Ожидался 0");

H:

if (chrFSelection == '0')

{

TakeSymbol();

goto FFin;

}

else throw new Exception("Ожидался 0");

}

case TCharType.ReservedSymbol:

{

if (chrFSelection == '/')

{

GetSymbol();

if (chrFSelection == '/')

{

while (enumFSelectionCharType != TCharType.EndRow)

GetSymbol();

}

GetSymbol();

}

if (chrFSelection == '(')

{

enumFToken = TToken.lxmLeftParenth;

GetSymbol();

return;

}

if (chrFSelection == ')')

{

enumFToken = TToken.lxmRightParenth;

GetSymbol();

return;

}

if (chrFSelection == '[')

{

enumFToken = TToken.lxmls;

GetSymbol();

return;

}

if (chrFSelection == ']')

{

enumFToken = TToken.lxmrs;

GetSymbol();

return;

}

if (chrFSelection == ',')

{

enumFToken = TToken.lxmComma;

GetSymbol();

return;

}

if (chrFSelection == ':')

{

enumFToken = TToken.lxmdt;

GetSymbol();

return;

}

if (chrFSelection == '=')

{

enumFToken = TToken.lxmr;

GetSymbol();

return;

}

break;

}

case TCharType.EndText:

{

enumFToken = TToken.lxmEmpty;

break;

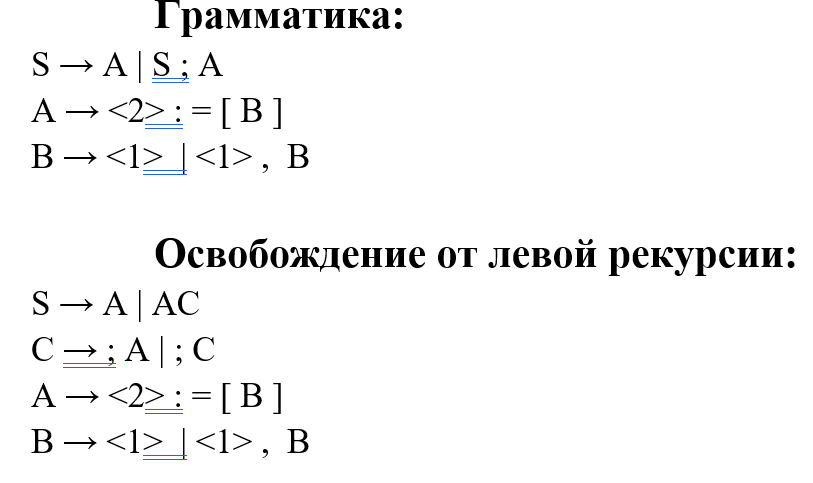
}

}

}

}

}



**uSyntAnalyzer.cs**

using System;

using System.Collections.Generic;

using System.Text;

using nsLex;

namespace nsSynt

{

class uSyntAnalyzer

{

private String[] strFSource;

private String[] strFMessage;

public String[] strPSource { set { strFSource = value; } get { return strFSource; } }

public String[] strPMessage { set { strFMessage = value; } get { return strFMessage; } }

public CLex Lex = new CLex();

public void S()

{

A();

if (Lex.enumPToken == TToken.lxmtz)

C();

throw new Exception("Конец слова, текст верный. Для продолжения ожидается ;");

}

public void C()

{

if (Lex.enumPToken == TToken.lxmtz)

{

Lex.NextToken();

A();

if (Lex.enumPToken == TToken.lxmtz)

C();

}

}

public void A()

{

if (Lex.enumPToken == TToken.lxmIdentifier)

{

Lex.NextToken();

if (Lex.enumPToken == TToken.lxmdt)

{

Lex.NextToken();

if (Lex.enumPToken == TToken.lxmr)

{

Lex.NextToken();

if (Lex.enumPToken == TToken.lxmls)

{

Lex.NextToken();

B();

if (Lex.enumPToken == TToken.lxmrs)

{

Lex.NextToken();

}

else throw new Exception("Ожидалась , или ]");

}

else throw new Exception("Ожидалась [");

}

else throw new Exception("Ожидалось =");

}

else throw new Exception("Ожидалось :");

}

else throw new Exception("Ожидался идентификатор");

}

public void B()

{

if (Lex.enumPToken == TToken.lxmNumber)

{

Lex.NextToken();

if (Lex.enumPToken == TToken.lxmComma)

{

Lex.NextToken();

B();

}

}

else throw new Exception("Ожидался идентификатор");

}

}

}

**Результаты тестирования:**

