**Содержание**

[Введение 5](#_Toc27328844)

[Данный курсовой проект представляет собой транслятор для языка программирования GMP-2019. Данный язык предназначен для выполнения сравнений беззнаковых чисел и строк, и выполнения определенных действий в зависимости от результата. 5](#_Toc27328845)

[Задание на курсовой проект можно разделить на следующие задачи: 5](#_Toc27328846)

[- Разработка спецификации языка GMP-2019; 5](#_Toc27328847)

[- Разработка лексического анализатора; 5](#_Toc27328848)

[- Разработка синтаксического анализатора; 5](#_Toc27328849)

[- Разработка семантического анализатора; 5](#_Toc27328850)

[- Разработка генератора кода; 5](#_Toc27328851)

[- Тестирование транслятора. 5](#_Toc27328852)

[1. Спецификация языка программирования 6](#_Toc27328853)

[**1.1 Характеристика языка программирования** 6](#_Toc27328854)

[**1.2 Алфавит языка** 6](#_Toc27328855)

[**1.3 Применяемые сепараторы** 6](#_Toc27328856)

[**1.5 Типы данных** 7](#_Toc27328857)

[**1.6 Преобразование типов данных** 8](#_Toc27328858)

[**1.7 Идентификаторы** 8](#_Toc27328859)

[**1.8 Литералы** 8](#_Toc27328860)

[**1.9 Объявление данных и область видимости** 8](#_Toc27328861)

[**1.10 Инициализация данных** 9](#_Toc27328862)

[**1.11 Инструкции языка** 9](#_Toc27328863)

[**1.11 Операции языка** 10](#_Toc27328864)

[**1.12 Выражения и их вычисление** 10](#_Toc27328865)

[**1.13 Программные конструкции языка** 10](#_Toc27328866)

[**1.14 Область видимости идентификаторов** 11](#_Toc27328867)

[**1.15 Семантические проверки** 11](#_Toc27328868)

[**1.16 Распределение оперативной памяти на этапе выполнения** 11](#_Toc27328869)

[**1.17 Стандартная библиотека и её состав** 12](#_Toc27328870)

[**1.18 Ввод и вывод данных** 12](#_Toc27328871)

[**1.19 Точка** **входа** 12](#_Toc27328872)

[**1.20 Препроцессор** 12](#_Toc27328873)

[**1.21 Соглашения о вызовах** 13](#_Toc27328874)

[**1.22 Объектный код** 13](#_Toc27328875)

[**1.23 Классификация сообщений транслятора** 13](#_Toc27328876)

[**1.24 Контрольный пример** 14](#_Toc27328877)

[2. Структура транслятора 16](#_Toc27328878)

[**2.1 Компоненты транслятора, их назначение и принципы взаимодействия** 16](#_Toc27328879)

[**2.2 Перечень входных параметров транслятора** 16](#_Toc27328880)

[**2.3 Перечень протоколов, формируемых транслятором и их содержимое** 17](#_Toc27328881)

[3. Разработка лексического анализатора 18](#_Toc27328882)

[**3.1 Структура лексического анализатора** 18](#_Toc27328883)

[**3.2 Контроль входных символов** 18](#_Toc27328884)

[**3.3 Удаление избыточных символов** 19](#_Toc27328885)

[**3.4 Перечень ключевых слов, сепараторов, символов операций и соответствующих им лексем, регулярных выражений и конечных автоматов** 19](#_Toc27328886)

[**3.5 Основные структуры данных** 20](#_Toc27328887)

[**3.6 Структура и перечень сообщений лексического анализатора** 20](#_Toc27328888)

[**3.7 Принцип обработки ошибок** 20](#_Toc27328889)

[**3.8 Параметры лексического анализатора** 21](#_Toc27328890)

[**3.9 Алгоритм лексического анализа** 21](#_Toc27328891)

[**3.10 Контрольный пример** 21](#_Toc27328892)

[4. Разработка синтаксического анализатора 22](#_Toc27328893)

[**4.1 Структура синтаксического анализатора** 22](#_Toc27328894)

[**4.2 Контекстно-свободная грамматика, описывающая синтаксис языка** 22](#_Toc27328895)

[**4.3 Построение конечного магазинного автомата** 23](#_Toc27328896)

[**4.4 Основные структуры данных** 24](#_Toc27328897)

[**4.5 Описание алгоритма синтаксического разбора** 24](#_Toc27328898)

[**4.6 Структура и перечень сообщений синтаксического анализатора** 25](#_Toc27328899)

[**4.7 Параметры синтаксического анализатора и режимы его работы** 25](#_Toc27328900)

[**4.8 Принцип обработки ошибок** 25](#_Toc27328901)

[**4.9 Контрольный пример** 25](#_Toc27328902)

[5. Разработка семантического анализатора 26](#_Toc27328903)

[**5.1 Структура семантического анализатора** 26](#_Toc27328904)

[**5.2 Функции семантического анализатора** 26](#_Toc27328905)

[**5.3 Структура и перечень сообщений семантического анализатора** 26](#_Toc27328906)

[**5.4 Принцип обработки ошибок** 26](#_Toc27328907)

[**5.5 Контрольный пример** 26](#_Toc27328908)

[6. Вычисление выражений 28](#_Toc27328909)

[**6.1 Выражения, допускаемые языком** 28](#_Toc27328910)

[**6.2 Польская запись** 28](#_Toc27328911)

[7. Генерация кода 28](#_Toc27328912)

[**7.1 Структура генератора кода** 28](#_Toc27328913)

[**7.2 Представление типов данных в оперативной памяти** 28](#_Toc27328914)

[**7.3 Статическая библиотека** 29](#_Toc27328915)

[**7.4 Особенности алгоритма генерации кода** 29](#_Toc27328916)

[**7.5 Входные параметры генератора кода** 30](#_Toc27328917)

[**7.6 Контрольный пример** 30](#_Toc27328918)

[Приложение А 31](#_Toc27328919)

[Приложение Б 35](#_Toc27328920)

[Приложение В 37](#_Toc27328921)

[Приложение Г 38](#_Toc27328922)

[Приложение Д 40](#_Toc27328923)

[Приложение И 42](#_Toc27328924)

**Введение**

Данный курсовой проект представляет собой транслятор для языка программирования GMP-2019. Данный язык предназначен для выполнения сравнений беззнаковых чисел и строк, и выполнения определенных действий в зависимости от результата.

Задание на курсовой проект можно разделить на следующие задачи:

* Разработка спецификации языка GMP-2019;
* Разработка лексического анализатора;
* Разработка синтаксического анализатора;
* Разработка семантического анализатора;
* Разработка генератора кода;
* Тестирование транслятора.

# **1. Спецификация языка программирования**

## **1.1 Характеристика языка программирования**

Язык программирования GMP-2019 предназначен для выполнения сравнения целых чисел и операций над строками.

Язык программирования GMP-2019 является процедурным, строго типизированным, не объектно-ориентированным, компилируемым.

**1.2 Алфавит языка**

Алфавит языка GMP-2019 основан на кодировке ASCII. Таблица кодировки ASCII представлена на рисунке 1.1.

Исходный код может содержать символы латинского алфавита, цифры десятичной системы счисления, символы пробела, табуляции, перевода строки, спецсимволы [], символы операторов: “< > = !” и символы сепараторов: , ; {}()

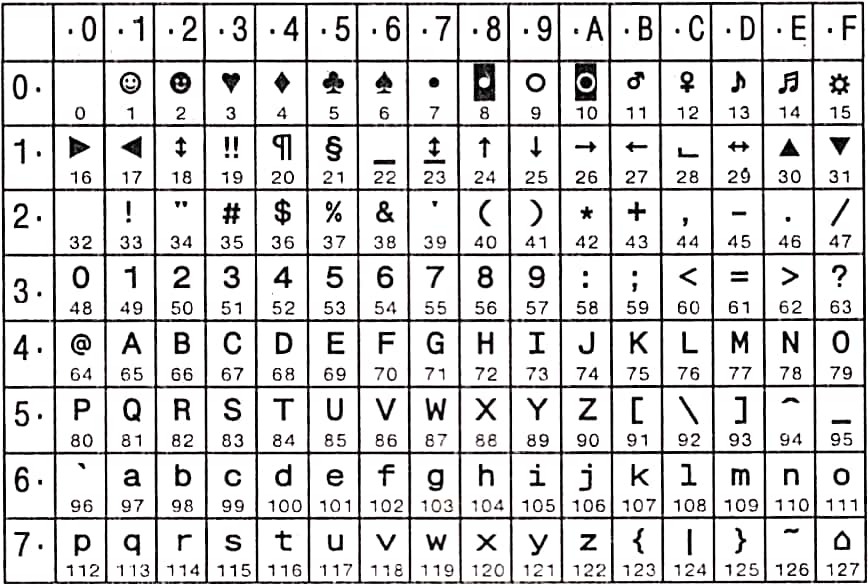


Рисунок 1.1 – Таблица кодировки ASCII

**1.3 Применяемые сепараторы**

Применяемые сепараторы в языке GMP-2019, приведены в таблице 1.1.

