МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Учреждение образования «БЕЛОРУССКИЙ

ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Факультет Информационных Технологий

Кафедра Программной инженерии

Специальность 1-40 01 01 Программное обеспечение информационных технологий

Специализация Программирование интернет-приложений

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

**К КУРСОВОМУ ПРОЕКТУ НА ТЕМУ:**

«Разработка компилятора SMA-2023»

Выполнил студент Станчик Максим Андреевич

(Ф.И.О.)

Руководитель проекта ст.пр. Наркевич Аделина Сергеевна

(учен. степень, звание, должность, подпись, Ф.И.О.)

Заведующий кафедрой к.т.н., доц. Смелов В.В.

(учен. степень, звание, должность, подпись, Ф.И.О.)

Консультанты ст.пр. Наркевич Аделина Сергеевна

(учен. степень, звание, должность, подпись, Ф.И.О.)

(учен. степень, звание, должность, подпись, Ф.И.О.)

Нормоконтролер ст.пр. Наркевич Аделина Сергеевна

(учен. степень, звание, должность, подпись, Ф.И.О.)

Курсовой проект защищен с оценкой

Минск 2023

**Содержание**

[**Введение** 4](#_Toc58811850)

[**Глава 1. Спецификация языка программирования** 5](#_Toc58811851)

[**1.1.** **Характеристика языка программирования** 5](#_Toc58811852)

[**1.2.** **Алфавит языка** 5](#_Toc58811853)

[**1.3.** **Символы сепараторы** 6](#_Toc58811854)

[**1.4.** **Применяемые кодировки** 6](#_Toc58811855)

[**1.5.** **Типы данных** 6](#_Toc58811856)

[**1.6.** **Преобразование типов данных** 7](#_Toc58811857)

[**1.7.** **Идентификаторы** 7](#_Toc58811858)

[**1.8.** **Литералы** 7](#_Toc58811859)

**1.9. Объявление данных……………………………………………………………...**

[**1.10.Область видимости идентификаторов** 8](#_Toc58811860)

[**1.11.** **Инициализация данных** 8](#_Toc58811861)

[**1.12.** **Инструкции языка** 8](#_Toc58811862)

[**1.13.** **Операции языка** 8](#_Toc58811863)

[**1.14.** **Выражения и их вычисления** 9](#_Toc58811864)

[**1.15.** **Программные конструкции языка** 9](#_Toc58811865)

[**1.16.** **Область видимости** 9](#_Toc58811866)

[**1.17.** **Семантические проверки** 10](#_Toc58811867)

[**1.18.** **Распределение оперативной памяти на этапе выполнения** 10](#_Toc58811868)

[**1.19.** **Стандартная библиотека и её состав** 10](#_Toc58811869)

[**1.20.** **Ввод и вывод данных** 11](#_Toc58811870)

[**1.21.** **Точка входа** 11](#_Toc58811871)

[**1.22.** **Препроцессор** 11](#_Toc58811872)

[**1.23.** **Соглашения о вызовах** 11](#_Toc58811873)

[**1.24.** **Объектный код** 11](#_Toc58811874)

[**1.25.** **Классификация сообщений транслятора** 12](#_Toc58811875)

**[1.26.](#_Toc58811876)****[Контрольный пример](#_Toc58811876)** [12](#_Toc58811876)

[**Глава 2. Структура транслятора** 13](#_Toc58811877)

[**2.1 Компоненты транслятора, их назначение и принципы взаимодействия** 13](#_Toc58811878)

[**2.2 Перечень входных параметров транслятора** 14](#_Toc58811879)

[**2.3 Перечень протоколов, формируемых транслятором и их содержимое** 14](#_Toc58811880)

[**Глава 3. Разработка лексического анализатора** 15](#_Toc58811881)

[**3.1 Структура лексического анализатора** 15](#_Toc58811882)

[**3.2 Контроль входных символов** 15](#_Toc58811883)

[**3.3 Удаление избыточных символов** 16](#_Toc58811884)

[**3.4 Перечень ключевых слов, сепараторов, символов операций и соответствующих им лексем, регулярных выражений и конечных автоматов** 16](#_Toc58811885)

[**3.5 Основные структуры данных** 17](#_Toc58811886)

[**3.6 Принцип обработки ошибок** 18](#_Toc58811887)

[**3.7 Структура и перечень сообщений лексического анализатора** 18](#_Toc58811888)

[**3.8 Параметры лексического анализатора и режимы его работы** 18](#_Toc58811889)

[**3.9 Алгоритм лексического анализа** 18](#_Toc58811890)

[**3.10 Контрольный пример** 19](#_Toc58811891)

[**Глава 4. Разработка синтаксического анализатора** 20](#_Toc58811892)

[**4.1 Структура синтаксического анализатора** 20](#_Toc58811893)

[**4.2 Контекстно свободная грамматика, описывающая синтаксис языка** 20](#_Toc58811894)

[**4.3 Построение конечного магазинного автомата** 23](#_Toc58811895)

[**4.4 Основные структуры данных** 23](#_Toc58811896)

[**4.5 Описание алгоритма синтаксического разбора** 23](#_Toc58811897)

[**4.6 Структура и перечень сообщений синтаксического анализатора** 24](#_Toc58811898)

[**4.7 Параметры синтаксического анализатора и режимы его работы** 24](#_Toc58811899)

[**4.8 Принцип обработки ошибок** 24](#_Toc58811900)

[**4.9 Контрольный пример** 25](#_Toc58811901)

[**Глава 5. Разработка семантического анализатора** 26](#_Toc58811902)

[**5.1 Структура семантического анализатора** 26](#_Toc58811903)

[**5.2 Функции семантического анализатора** 26](#_Toc58811904)

[**5.3 Структура и перечень сообщений семантического анализатора** 26](#_Toc58811905)

[**5.4 Принцип обработки ошибок** 26](#_Toc58811906)

[**5.5 Контрольный пример** 27](#_Toc58811907)

[**Глава 6. Преобразование выражений** 28](#_Toc58811908)

[**6.1 Выражения, допускаемые языком** 28](#_Toc58811909)

[**6.2 Польская запись** 28](#_Toc58811910)

[**6.3 Программная реализация обработки выражений** 29](#_Toc58811911)

[**6.4 Контрольный пример** 29](#_Toc58811912)

[**Глава 7. Генерация кода** 30](#_Toc58811913)

[**7.1 Структура генератора кода** 30](#_Toc58811914)

[**7.2 Представление типов данных в оперативной памяти** 30](#_Toc58811915)

[**7.3 Статическая библиотека** 30](#_Toc58811916)

[**Глава 8. Тестирование транслятора** 32](#_Toc58811917)

[**8.1 Тестирование фазы проверки на допустимость символов** 32](#_Toc58811918)

[**8.2** **Тестирование лексического анализатора** 32](#_Toc58811919)

[**8.3 Тестирование синтаксического анализатора** 32](#_Toc58811920)

[**8.4 Тестирование семантического анализатора** 32](#_Toc58811921)

[**Приложение А** 33](#_Toc58811922)

[**Приложение Б** 35](#_Toc58811923)

[**Приложение В** 38](#_Toc58811924)

[**Приложение Г** 42](#_Toc58811925)

[**Приложение Д** 45](#_Toc58811926)

[**Приложение Е** 47](#_Toc58811927)

[**Приложение Ж** 49](#_Toc58811928)

[**Приложение З** 50](#_Toc58811929)

[**Литература** 53](#_Toc58811930)

# **Введение**

Задачей данного курсового проекта является разработка транслятора для своего языка программирования: SMA-2023. Этот язык программирования предназначен для выполнения простейших операций и арифметических действий над строками и числами.

Главная задача транслятора заключается в том, чтобы сделать программу, написанную языке программирования SMA-2023, понятной компьютеру. В данном курсовом проекте исходный код транслируется на язык ассемблера. Язык ассемблера – тип языка программирования низкого уровня, представляющий собой формат записи машинных команд, удобный для восприятия человеком.

Исходя из цели курсового проекта, были определены следующие задачи:

– разработка спецификации языка программирования;

– разработка структуры транслятора;

– разработка лексического анализатора;

– разработка синтаксического анализатора;

– разработка семантического анализатора;

– обработка выражений;

– генерация кода на язык Assembler;

– тестирование транслятора.

Решения каждой из поставленных задач буду приведены в соответствующих главах курсового проекта:

В первой главе работы определена спецификация языка программирования, т.е. описан синтаксис и семантика языка.

Во второй главе работы представлена структура транслятора, т.е. перечислены компоненты транслятора, их назначение и принципы взаимодействия, перечень входных параметров, перечень протоколов, формируемых транслятором и их содержимое.

В третьей главе работы показана разработка лексического анализатора, порождающего таблицы лексем и идентификаторов.

В четвертой главе работы рассказывается о синтаксическом анализаторе, который выполняет синтаксический разбор текста с распечаткой протокола разбора и дерева разбора на основе таблицы лексем.

В пятой главе описан семантический анализатор, показана его работа (распечатка выданных сообщений в трёх примерах на разных этапах).

В шестой главе решены вопросы преобразования выражений, допускаемых языком и приведена часть протокола для контрольного примера, отображающая результаты преобразования выражений в польский формат.

В седьмой главе представлена генерация кода, где из промежуточного представления порождается код на целевом языке.

В восьмой главе описывается тестирование транслятора.

# **Глава 1. Спецификация языка программирования**

## **Характеристика языка программирования**

Язык программирования SMA-2023 – это процедурный язык высокого уровня, который транслируется в язык ассемблера. это процедурный, статически типизированный язык с ограниченным набором инструкций язык, который транслируется в язык ассемблера.

* 1. **Алфавит языка**

Язык программирования SMA-2023 использует таблицу символов Windows-1251. Исходный код может содержать символы латинского алфавита верхнего и малого регистров, цифры десятичной системы счисления от 0 до 9.

Таблица 1.1 – Алфавит языка

|  |  |
| --- | --- |
| Название подгруппы | Символы подгруппы |
| Символы латинского языка и кириллицы | [a-z|A-Z|А-Я|а-я]\* |
| Знаковые символы и числовые символы | [0-9|?|!|;|:|-|+|\*|/|%]\* |

## **Символы сепараторы**

Сепараторы необходимы для разделения операций языка. Сепараторы, используемые в языке программирования SMA-2023, приведены в таблице 1.2.