Таблица 1.1 — Применяемые сепараторы

|  |  |
| --- | --- |
| Сепаратор | Назначение сепаратора |
| ; | Разделитель инструкций |
| { } | Программный блок |
| ( ) | Параметры |
| пробел | Служит для разделения. Допускается везде, кроме идентификаторов и ключевых слов |
| , | Разделитель параметров в функции |

**1.4 Применяемые кодировки**

Для написания исходного кода на языке программирования GMP-2019 используется кодировка ASCII, которая представлена на рисунке 1.1.

## **1.5 Типы данных**

В языке GMP-2019 есть 2 типа данных: целочисленный и строковый. Описание типов данных, предусмотренных в данным языке представлено в таблице 1.2.

Таблица 1.2 – Типы данных языка GMP-2019

|  |  |
| --- | --- |
| Тип данных | Описание типа данных |
| ushort | Фундаментальный тип данных. Предусмотрен для объявления беззнаковых целочисленных данных (2 байта).  Автоматически инициализируется нулевым значением.  Операции:  <= бинарная операция сравнения меньше либо равно;  >= бинарная операция сравнения больше либо равно;  == бинарная операция сравнения на равенство;  != бинарная операция сравнения на не равенство;  < бинарная операция сравнения меньше;  > бинарная операция сравнения больше; |
| line | Является строковым типом данных. Предназначен для работы с символами, каждый символ занимает 1 байт. Максимальное количество символов – 255, включая символ окончания строки.  Инициализация по умолчанию: нулевой символ. |

## **1.6 Преобразование типов данных**

Язык GMP-2019 строго типизированный, преобразование типов данных не поддерживается.

## **1.7 Идентификаторы**

Общее количество идентификаторов ограничено максимальным размером таблицы идентификаторов. Идентификаторы должны начинаться только с символов латинского алфавита, могут содержать цифры. Максимальная длина идентификатора равна 8 символам. Идентификаторы, объявленные внутри функционального блока, получают префикс, идентичный имени функции, внутри которой они объявлены. Префикс занимает 8 дополнительных символов. В случае превышения заданной длины, идентификаторы усекаются до длины, равной 16 символов (8 символов на имя идентификатора, 8 символов на префикс). Данные правила действуют для всех типов идентификаторов.

<буква> ::= a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | t | u | v | w | x | y | z

<цифра> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

<идентификатор> ::= <буква>{ (<цифра> |<буква> ) }

## **1.8 Литералы**

В языке существует 2 типа литералов: целого и символьного типов. Краткое описание литералов представлено в таблице 1.3.

Таблица 1.3 – Описание литералов

|  |  |
| --- | --- |
| Тип литерала | Описание |
| Литералы целого типа | Целочисленные беззнаковые литералы, десятичное и шестнадцатеричное представления. Литералы только rvalue. |
| Строковые литералы | Состоит из символов латинского алфавита, заключенных в "…" (двойные кавычки). Только rvalue. |

<цифра> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

<шестнадцатеричное число> ::= <цифра> | a | b | c | d | e | f

<целочисленный литерал> ::= (0x<шестнадцатеричное число><шестнадцатеричное число><шестнадцатеричное число><шестнадцатеричное число>)|({<цифра>})

## **1.9 Объявление данных и область видимости**

В языке GMP-2019 объявление данных начинается с ключевого слова var, указывается тип данных и имя идентификатора. Требуется обязательное объявление переменной перед её использованием.

Примеры: var ushort a, var line b;

Все переменные в языке GMP-2019 имеют область видимости, а именно префикс — название функции, в которой они находятся, что разрешает использование в различных функциях переменных с одинаковым именем. Параметры функции видны только внутри неё. Переменные, объявленные в одной функции, недоступны в другой.

## **1.10 Инициализация данных**

В момент объявления переменных в языке GMP-2019 происходит автоматическая инициализация в зависимости от типа данных. Инициализация другими значениями в момент объявления не допускается. Виды инициализации представлены в таблице 1.4.

Таблица 1.4 — Способы инициализации переменных

|  |  |
| --- | --- |
| Вид инициализации | Примечание |
| <идентификатор> = <значение>; | Присваивание переменной значения. |

## **1.11 Инструкции языка**

Инструкции языка GMP-2019 представлена в таблице 1.5.

Таблица 1.5 – Инструкции языка

|  |  |
| --- | --- |
| Инструкция | Форма записи |
| Объявление переменной | var <тип данных> <идентификатор> |
| Объявление функции | <тип данных> function <идентификатор> ({<тип данных> <идентификатор>{,}})  {  < программный блок>  return <идентификатор>|<литерал>.  } |
| Вызов функции | <идентификатор> (<идентификатор>, …) |
| Присвоение значения | <идентификатор> = <значение>; |
| Печать данных | print <литерал>|<идентификатор> |
| Функции стандартной библиотеки  (применяются при инициализации и в выражениях) | compare(line, line) — лексикографическое сравнение строк  pow(ushort, ushort) — возводит первый операнд типа ushort в степень, равную значению второго операнда типа ushort. (перенести) |
| Возвращаемое значение | return <литерал>|<идентификатор> |

## **Операции языка**

Операции сравнения, используются в условной конструкции, которые можно использовать в языке GMP-2019, представлены в таблице 1.6.

Таблица 1.6 — Операции языка

|  |  |
| --- | --- |
| Инструкция | Форма записи |
| Операции сравнения языка | > — бинарное больше  < — бинарное меньше  >= — бинарное больше либо равно  <= — бинарное меньше либо равно  == — бинарное равно  != — бинарное не равно |

## **Выражения и их вычисление**

В языке присутствуют выражения сравнения, использующиеся в условной конструкции

## **Программные конструкции языка**

Основные программные конструкции языка GMP-2019 представлены в таблице 1.7.

Таблица 1.7 — Основные конструкции языка

|  |  |
| --- | --- |
| Конструкция | Реализация |
| Главная функция  (точка входа) | main  {  <программный блок>  } |
| Функции | <тип данных> function <идентификатор>  (<тип данных> <идентификатор>, …)  {  <программный блок>  return <идентификатор>|<литерал>.  } |

Продолжение таблицы 1.7

|  |  |
| --- | --- |
| Условный оператор | if(<литерал>|<идентификатор><логический оператор><литерал>|<идентификатор>)  {  <программный блок>  }  else  {  <программный блок>  } |

## **Область видимости идентификаторов**

Область видимости «сверху вниз». В GMP-2019 требуется обязательное объявление переменной перед её использованием. Все переменные должны находиться внутри программного блока языка. Имеется возможность объявления одинаковых переменных в разных функциях. Каждая переменная получает префикс – название функции, в которой она объявлена

## **Семантические проверки**

В языке программирования GMP-2019 выполняются следующие семантические проверки:

1. Наличие функции main – точки входа в программу;

2. Единственность точки входа;

3. Переопределение идентификаторов;

4. Использование идентификаторов без их объявления;

5. Проверка соответствия типа функции и возвращаемого параметра;

6. Правильность передаваемых в функцию параметров: количество, типы;

7. Правильность строковых выражений;

8. Превышение размера строковых и числовых литералов;

9. Правильность составленного условия цикла/условного оператора.

## **Распределение оперативной памяти на этапе выполнения**

Транслированный код использует две области памяти. В сегмент констант заносятся все литералы. В сегмент данных заносятся переменные и параметры функций. Локальная область видимости в исходном коде определяется за счет использования префиксов.

## **Стандартная библиотека и её состав**

В языке GMP-2019 предусмотрена стандартная библиотека. Функции, входящие в состав библиотеки, описаны в табл. 1.7. Стандартная библиотека подключается автоматически на этапе генерации кода.