Таблица 1.2 – Сепараторы

|  |  |  |
| --- | --- | --- |
| Сепаратор | Название | Область применения |
| ‘ ‘ | Пробел | Допускается везде, кроме идентификаторов и ключевых слов |
| ; | Точка с запятой | Разделение конструкций |
| {…} | Фигурные скобки | Заключение программного блока |
| […] | Квадратные кавычки | Блок кода |
| (…) | Круглые скобки | Приоритет операций, параметры функции |
| ‘…’ | Одинарные кавычки | Символьный литерал |
| = | Знак «равно» | Присваивание значения |
| , | Запятая | Разделение параметров |

## **Применяемые кодировки**

В основе алфавита SMA-2023 лежит таблица символов WINDOWS-1251. Исходный код может содержать символы кириллицы, латинского алфавита верхнего и малого регистров, цифры десятичной системы счисления от 0 до 9, спецсимволы.

## **Типы данных**

В языке используется 4 основных типов данных, которые описываются в таблице 1.3.

Таблица 1.3 – Типы данных языка SMA-2023

|  |  |
| --- | --- |
| Тип данных | Описание типа данных |
| int | Фундаментальный тип данных. Предусмотрен для объявления целочисленных знаковых данных (4 байта).  Автоматически инициализируется нулевым значением.  Возможные операции:  *Арифметические*  + – бинарный, суммирование;  - – бинарный, вычитание;  \* – бинарный, умножение;  / – бинарный, деление;  % – бинарный, нахождение остатка от деления; |
| string | Фундаментальный тип данных. Предусмотрен для объявления строк. (1 символ – 1 байт). Автоматическая инициализация строкой нулевой длины. Максимальное количество символов в строке – 255. |
| symbol | Фундаментальный тип данных. Предусмотрен для объявления символов (2 байтa). Автоматическая инициализация нулевым символом. |
| boolean | Фундаментальный тип данных (2 байта), используемый для объявления логической переменной, которая принимает одно из двух значений: true или false. Без явно указанной инициализации переменной, присваивается нулевое значение (false). |

## **Преобразование типов данных**

В языке программирования SMA-2023 поддерживается преобразование строки в число.

## **Идентификаторы**

Идентификаторы используются для именования переменных, функций, структур и других элементов программы. Вот некоторые основные правила идентификаторов в SMA-2023:

* состоит из символов латинского алфавита [A-Z|a-z]+.
* максимальная длина идентификатора равна 15 и не должна превышать это значение.

## **Литералы**

В языке программирования SMA-2023 существует только 4 типа литералов: целые, символьные, строковые и логические. Их краткое описание представлено в таблице 1.4.

Таблица 1.4 – Литералы

|  |  |
| --- | --- |
| Тип литерала | Описание |
| Литералы целого типа | Целочисленные литералы. Могут быть представлены как в десятичном, двоичном (первый символ: ^), так и в восьмеричном (первый символ: 0)представлении. |
| Строковые литералы | Символы, заключённые в " " (двойные кавычки), инициализируются пустой строкой, строковые переменные. |
| Символьные литералы | Символы, заключённые в ' ' (одинарные кавычки), инициализируются нулевым символом, строковые переменные. |
| Логические литералы | true и false являются логическими литералами. true интерпретируется, как 1, а false как 0. Логический тип относится к целым типам. |

## **Объявление данных**

Объявление данных позволяет программе резервировать необходимую память и создавать переменные или объекты для хранения значений. После объявления данных можно использовать переменную или объект для хранения, изменения и получения значений в программе.

Таблица 1.5 – Способы объявления данных

|  |  |
| --- | --- |
| Способ | Описание |
| Объявление переменных | Чтобы объявить переменную, нужно написать var, потом использовать конкретный тип данных, после которого следует имя переменной, а потом символ “;” (точка с запятой).  Пример: var boolean h; |
| Задание начального значения переменной | Можно назначить начальное значение переменной при её объявлении, используя оператор присваивания =.  Пример: var int k = 9; |
| Объявление комментария | Для того, чтобы добавить комментарий в исходный код программы, написанной на SMA-2023, нужно использовать символ ?, после которого можно использовать символы Windows-1251 для заполнения комментария содержимым. |

## **Область видимости идентификаторов**

Перед использованием переменной необходимо её объявление. Допускается использование переменной только внутри её области видимости. Допускается объявление переменных с одинаковыми именами в разных программных блоках.

## **Инициализация данных**

Способы инициализации переменных языка программирования SMA-2023 представлены в таблице 1.5.

Таблица 1.5 – Способы инициализации переменных

|  |  |
| --- | --- |
| Вид инициализации | Примечание |
| var <тип данных> <идентификатор>; | Объявление(по умолчанию) |
| var <тип данных> <идентификатор> = <значение>; | Инициализация переменной с присваиванием значения. |

## **Инструкции языка**

Инструкции языка программирования SMA-2023 представлены в общем виде в таблице 1.6.

Таблица 1.6 – Инструкции языка программирования SMA-2023

|  |  |
| --- | --- |
| Инструкция | Запись |
| Объявление переменной | var <тип данных> <идентификатор>; |
| Присваивание | <идентификатор> = <значение>|<идентификатор>; |
| Объявление метода | <тип данных> method <идентификатор> (<тип данных> <идентификатор>, …) {…} |
| Объявление процедуры | procedure <идентификатор> (<тип данных> <идентификатор>, …) {…} |
| Блок инструкций | {  …  } |
| Возврат из подпрограммы | recive <выражение> |
| Вывод данных | write <выражение> |
| Вызов функций | <идентификатор функции>(<идентификатор | литерал>,…) |
| Оператор цикла | cycle (<условие>)  {…} |
| Условный оператор | condition(<условие>)  {…}  otherwise  {…} |

## **Операции языка**

## Язык программирования SMA-2023 может выполнять арифметические операции, представленные в таблице 1.7.

Таблица 1.7 – Операции и их приоритеты

|  |  |
| --- | --- |
| Операция | Приоритет операции |
| (  ) | 0 |
| , | 1 |
| +  - | 2 |
| \*  / (деление нацело)  ? (деление по модулю) | 3 |

## **Выражения и их вычисления**

Круглые скобки в выражении используются для изменения приоритета операций. Также не допускается запись двух подряд идущих арифметических операций. Выражение может содержать вызов функции. В математических и программных выражениях круглые скобки используются для изменения приоритета операций. Они позволяют установить порядок выполнения операций, определяя, какие операции должны быть выполнены первыми. Выражения, заключенные в круглые скобки, вычисляются первыми, а их результат может быть использован в других частях выражения.

## **Программные конструкции языка**

Ключевые программные конструкции языка программирования SMA-2023 представлены в таблице 1.8

Таблица 1.8 – программные конструкции

|  |  |
| --- | --- |
| Главная функция (точка входа в приложение) | primary  {…} |
| Метод | method <тип данных> <идентификатор> (<тип данных> <идентификатор>, …)  {  …  recive <выражение>;  } |
| Процедура | procedure <идентификатор> (<тип данных> <идентификатор>, …)  { … } |
| Цикл | cycle (<условие>)  {…} |
| Условная конструкция condition | condition(<условие>)  {…}  otherwise{…} |

## **Область видимости**

В языке программирования SMA-2023 переменные обязаны находиться внутри программного блока функций. Внутри разных областей видимости разрешено объявление переменных с одинаковыми именами. Все переменные, параметры или функции внутри области видимости получают префикс, который отображается в таблице идентификаторов.

## **Семантические проверки**

Таблица с перечнем семантических проверок, предусмотренных языком, приведена в таблице 1.8.

Таблица 1.8 – Семантические проверки

|  |  |
| --- | --- |
| Номер | Правило |
| 1 | Идентификаторы функций не должны повторяться |
| 2 | Операнды в операторах ветвления и выхода из функции должны быть целочисленного или логического типа |
| 3 | Тип данных передаваемых значений в функцию должен совпадать с типом параметров при её объявлении |
| 4 | Тип данных передаваемых значений в функцию стандартной библиотеки должен соответствовать заявленному. |
| 5 | Идентификатор должен быть объявлен до его использования. |
| 6 | Операнды в арифметическом выражении не могут быть разных типов |

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

Все переменные размещаются в стеке. Таблица лексем и таблица идентификаторов сохраняются в структуры с выделенной под них динамической памятью, которая очищается по окончанию работы транслятора.

## **Стандартная библиотека и её состав**

Функции стандартной библиотеки с описанием представлены в таблице 1.9. Стандартная библиотека написана на языке программирования C++.

Таблица 1.9 – Состав стандартной библиотеки

|  |  |  |  |
| --- | --- | --- | --- |
| Имя функции | Возвращаемое значение | Принимаемые параметры | Описание |
| strlength | int | string a | Функция возвращает длину строки а. |
| stoi | int | string a | Функция преобразует строку в число до первого символа не цифры, если преобразование не удается возвращает 0 |
| strcomp | int | string a, string b | Функция сравнивает две строки. Возвращает 0 (если равны), 1(первая больше второй), -1(вторая больше первой) |
| OutputStr | - | string a | Функция выводит в консоль строку a |
| BREAKL | - | - | Переход на следующую строку |
| OutputInt | - | int a | Функция выводит в консоль число a |
| OutputChar | - | int a | Функция выводит в консоль символ a |
| OutputBoolean | - | int a | Функция выводит в консоль true(false)  в зависимости от параметра а - 1(0) |

## **Ввод и вывод данных**

В языке программирования SMA-2023 ввод данных не поддерживается. Вывод данных происходит с помощью функции write(<идентификатор>|<литерал>);

## **Точка входа**

Точкой входа является функция primary.

## **Препроцессор**

Препроцессор в языке программирования SMA-2023 не предусмотрен.

## **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**

SMA-2023 транслируется в язык ассемблера.

## **Классификация сообщений транслятора**

В случае возникновения ошибки в коде программы на языке SMA-2023 и выявления её транслятором в текущий файл протокола выводится сообщение. Их классификация сообщений приведена в таблице 1.10.

Таблица 1.10. Классификация сообщений транслятора

|  |  |
| --- | --- |
| Интервал | Описание ошибок |
| 0-110 | Системные ошибки, ошибки параметров |
| 200-220 | Ошибки лексического анализа |
| 300-400 | Ошибки семантического анализа |
| 600-620 | Ошибки синтаксического анализа |

## **Контрольный пример**

Контрольный пример представлен в приложении А.