Таблица 1.7 - Функции стандартной библиотеки языка GMP-2019

|  |  |
| --- | --- |
| Функция | Описание |
| compare(line a, line b) | Данная функция целочисленного типа, принимает два строковых параметра. Сравнивает строки и возвращает 1, если равны, 0, если a меньше b, 2, если a больше b |
| pow(ushort a, ushort b) | Данная функция целочисленного типа принимает два целочисленных параметра. Функция возводит число a в степень b и возвращает результат. |

## **Ввод и вывод данных**

Ввод данных языком программирования GMP-2019 не поддерживается.

Для вывода данных используется функция print(<имя идентификатора>|<литерал>). Пример: print(a);

## **Точка** **входа**

Функция точки входа в языке программирования GMP-2019 представлена в таблице 1.10.

Таблица 1.10 — Точка входа

|  |  |
| --- | --- |
| Конструкция | Реализация |
| Главная функция  (точка входа) | main  {  <программный блок>  } |

## **Препроцессор**

Препроцессор, принимающий и выдающий некоторые данные на вход транслятору, в языке GMP-2019 отсутствует.

## **Соглашения о вызовах**

В языке вызов функций происходит по соглашению о вызовах stdcall. Особенности stdcall:

– все параметры функции передаются через стек;

– память высвобождает вызываемый код;

– занесение в стек параметров идёт справа налево

## **Объектный код**

GMP-2019 транслируется в язык ассемблера.

## **Классификация сообщений транслятора**

Генерируемые транслятором сообщения определяют степень его информативности, то есть сообщения транслятора должны давать максимально полную информацию о допущенной пользователем ошибке при написании программы. Сообщения транслятора приведены в таблице 1.10, а также в приложении А.

Таблица 1.10 Классификация ошибок(диапазон)

|  |  |
| --- | --- |
| Префикс ошибки | Описание ошибки |
| Ошибка ###: LEX: | ### - код ошибки. Сообщение, генерируемое на этапе лексического анализа. Диапазон: 200-299 |
| Ошибка ###: SYN: | ### - код ошибки. Сообщение, генерируемое на этапе синтаксического анализа. Диапазон: 600-699 |
| Ошибка ###: SEM: | ### - код ошибки. Сообщение, генерируемое на этапе семантического анализа. Диапазон: 500-599 |
| Ошибка ###: SYS: | ### - код ошибки. Сообщение, генерируемое при критической ошибке. Диапазон: 0-99 |

## **Контрольный пример**

uselib "../Debug/StaticLib.lib"

ushort function min(ushort a, ushort b)

{

var ushort result;

if (b < a) {

result = b;

} else {

result = a;

}

return result;

}

line function ismore(ushort a, ushort b)

{

var line result;

if (a > b) {

result = "First argument more than second";

} else {

result = "Second argument more than first";

}

return result;

}

main

{

[var line s234;]

var line str1;

var line str2;

var ushort a;

var ushort b;

var ushort c;

var ushort result;

b = 0x0011;

c = 18;

str1 = "Mikhail";

str2 = str1;

result = compare(str1, str2);

if(result == 1) {

a = min(b, c);

} else {

print("strings are not equal");

}

if(a >= b) {

b = pow(c, 2);

print(b);

}

str1 = ismore(b, 300);

print(str1);

return 0;

}

# **2. Структура транслятора**

## **2.1 Компоненты транслятора, их назначение и принципы взаимодействия**

Транслятор преобразует программу, написанную на языке GMP-2019 в программу на языке ассемблера. Компонентами транслятора являются лексический, синтаксический и семантический анализаторы, а также генератор кода на язык ассемблера. Принцип их взаимодействия представлен на рисунке 2.1.

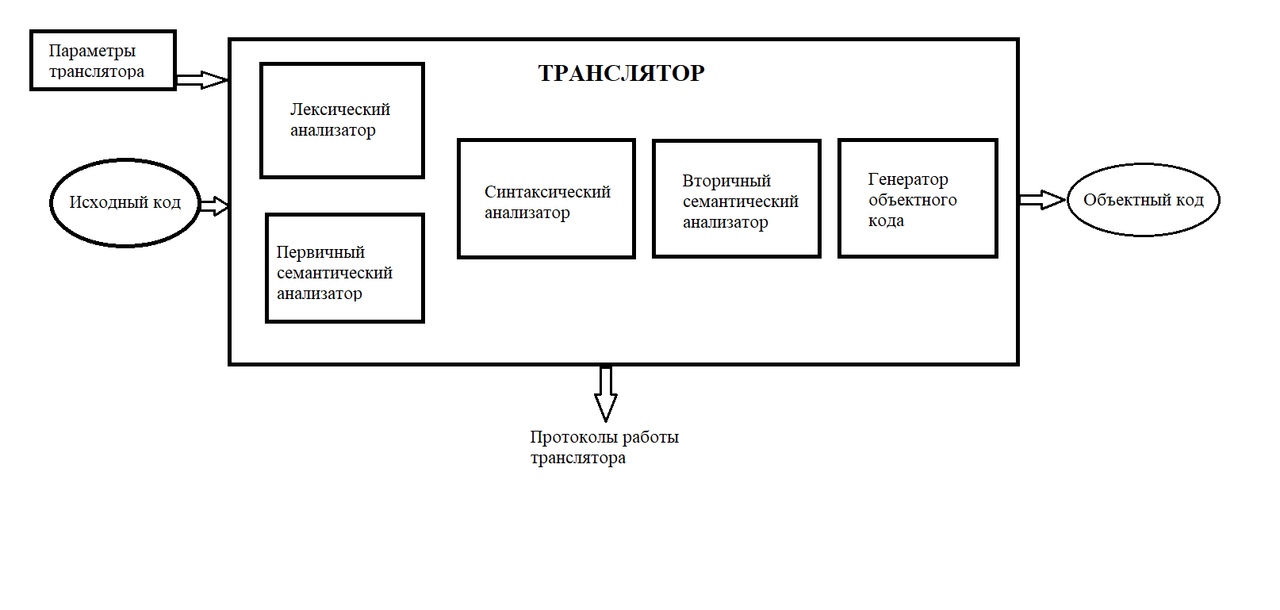


Рисунок 2.1 – Структура транслятора

Лексический анализ – первая фаза трансляции. Назначением лексического анализатора является нахождение ошибок лексики языка и формирование таблицы лексем и таблицы идентификаторов. Подробнее описан в 3 главе.

Синтаксический анализ – часть компилятора, выполняющая синтаксический анализ, то есть проверку исходного кода на соответствие правилам грамматики. Входной информацией для синтаксического анализа является таблица лексем и таблица идентификаторов. Выходной информацией является дерево разбора.

Семантический анализ в свою очередь является проверкой исходной программы на семантическую согласованность с определением языка, т.е. проверяет правильность текста исходной программы с точки зрения семантики. Подробное описание представлено в 5 главе.

Генератор кода – этап транслятора, выполняющий генерацию ассемблерного кода на основе полученных данных на предыдущих этапах трансляции. Генератор кода принимает на вход таблицы идентификаторов и лексем и транслирует код на языке GMP-2019, прошедший все предыдущие этапы, в код на языке Ассемблера. Более полно описан в главе 7.

## **2.2 Перечень входных параметров транслятора**

Входные параметры представлены в таблице 2.1.

Таблица 2.1 Входные параметры транслятора языка GMP-2019

|  |  |  |
| --- | --- | --- |
| Входной параметр | Описание параметра | Значение по умолчанию |
| -in:<путь к in-файлу> | Файл с исходным кодом на языке GMP-2019, имеющий расширение .txt | Не предусмотрено |
| -log:<путь к log-файлу> | Файл журнала для вывода протоколов работы программы. | Значение по умолчанию:  <имя in-файла>.log |
| -out:<путь к out-файлу> | Выходной файл – результат работы транслятора. Содержит исходный код на языке асемблера. | Значение по умолчанию:  <имя in-файла>.asm |
| -trace | Флаг для включения трассировки синтаксического анализа | выключен | |

## **2.3 Перечень протоколов, формируемых транслятором и их содержимое**

Таблица с перечнем протоколов, формируемых транслятором языка

GMP-2019 и их назначением представлена в таблице 2.2

Таблица 2.2 Протоколы, формируемые транслятором языка GMP-2019

|  |  |
| --- | --- |
| Формируемый протокол | Описание выходного протокола |
| Файл журнала, заданный параметром "-log:" | Файл с протоколом работы транслятора языка программирования GMP-2019. Содержит таблицу лексем и таблицу идентификаторов |
| Выходной файл, заданный параметром "-out:" | Результат работы программы – файл, содержащий исходный код на языке ассемблера. |