# **Глава 2. Структура транслятора**

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

Транслятор преобразует программу, написанную на языке SMA-2023 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 SMA-2023 и формирование таблицы лексем и таблицы идентификаторов.

Синтаксический анализ – это основная часть транслятора, предназначенная для распознавания синтаксических конструкций и формирования промежуточного кода SMA-2023. Для этого используются таблица лексем и идентификаторов. Синтаксический анализатор распознаёт синтаксические конструкции, выявляет синтаксические ошибки при их наличии и формирует дерево разбора

Семантический анализ в свою очередь является проверкой исходной программы SMA-2023 на семантическую согласованность с определением языка, т.е. проверяет правильность текста исходной программы с точки зрения семантики.

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 SMA-2023, прошедший успешно все предыдущие этапы, в код на языке Ассемблера.

## **2.2 Перечень входных параметров транслятора**

Входные параметры представлены в таблице 2.1.

Таблица 2.1 – Входные параметры транслятора языка SMA-2023

|  |  |  |
| --- | --- | --- |
| Входной параметр | Описание | Значение по умолчанию |
| -in:<имя\_файла> | Входной файл с любым расширением, в котором содержится исходный код на языке SMA-2023. Данный параметр должен быть указан обязательно. В случае если он не будет задан, то выполнение этапа трансляции не начнётся. | Не предусмотрено |
| -log:<имя\_файла> | Файл содержит в себе краткую информацию об исходном коде на языке SMA-2023. В этот файл могут быть выведены таблицы идентификаторов, лексем, а также дерево разбора. | <имя\_файла>.log |
| -out:<имя\_файла> | В этот файл будет записан результат трансляции кода на язык assembler | <имя\_файла>.asm |

Таблицы лексем и дерево разбора синтаксического анализатора выводятся в logOut журнал.

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

Таблица с перечнем протоколов, формируемых транслятором языка SMA-2023 и их назначением представлена в таблице 2.2.

Таблица 2.2 – Протоколы, формируемые транслятором языка SMA-2023

|  |  |
| --- | --- |
| Формируемый протокол | Описание протокола |
| Файл журнала, log | Файл содержит в себе краткую информацию об исходном коде на языке SMA-2023. В этот файл выводится различные ошибки, а также результат работы анализаторов |
| Файл журнала, logOut | Файл содержит в себе краткую информацию об исходном коде на языке SMA-2023. В этот файл выводится протокол работы анализаторов. |
| asm | Содержит сгенерированный код на языке Ассемблера. |

В log файл выводятся все ошибки, за исключением тех, что связаны с открытием файла log или считывания параметров.

# **Глава 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 SMA-2023. На выходе формируется таблица лексем и таблица идентификаторов. Структура лексического анализатора представлена на рисунке 3.1.



Рисунок 3.1 Структура лексического анализатора SMA-2023

## **3.2 Контроль входных символов**

Таблица для контроля входных символов представлена на рисунке 3.2

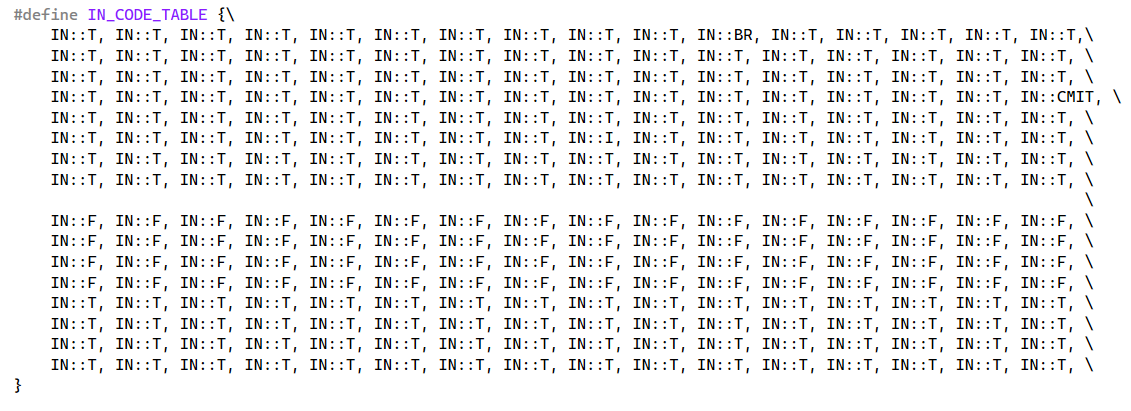


Рисунок 3.2. Таблица контроля входных символов

Принцип работы таблицы заключается в соответствии значения каждому элементу в шестнадцатеричной системе счисления значению в таблице WINDOWS-1251.

Описание значения символов: T – разрешённый символ, F – запрещённый символ, I – игнорируемый символ, BR – символ перехода на новую строку (\n), CMIT – символ комментария(?).

## **3.3 Удаление избыточных символов**

Избыточными символами являются символы табуляции и пробелы.

Избыточные символы удаляются на этапе разбиения исходного кода на лексемы.

Описание алгоритма удаления избыточных символов:

1. Посимвольно считываем исходный код, занесенный в структуру In.

2. Встреча пробела или знака табуляции вне пределов строкового литерала является своего рода встречей символа-сепаратора.

3. В отличие от других символов-сепараторов не записываем в таблицу лексем эти символы, т.е. игнорируем.

## **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 цепочки | Цепочка | Лексема |
| Ключевые слова | var | d |
| int, string, boolean, symbol | t |
| primary | m |
| method | f |
| procedure | p |
| recive | r |
| write | o |
| cycle | u |
| condition | w |
| otherwise | ! |
| Иное | Идентификатор | i |
| Литерал | l |

Продолжение таблицы 3.1

|  |  |  |
| --- | --- | --- |
| Сепараторы | ; | ; |
| , | , |
| { | { |
| } | } |
| ( | ( |
| ) | ) |
| Операторы | Арифметические (+, -, \*, /, %), Cравнения (!,~,<,>) | v |
| Присваивание (=) | = |

Пример реализации таблицы лексем представлен в приложении Б.

Также в приложении В находятся конечные автоматы, соответствующие лексемам языка SMA-2023.

## **3.5 Основные структуры данных**

Описание основных структур данных, используемых для хранения таблиц идентификаторов, представлено на рис. 3.3.

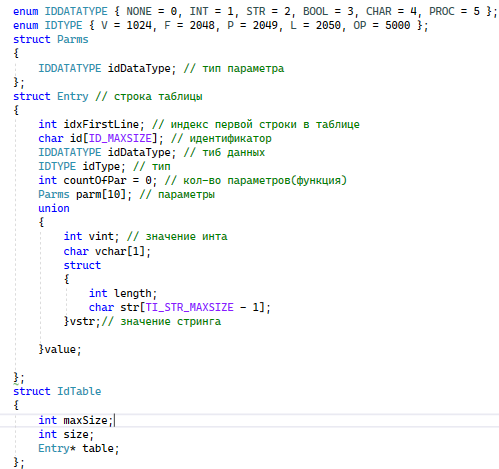


Рисунок 3.3. — Структуры таблиц идентификаторов SMA-2023

Описание основных структур данных, используемых для хранения таблиц лексем, представлено на рис. 3.4.

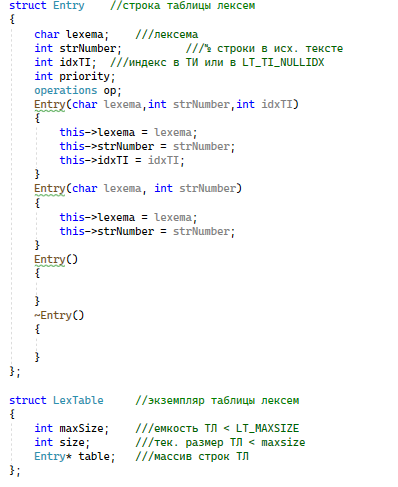


Рисунок 3.4 — Структуры таблиц лексем SMA-2023

## **3.6 Принцип обработки ошибок**

Когда возникает ошибка – работа транслятора прекращается, а ошибка записывается в log журнал.

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

Перечень сообщений лексического анализатора представлен на рисунке 3.5.

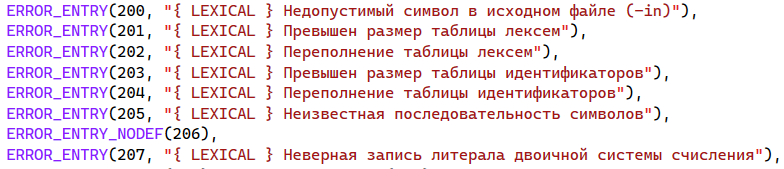


Рисунок 3.5 – Перечень ошибок лексического анализатора

## **3.8 Параметры лексического анализатора и режимы его работы**

Входным параметром лексического анализатора является структура IN, которая содержит исходный текст программы, написанный на языке SMA-2023, а также структура LOG, которая содержит файл протокола.

## **3.9 Алгоритм лексического анализа**

Лексический анализ выполняется программой (входящей в состав транслятора), называемой лексическим анализатором. Цель лексического анализа — выделение и классификация лексем в тексте исходной программы. Лексический анализатор производит распознаёт и разбирает цепочки исходного текста программы. Это основывается на работе конечных автоматов, которую можно представить в виде графов. Регулярные выражения — аналитический или формульный способ задания регулярных языков. Они состоят из констант и операторов, которые определяют множества строк и множество операций над ними. Любое регулярное выражение можно представить в виде графа. Для ускорения работы анализатора я добавил просмотр первого символа слова, за счет этого отсеиваются неподходящие графы. Результат работы лексического анализатора – сформированные таблицы лексем и идентификаторов.

Пример. Регулярное выражение для ключевого слова primary.

Граф конечного автомата для этой лексемы представлен на рисунке 3.6. S0 – начальное состояние, S4 – конечное состояние автомата. В виде кода представлен на рисунке 3.7.

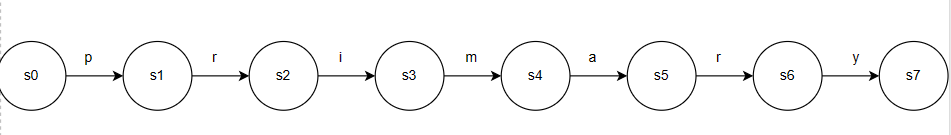


Рисунок 3.6 — Граф переходов для цепочки “primary”

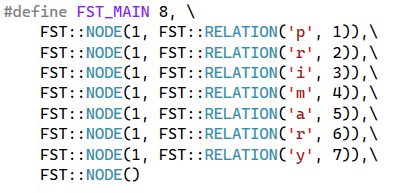


Рисунок 3.7 — Граф переходов для цепочки “ primary ”

## **3.10 Контрольный пример**

Результат работы лексического анализатора –сформированные таблицы лексем и идентификаторов – представлены в приложении Б.