# **3. Разработка лексического анализатора**

## **3.1 Структура лексического анализатора**

Лексический анализатор – часть транслятора, выполняющая лексический анализ. Лексический анализатор принимает обработанный и разбитый на отдельные компоненты исходный код на языке GMP-2019. На выходе формируется таблица лексем и таблица идентификаторов. Структура лексического анализатора представлена на рисунке 3.1



Рисунок 3.1 Структура лексического анализатора GMP-2019

## **3.2 Контроль входных символов**

Таблица для контроля входных символов представлена на рисунке 3.2

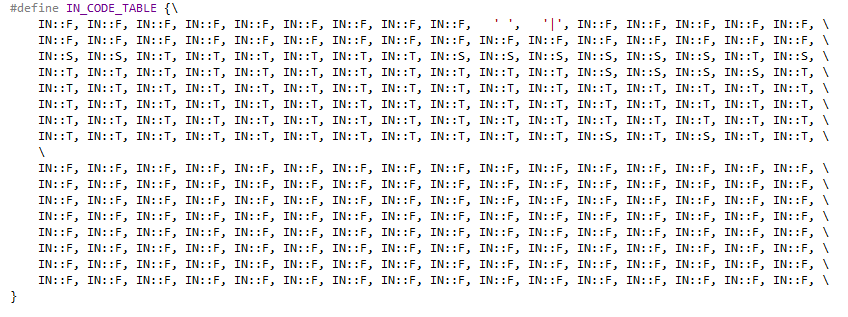


Рисунок 3.2. Таблица контроля входных символов

Принцип работы таблицы заключается в сопоставлении кода ASCII каждого входного символа значению в таблице.

Описание значения символов: T – разрешённый символ, F – запрещённый символ, S – сепаратор.

## **3.3 Удаление избыточных символов**

Избыточными символами являются символы табуляции и пробелы. Табуляция заменяется на пробел. Пробелы игнорируются на этапе разбиения исходного кода на слова.

Описание алгоритма удаления избыточных символов:

1. Посимвольно считываем файл с исходным кодом программы.

2. Пробелы выступаю в роли сепаратора и пропускаются.

## **3.4 Перечень ключевых слов, сепараторов, символов операций и соответствующих им лексем, регулярных выражений и конечных автоматов**

Лексемы – это символы, соответствующие ключевым словам, символам операций и сепараторам, необходимые для упрощения дальнейшей обработки исходного кода программы. Данное соответствие описано в таблице 3.1.

Таблица 3.1 Соответствие ключевых слов, символов операций и сепараторов с лексемами

|  |  |  |
| --- | --- | --- |
| Тип | Слова | Лексема |
| Ключевые слова | var | d |
| ushort | t |
| line | t |
| function | f |
| print | p |
| pow | b |
| compare | a |
| return | r |
| main | m |
| if | c |
| else | e |
| uselib | u |
| Иное | Идентификатор | i |
| Целочисленный литерал | l |
| Строковый литерал | l |

Продолжение таблицы 3.1

|  |  |  |
| --- | --- | --- |
| Сепараторы | ; | ; |
| , | , |
|  |  |
| { | { |
| } | } |
| ( | ( |
| ) | ) |
| Операторы | = | = |
| Логические операторы | o |

Пример реализации таблицы лексем представлен в приложении А.

Также в приложении А находятся некоторые конечные автоматы, соответствующие лексемам языка GMP-2019.

## **3.5 Основные структуры данных**

Основные структуры таблиц лексем и идентификаторов данных языка GMP-2019, используемых для хранения, представлены в приложении Б. В таблице лексем содержится лексема, её номер, полученный при разборе, номер строки в исходном коде. В таблице идентификаторов содержится имя идентификатора, номер в таблице лексем, тип данных, смысловой тип идентификатора и его значение.

## **3.6 Структура и перечень сообщений лексического анализатора**

Перечень сообщений лексического анализатора представлен на рисунке 3.3.

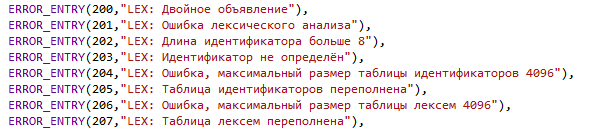


Рисунок 3.3 – Перечень ошибок лексического анализатора

## **3.7 Принцип обработки ошибок**

При возникновении ошибки транслятор завершает свою работу. Для обработки ошибок лексический анализатор использует таблицу с сообщениями. Структура сообщений содержит информацию о номере сообщения, номер строки и позицию, где было вызвано сообщение в исходном коде, информацию об ошибке.

## **3.8 Параметры лексического анализатора**

Входным параметром лексического анализа является структура, полученная после чтения входного файла на этапе проверки исходного кода на допустимость символов.

## **3.9 Алгоритм лексического анализа**

* проверяет входной поток символов программы на исходном языке на допустимость, удаляет лишние пробелы и добавляет сепаратор для вычисления номера строки для каждой лексемы;
* для выделенной части входного потока выполняется функция распознавания лексемы;
* при успешном распознавании информация о выделенной лексеме заносится в таблицу лексем и таблицу идентификаторов, и алгоритм возвращается к первому этапу;
* формирует протокол работы;
* при неуспешном распознавании выдается сообщение об ошибке.

Распознавание цепочек основывается на работе конечных автоматов. Работу конечного автомата можно проиллюстрировать с помощью графа переходов.

Пример. Регулярное выражение для ключевого слова main.

Граф конечного автомата для этой лексемы представлен на рисунке 3.4. S0 – начальное состояние, S4 – конечное состояние автомата.

n

m

a

i

Рисунок 3.4 – Граф переходов для цепочки ‘main’

## **3.10 Контрольный пример**

Результат работы лексического анализатора – таблицы лексем и идентификаторов – представлен в приложении А.

# **4. Разработка синтаксического анализатора**

## **4.1 Структура синтаксического анализатора**

Синтаксический анализ – это фаза трансляции, выполняемая после лексического анализа и предназначенная для распознавания синтаксических конструкций. Входом для синтаксического анализа является таблица лексем и таблица идентификаторов, полученные после фазы лексического анализа. Выходом – дерево разбора. Структура синтаксического анализатора представлена на рисунке  4.1.



Рисунок 4.1 – Структура синтаксического анализатора

## **4.2 Контекстно-свободная грамматика, описывающая синтаксис языка**

В синтаксическом анализаторе транслятора языка GMP-2019 используется контекстно-свободная грамматика , где

T – множество терминальных символов (было описано в разделе 1.2 данной пояснительной записки),

N – множество нетерминальных символов (первый столбец таблицы 4.1),

P – множество правил языка (второй столбец таблицы 4.1),

S – начальный символ грамматики, являющийся нетерминалом.

Эта грамматика имеет нормальную форму Грейбах, т.к. она не леворекурсивная (не содержит леворекурсивных правил) и правила  имеют вид:

1. , где ; (или , или )
2. , где — начальный символ, при этом если такое правило существует, то нетерминал  не встречается в правой части правил.

Грамматика языка GMP-2019 представлена в приложении B.

TS – терминальные символы, которыми являются сепараторы, знаки арифметических операций и некоторые строчные буквы, которые приведены в разделе 1.2

NS – нетерминальные символы, представленные несколькими заглавными буквами латинского алфавита.

Таблица 4.1 – Перечень правил, составляющих грамматику языка и описание нетерминальных символов GMP-2019

|  |  |  |
| --- | --- | --- |
| Нетерминал | Цепочки правил | Описание |
| S | S→m{NrE;};| tfi(F){NrE;};S| ul | ulS | Порождает правила, описывающее общую структуру программы |
| N | N→dti;|i=E;|dti;N| i=E;N|p(l);N| p(i);N|p(l);| p(i);|  c(C){N}e{N}N| c(C){N}e{N}|  c(C){N}| c(C){N}N | Порождает правила, описывающие инструкции языка |
| E | E→i|l|(E)|i(W)| b(W)| a(W)| | Порождает правила, описывающие выражения |
| F | F→ti|ti,F | Порождает правила, описывающие параметры локальной функции при её объявлении |
| W | W→i|l|i,W|l,W | Порождает правила, описывающие параметры вызываемой функции |
| C | С→ioi|iol|loi|lol | Порождает правила, описывающие условия оператора if |

## **4.3 Построение конечного магазинного автомата**

Конечный автомат с магазинной памятью представляет собой семерку, описание которой представлено в таблице 4.2. Структура данного автомата показана в приложении Г.

Таблица 4.2 – Описание компонентов магазинного автомата

|  |  |  |
| --- | --- | --- |
| Компонента | Определение | Описание |
|  | Множество состояний автомата | Состояние автомата представляет из себя структуру, содержащую позицию на входной ленте, номера текущего правила и цепочки и стек автомата |
|  | Алфавит входных символов | Алфавит является множеством терминальных и нетерминальных символов, описание которых содержится в разделе 1.2 и в таблице 4.1. |

Продолжение таблицы 4.2

|  |  |  |
| --- | --- | --- |
|  | Алфавит специальных магазинных символов | Алфавит магазинных символов содержит стартовый символ и маркер дна стека |
|  | Функция переходов автомата | Функция представляет из себя множество правил грамматики, описанных в таблице 4.1. |
|  | Начальное состояние автомата | Состояние, которое приобретает автомат в начале своей работы. Представляется в виде стартового правила грамматики (нетерминальный символ S) |
|  | Начальное состояние магазина автомата | Символ маркера дна стека ($) |
|  | Множество конечных состояний | Конечные состояние заставляют автомат прекратить свою работу. Конечным состоянием является пустой магазин автомата и совпадение позиции на входной ленте автомата с размером ленты |

## **4.4 Основные структуры данных**

Основные структуры данных синтаксического анализатора включают в себя структуру магазинного автомата и структуру грамматики Грейбах, описывающей правила языка GMP-2019. Данные структуры представлены в приложении В.

## **4.5 Описание алгоритма синтаксического разбора**

Принцип работы автомата:

1. В магазин записывается стартовый символ;
2. На основе полученных ранее таблиц формируется входная лента;
3. Запускается автомат;
4. Выбирается цепочка, соответствующая нетерминальному символу, записывается в магазин в обратном порядке;
5. Если терминалы в стеке и в ленте совпадают, то данный терминал удаляется из ленты и стека. Иначе возвращаемся в предыдущее сохраненное состояние и выбираем другую цепочку нетерминала;
6. Если в магазине встретился нетерминал, переходим к пункту 4;
7. Если наш символ достиг дна стека, и лента в этот момент пуста, то синтаксический анализ выполнен успешно. Иначе генерируется исключение.

## **4.6 Структура и перечень сообщений синтаксического анализатора**

Перечень сообщений синтаксического анализатора представлен на рисунке 4.1.

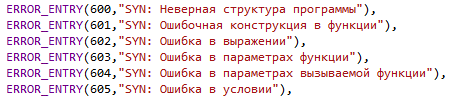


Рисунок 4.1 – Перечень сообщений синтаксического анализатора

## **4.7 Параметры синтаксического анализатора и режимы его работы**

Входным параметром синтаксического анализатора является таблица лексем, полученная на этапе лексического анализа, а также правила контекстно-свободной грамматики в форме Грейбах.

Выходными параметрами являются трассировка прохода таблицы лексем, при наличии разрешающего ключа -trace, и правила разбора, которые выводятся на консоль.

## **4.8 Принцип обработки ошибок**

Синтаксический анализатор выполняет разбор исходной последовательности лексем до тех пор, пока не дойдёт до конца цепочки лексем или не найдёт ошибку. Тогда анализ останавливается и выводится сообщение об ошибке (если она найдена). Если в процессе анализа находятся более трёх ошибок, то анализ останавливается. После всей процедуры трассировки в протокол будет выведено диагностическое сообщение.

## **4.9 Контрольный пример**

Пример разбора синтаксическим анализатором исходного кода на языке GMP-2019 представлен в приложении Д. Дерево разбора исходного кода также представлено в графическом материале.

# **5. Разработка семантического анализатора**

## **5.1 Структура семантического анализатора**

Семантический анализ происходит при выполнении фазы лексического анализа и реализуется в виде отдельных проверок текущих ситуаций в конкретных случаях: установки флага или нахождении в особом месте программы (оператор выхода из функции, оператор ветвления, вызов функции стандартной библиотеки). Так же некоторые проверки производятся после завершения лексического анализа.

## **5.2 Функции семантического анализатора**

Семантический анализатор выполняет проверку на основные правила языка (семантики языка), которые описаны в разделе 1.16.

## **5.3 Структура и перечень сообщений семантического анализатора**

Сообщения, формируемые семантическим анализатором, представлены на рисунке 5.1.

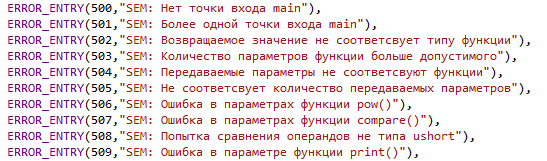


Рисунок 5.1 – Перечень сообщений семантического анализатора

## **5.4 Принцип обработки ошибок**

Принцип обработки ошибок идентичен принципу обработки ошибок на этапе лексического анализа (раздел 3.7).

## **5.5 Контрольный пример**

Соответствие примеров некоторых ошибок в исходном коде и диагностических сообщений об ошибках приведено в таблице 5.1.

Таблица 5.1. Примеры диагностики ошибок

|  |  |
| --- | --- |
| Текст ошибки | Исходный код |
| Ошибка 501: SEM: Более одной точки входа main, строка 0, позиция 0 | main  {  return 0;  }  main  {  return 0;  } |
| Ошибка 502: SEM: Возвращаемое значение не соответствует типу функции, строка 2, позиция 0 | line function func(ushort a){  return a;  }  main  {  return 0;  } |
| Ошибка 504: SEM: Передаваемые параметры не соответствуют функции, строка 7, позиция 0 | line function func(ushort a){  return a;  }  main  {  var line str;  str = func("b");  return 0;  } |

# **6. Вычисление выражений**

## **6.1 Выражения, допускаемые языком**

В языке GMP-2019 допускаются выражения, применимые к целочисленным типам данных. В выражениях поддерживаются логические операции, такие как < > == <= >= !=

## **6.2 Польская запись**

В языке GMP-2019 польская запись не используется.

# **7. Генерация кода**

## **7.1 Структура генератора кода**

Генерация объектного кода — это перевод компилятором внутреннего представления исходной программы в цепочку символов выходного языка. На вход генератора подаются таблицы лексем и идентификаторов, на основе которых генерируется файл с ассемблерным кодом.

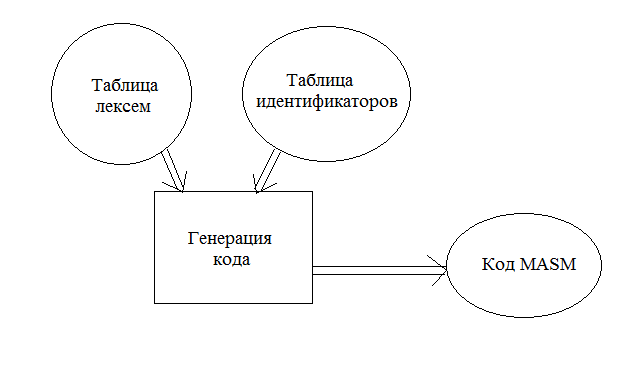


Рисунок 7.1 Структура генератора кода

## **7.2 Представление типов данных в оперативной памяти**

Элементы таблицы идентификаторов расположены в разных сегментах языка ассемблера – .data и .const. Идентификаторы языка GMP-2019 размещены в сегменте данных(.data). Литералы – в сегменте констант (.const). Соответствия между типами данных идентификаторов на языке GMP-2019 и на языке ассемблера приведены в таблице 7.1.

Таблица 7.1 – Соответствия типов идентификаторов языка GMP-2019 и языка Ассемблера

|  |  |  |
| --- | --- | --- |
| Тип идентификатора на языке GMP-2019 | Тип идентификатора на языке ассемблера | Пояснение |
| ushort | WORD | Хранит целочисленный тип данных со знаком. |
| line | DWORD | Хранит указатель на начало строки. |
| L(0-99) | BYTE  WORD | Литералы: символьные,  целочисленные |

## **7.3 Статическая библиотека**

В языке GMP-2019 предусмотрена статическая библиотека. Статическая библиотека содержит функции, написанные на языке C++. Объявление функций статической библиотеки генерируется автоматически в коде ассемблера. Объявление функций статической библиотеки генерируется автоматически.

Таблица 7.3 – Функции статической библиотеки

|  |  |
| --- | --- |
| Функция | Назначение |
| \_pow(ushort a, ushort b) | Возводит число a в степень b |
| \_printN(ushort a)  \_printS(line str) | Вывод на консоль целочисленной переменной a и строки str |
| \_compare(line a, line b) | Лексикографическое сравнение строк, если строка a меньше b возвращает 0, равна 1, больше 2 |

## **7.4 Особенности алгоритма генерации кода**

В языке GMP-2019 генерация кода строится на основе таблиц лексем и идентификаторов. Преобразования происходят по мере прохождения по таблицам. Функции статической библиотеки начинаются с нижнего подчёркивания для исключения их переопределения. Перед началом основной трансляции производится запись литералов и идентификаторов в сегменты констант и данных соответственно.