# **Глава 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 SMA-2023 используется контекстно-свободная грамматика , где

T – множество терминальных символов (представлены в приложении З)

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.

Правила языка SMA-2023 представлена в приложении Г.

TS – терминальные символы, которыми являются сепараторы, знаки арифметических операций и некоторые строчные буквы.

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

Таблица 4.1 – Перечень правил, составляющих грамматику языка и описание нетерминальных символов SMA-2023

|  |  |  |
| --- | --- | --- |
| Нетерминал | Цепочки правил | Описание |
| S | ftiFBS  piFUS  m{N}  ftiFB  piFU | Проверка правильности структуры программы |
| F | (P)  () | Проверка наличия списка параметров функции |
| P | ti  ti,P | Проверка на ошибку в параметрах функции при ее объявлении |
| B | {NrI;}  {rI;} | Проверка наличия тела функции |
| I | l  i | Проверка на недопустимое выражение (ожидается только литерал или идентификатор) |
| N | dti;N  dti;  dti=E;N  dti=E;  i=E;N  i=E;  u(R){L}N  u(R){ L }  w(R){ L }N  w(R){ L }  w(R){ L }!{ L }N  w(R){ L }!{ L }  o[I];N  o[I];  b;N  b;  iK;N  iK; | Проверка на неверную конструкцию в теле функции |

Продолжение таблицы 4.1

|  |  |  |
| --- | --- | --- |
| K | (W)  () | Проверка на ошибку в вызове функции |
| E | i  iM  l  lM  (E)  (E)M  iK  iKM | Проверка на ошибку в арифметическом выражении |
| W | i  l  i,W  l,W | Проверка на ошибку в параметрах вызываемой функции |
| M | vE  vEM | Проверка арифметических действий |
| L | i=E;N  i=E;  oi;N  oK;  b;N  b;  iK;N  iK; | Проверка на неверную конструкцию в теле цикла или условного выражения |

## **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 Д.

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 А) |
|  | 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 SMA-2023. Данные структуры представлены в приложении Д.

## **4.5 Описание алгоритма синтаксического разбора**

Принцип работы автомата следующий:

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.

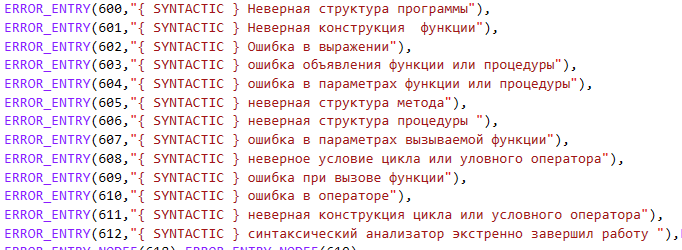


Рисунок 4.1 – Перечень сообщений синтаксического анализатора

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

Входным параметром синтаксического анализатора является структура LEX, которая содержит сформированную таблицу лексем, полученную на этапе лексического анализа, потоки вывода протокола, а также правила контекстно-свободной грамматики в форме Грейбах.

Выходными параметрами являются трассировка прохода таблицы лексем и правила разбора, которые записываются в файл протокола.

## **4.8 Принцип обработки ошибок**

Обработка ошибок происходит следующим образом:

1. Синтаксический анализатор перебирает все правила и цепочки правила грамматики для нахождения подходящего соответствия с конструкцией, представленной в таблице лексем.
2. Если невозможно подобрать подходящую цепочку, то генерируется соответствующая ошибка.
3. Синтаксический анализатор работает до 3 ошибок.
4. В случае нахождения ошибки, после всей процедуры трассировки в протокол будет выведено диагностическое сообщение.

## **4.9 Контрольный пример**

Пример разбора синтаксическим анализатором исходного кода предоставлен в приложении Е в виде фрагмента трассировки и дерева разбора исходного кода.

# **Глава 5. Разработка семантического анализатора**

## **5.1 Структура семантического анализатора**

Семантический анализ в трансляторе языка SMA-2023 выделен в отдельную фазу, 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 и т.д.).

## **5.2 Функции семантического анализатора**

За семантический анализ отвечает функция SemAnalyze. Ее входными параметрами является структура LEX, которая содержит таблицу лексем, идентификаторов и поток вывода в протокол.

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.