## **7.5 Входные параметры генератора кода**

На вход генератору кода поступают таблицы лексем и идентификаторов исходного код программы на языке GMP-2019. Результаты работы генератора кода выводятся в файл с расширением .asm

## **7.6 Контрольный пример**

Результат генерации ассемблерного кода на основе контрольного примера приведен в приложении И. Результат работы контрольного примера приведён на рисунке 7.1.



Рисунок 7.1 Результат работы программы на языке GMP-2019

# **8. Тестирование транслятора**

## **8.1 Общие положения**

При возникновении ошибки на каком-либо этапе трансляции, она обрабатывается в главном файле программы: ошибка выводится на консоль и записывается в файл логирования.

## **8.2 Результаты тестирования**

Результаты тестирования приведены в таблице 8.1

Таблица 8.1 – Результаты тестирования транслятора

|  |  |
| --- | --- |
| Исходный код | Диагностическое сообщение |
| э | Ошибка 111: SYS: Недопустимый символ в исходном файле (-in), строка 1, позиция 0 |
| main  {  var line str1;  var line str1;  } | Ошибка 200: LEX: Двойное объявление, строка 4, позиция 0 |
| main  {  A  } | Ошибка 201: LEX: Ошибка лексического анализа, строка 3, позиция 0 |
| var line a23456789; | Ошибка 202: LEX: Длина идентификатора больше 8, строка 1, позиция 0 |
| str = ""; | Ошибка 203: LEX: Идентификатор не определён, строка 1, позиция 0 |
| main  {  var ushort a;  a = 65536;  } | Ошибка 208: LEX: Максимальное значение типа ushort 65535, строка 4, позиция 0 |
| var ushort a; | Ошибка 500: SEM: Нет точки входа main, строка 0, позиция 0 |

Продолжение таблицы 8.1

|  |  |
| --- | --- |
| main  {  }  main  {  } | Ошибка 501: SEM: Более одной точки входа main, строка 0, позиция 0 |
| line function func()  {  var ushort a;  return 0;  }  main  {  var ushort a;  return 0;  } | Ошибка 502: SEM: Возвращаемое значение не соответствует типу функции, строка 4, позиция 0 |
| ushort function func(ushort a, ushort b, ushort c, ushort d, ushort e, ushort f)  {  a = 3;  return a;  }  main  {  var ushort a;  a = func(1, 2, 3, 4, 5, 6);  return 0;  } | Ошибка 503: SEM: Количество параметров функции больше допустимого, строка 10, позиция 0 |

Продолжение таблицы 8.1

|  |  |
| --- | --- |
| ushort function func(ushort a)  {  a = 3;  return a;  }  main  {  var ushort a;  a = func("1");  return 0;  } | Ошибка 504: SEM: Передаваемые параметры не соответствуют функции, строка 10, позиция 0 |
| ushort function func(ushort a)  {  a = 3;  return a;  }  main  {  var ushort a;  a = func(1, 2);  return 0;  } | Ошибка 505: SEM: Не соответсвует количество передаваемых параметров, строка 10, позиция 0 |

**Заключение**

В ходе выполнения курсовой работы был разработан транслятор и генератор кода для языка программирования PAA-2018 со всеми необходимыми компонентами. Таким образом, были выполнены основные задачи данной курсовой работы:

1. Сформулирована спецификация языка PAA-2018;
2. Разработаны конечные автоматы и важные алгоритмы на их основе для эффективной работы лексического анализатора;
3. Осуществлена программная реализация лексического анализатора, распознающего допустимые цепочки спроектированного языка;
4. Разработана контекстно-свободная, приведённая к нормальной форме Грейбах, грамматика для описания синтаксически верных конструкций языка;
5. Осуществлена программная реализация синтаксического анализатора;
6. Разработан семантический анализатор, осуществляющий проверку используемых инструкций на соответствие логическим правилам;
7. Разработан транслятор кода на язык ассемблера;
8. Проведено тестирование всех вышеперечисленных компонентов.

Окончательная версия языка PAA-2018 включает:

1. 2 типа данных;
2. Поддержка операторов вывода и перевода строки;
3. Возможность вызова функций стандартной библиотеки;
4. Наличие 4 арифметических операторов для вычисления выражений;
5. Поддержка функций, процедур, операторов цикла и условия;
6. Структурированная и классифицированная система для обработки ошибок пользователя.

Проделанная работа позволила получить необходимое представление о структурах и процессах, использующихся при построении трансляторов, а также основные различия и преимущества тех или иных средств трансляции.

**Список использованных источников**

1. Ахо, А. Компиляторы: принципы, технологии и инструменты / А. Ахо, Р. Сети, Дж. Ульман. – M.: Вильямс, 2003. – 768с.

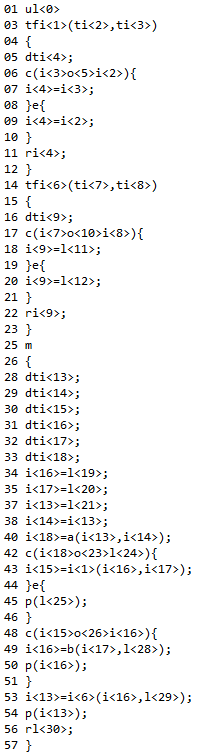
2. Ахо, А. Теория синтаксического анализа, перевода и компиляции /А. Ахо, Дж. Ульман. – Москва : Мир, 1998. – Т. 2 : Компиляция. - 487 с.

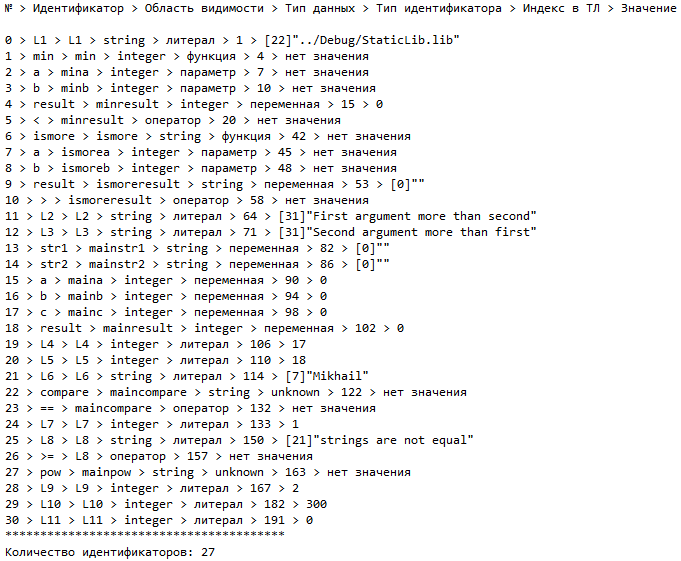
3. Герберт, Ш. Справочник программиста по C/C++ / Шилдт Герберт. - 3-е изд. – Москва : Вильямс, 2003. - 429 с.

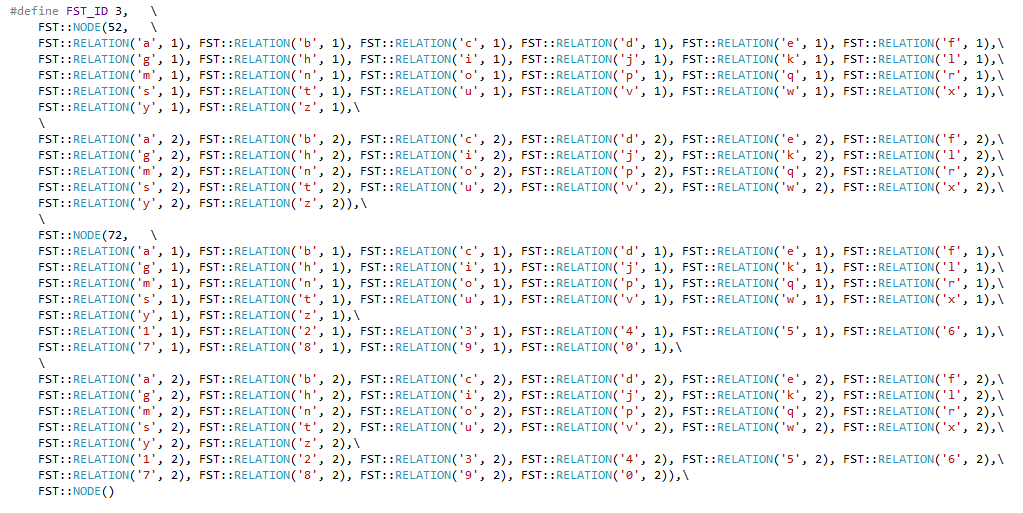
4. Прата, С. Язык программирования С++. Лекции и упражнения / С. Прата. – М., 2006 — 1104 c.

5. Страуструп, Б. Принципы и практика использования C++ / Б. Страуструп – 2009 – 1238 с

# **Приложение А**

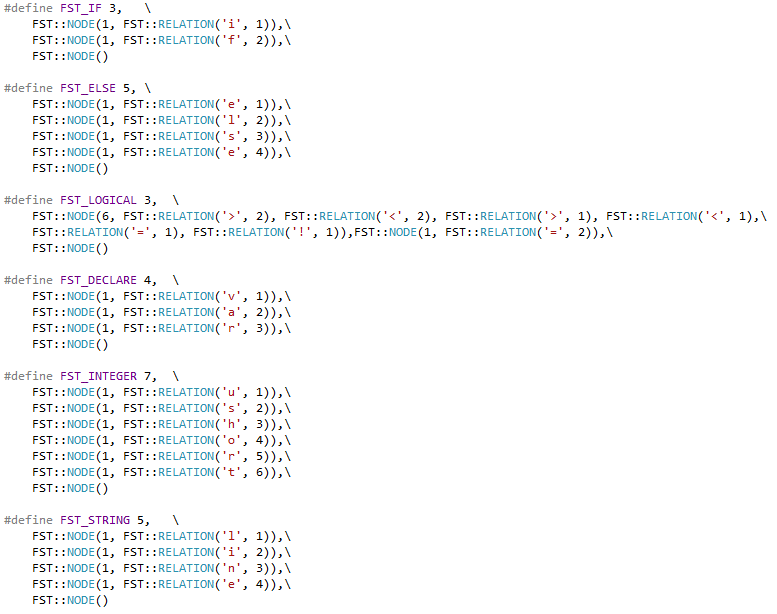






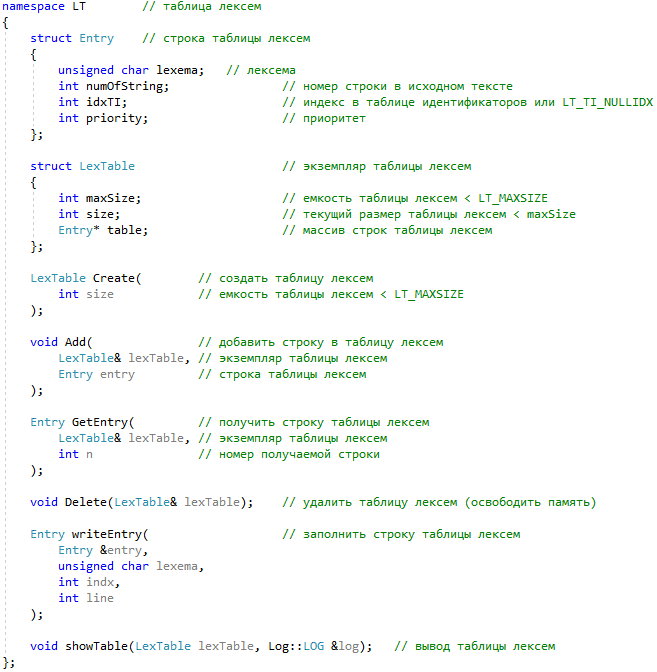


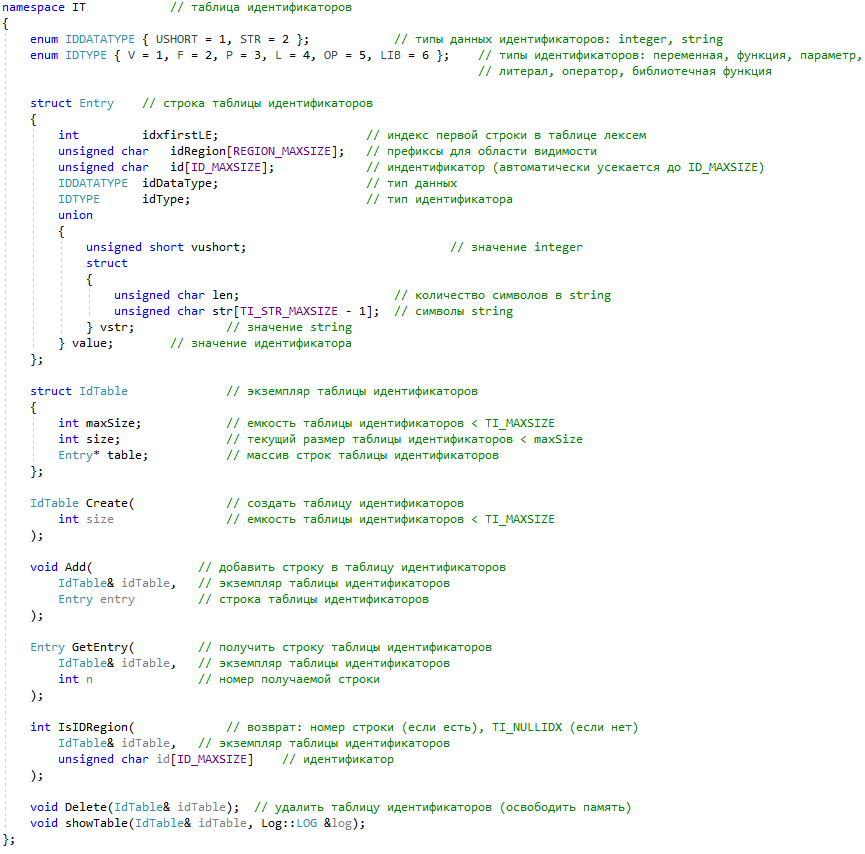




# **Приложение Б**

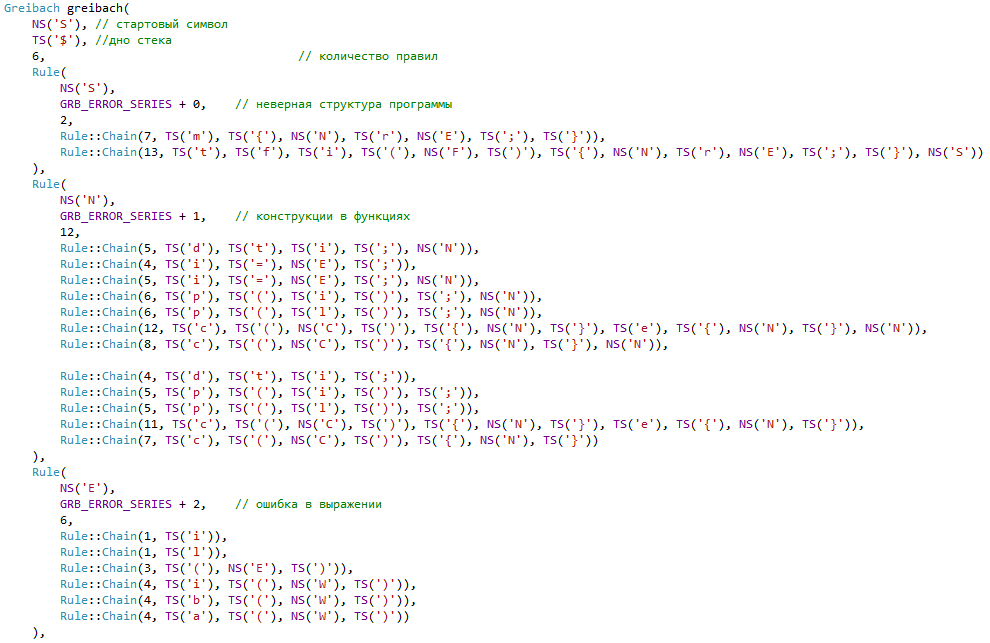
Основные структуры таблиц лексем и идентификаторов данных

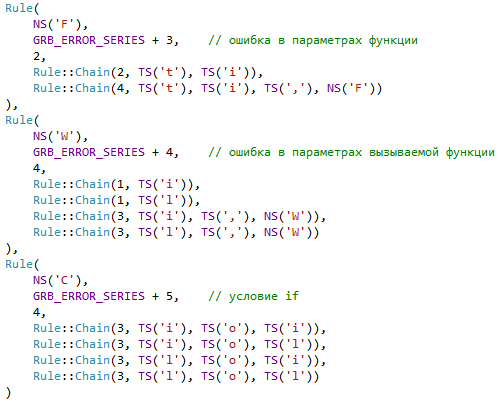




# **Приложение В**

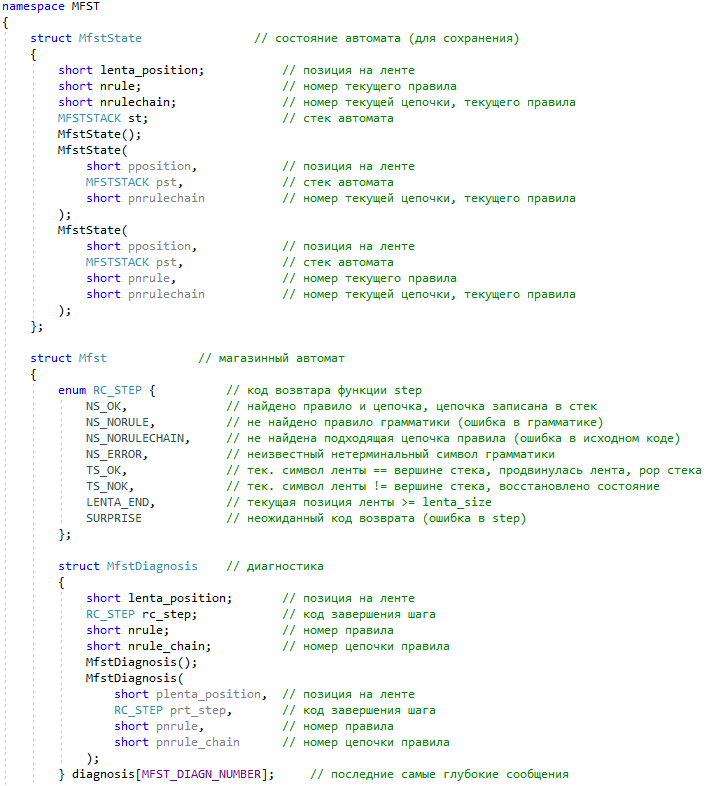
Грамматика языка

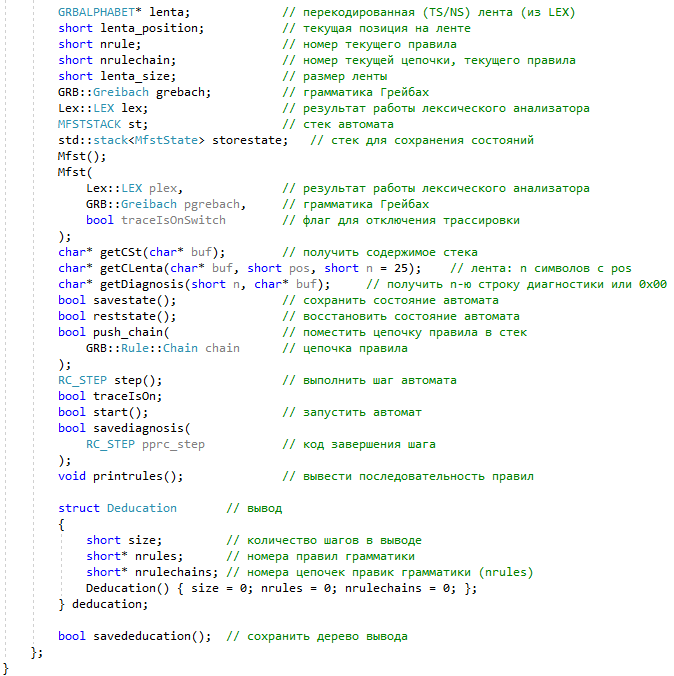




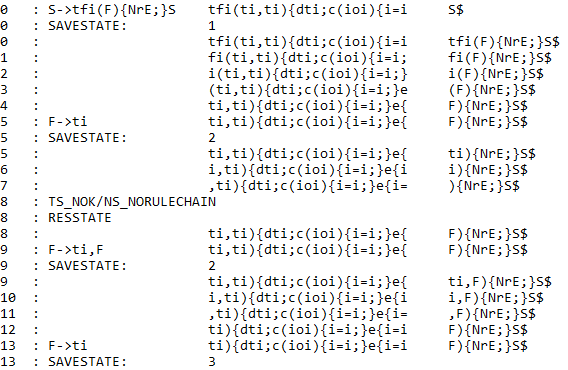
# **Приложение Г**

Структура магазинного автомата и грамматики Грейбах

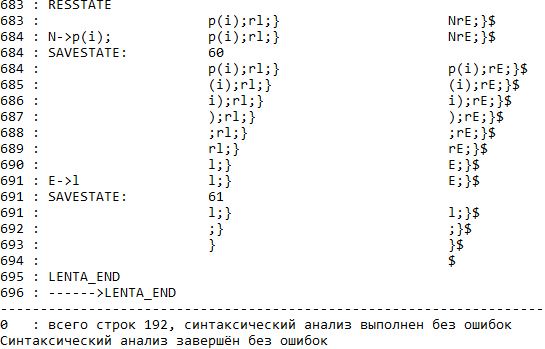




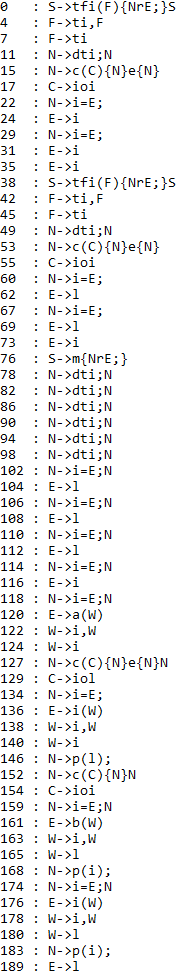
# **Приложение Д**

Начало разбора:

Конец разбора:



Правила, использованные при разборе:



# **Приложение И**

.586

.model flat, stdcall

includelib libucrt.lib

includelib kernel32.lib

includelib ../Debug/StaticLib.lib

EXTERN \_printS :PROC

EXTERN \_printN :PROC

EXTERN \_pow :PROC

EXTERN \_compare :PROC

ExitProcess PROTO :DWORD

.stack 4096

.const

L1 BYTE '../Debug/StaticLib.lib', 0

L2 BYTE 'First argument more than second', 0

L3 BYTE 'Second argument more than first', 0

L4 WORD 17

L5 WORD 18

L6 BYTE 'Mikhail', 0

L7 WORD 1

L8 BYTE 'strings are not equal', 0

L9 WORD 2

L10 WORD 300

L11 WORD 0

.data

minresult WORD 0

ismoreresult DWORD 0

mainstr1 DWORD 0

mainstr2 DWORD 0

maina WORD 0

mainb WORD 0

mainc WORD 0

mainresult WORD 0

.code

min PROC mina : WORD, minb : WORD

mov ax, minb

cmp ax, mina

jl p0

jg p1

je p1

p0:

push minb

pop minresult

jmp ife0

p1:

push mina

pop minresult

ife0:

push minresult

pop eax

ret

min ENDP

ismore PROC ismorea : WORD, ismoreb : WORD

mov ax, ismorea

cmp ax, ismoreb

jg p2

jl p3

je p3

p2:

push offset L2

pop ismoreresult

jmp ife1

p3:

push offset L3

pop ismoreresult

ife1:

push ismoreresult

pop eax

ret

ismore ENDP

main PROC

push L4

pop mainb

push L5

pop mainc

push offset L6

pop mainstr1

push mainstr1

pop mainstr2

push mainstr2

push mainstr1

call \_compare

push eax

pop mainresult

mov ax, mainresult

cmp ax, L7

je p4

jg p5

jl p5

p4:

movzx eax, mainc

push eax

movzx eax, mainb

push eax

call min

push eax

pop maina

jmp ife2

p5:

push offset L8

call \_printS

ife2:

mov ax, maina

cmp ax, mainb

je p6

jg p6

jl p7

p6:

movzx eax, L9

push eax

movzx eax, mainc

push eax

call \_pow

push eax

pop mainb

push mainb

call \_printN

p7:

movzx eax, L10

push eax

movzx eax, mainb

push eax

call ismore

push eax

pop mainstr1

push mainstr1

call \_printS

push L11

call ExitProcess

main ENDP

end main