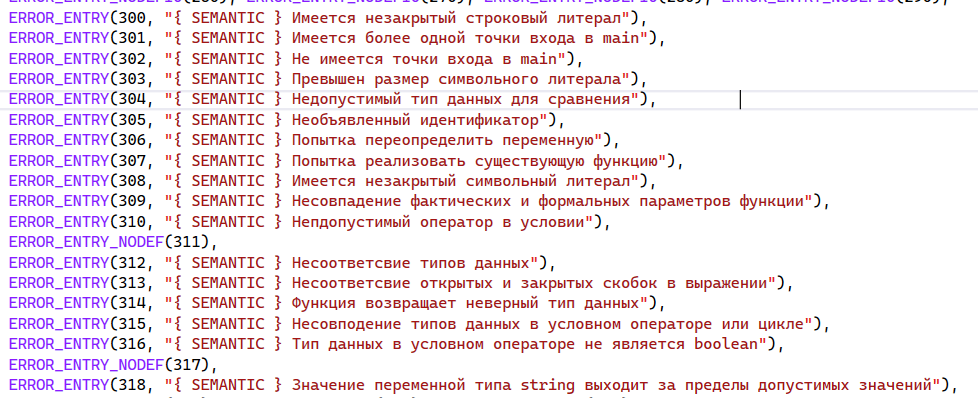


Рисунок 5.1 – Перечень сообщений семантического анализатора

## **5.4 Принцип обработки ошибок**

Семантический анализатор - проверяет, что объявления и утверждения программы семантически верны. Например: соответствие типов данных в выражении, совпадение фактических и формальных параметров функции.

Обработка ошибок происходит следующим образом:

1. Анализатор перебирает таблицу лексем.
2. Проверяет исключительные ситуации, которые могут быть связаны с данной лексемой.
3. В случае нахождения ошибки, она записывается в log журнал и дублируется в консоль
4. По завершению работы в log журнал печатается сообщение об успешной или ошибочной работе анализатора.

## **5.5 Контрольный пример**

Результат работы контрольного примера расположен в приложении Б, где показан результат лексического анализатора, т.к. представленные таблицы лексем и идентификаторов проходят лексическую и семантическую проверки одновременно.

Таблица 5.3 – Тестирование функций

|  |  |
| --- | --- |
| Исходный код с ошибкой | Генерируемое сообщение об ошибке |
| {  var int check = 5;  write check;  } | ERROR CODE 302: { SEMANTIC } Не имеется точки входа в primary  Строка -1 позиция -1 |
| main  {  var int check = 5;  write check;  write "hello;  } | ERROR CODE 300: { SEMANTIC } Имеется незакрытый строковый литерал  Строка -1 позиция -1 |
| main  {  var int check = 5;  var int check = 2;  } | ERROR CODE 306, { SEMANTIC } Попытка переопределить переменную  Строка n позиция -1 |

# **Глава 6. Преобразование выражений**

## **6.1 Выражения, допускаемые языком**

В языке SMA-2023 допускаются выражения, применимые к целочисленным типам данных. В выражениях поддерживаются арифметические операции, такие как +, -, \*, /, %(остаток от деления) и (), и вызовы функций как операнды арифметических выражений.

Приоритет операций представлен в таблице 6.1.

Таблица 6.1 – Приоритет операций в языке SMA-2023

|  |  |
| --- | --- |
| Приоритет | Операция |
| 0 | ( ) |
| 1 | , |
| 2 | + - |
| 3 | \* / % |

## **6.2 Польская запись**

Выражения в языке SMA-2023 преобразовываются к обратной польской записи.

Польская запись – это альтернативный способ записи арифметических выражений, преимущество которого состоит в отсутствии скобок.

Обратная польская запись — это форма записи математических выражений, в которой операторы расположены после своих операндов. Выражение в обратной польской нотации читается слева направо: операция выполняется над двумя операндами, непосредственно стоящими перед знаком этой операции.

Алгоритм построения:

– исходная строка: выражение;

– результирующая строка: польская запись;

– стек: пустой;

– результирующая строка: польская запись;

– исходная строка просматривается слева направо;

– операнды переносятся в результирующую строку в порядке их следования;

– операция записывается в стек, если стек пуст или в вершине стека лежит отрывающая скобка;

– операция выталкивает все операции с большим или равным приоритетом в результирующую строку;

– запятая не помещается в стек, если в стеке операции, то все выбираются в строку;

– отрывающая скобка помещается в стек;

– закрывающая скобка выталкивает все операции до открывающей скобки, после чего обе скобки уничтожаются;

– закрывающая скобка с приоритетом, равным 4, выталкивает все до открывающей с таким же приоритетом и генерирует @ – специальный символ, в которого записывается информация о вызываемой функции, а в поле приоритета для данной лексемы записывается число параметров вызываемой функции;

– по концу разбора исходной строки все операции, оставшиеся в стеке, выталкиваются в результирующую строку.

## **6.3 Программная реализация обработки выражений**

Программная реализация алгоритма преобразования выражений в обратный польский формат основана функциях PolishNotation и StartPolish. Функция StartPolish принимает как параметр адрес таблицы лексем и содержит цикл, в ходе которого перебираются все лексемы исходного кода. Если последовательность лексем соответствует началу выражения, вызывается функция PolishNotation, где и проводится точечное преобразование выражений к польской нотации.

## **6.4 Контрольный пример**

Пример преобразования выражения к польской записи представлен в таблице 6.2. Преобразование выражений в формат польской записи необходимо для построения более простых алгоритмов их вычисления.

В приложении Ж приведен измененное представление промежуточного кода, отображающее результаты преобразования выражений в польский формат.

Таблица 6.2 – Пример преобразования выражения в обратную польскую запись

|  |  |  |
| --- | --- | --- |
| Исходная строка | Результирующая строка | Стек |
| q\*2 - s(i) |  |  |
| \*2 - s(i) | q |  |
| 2 - s(i) | q | \* |
| - s(i) | q2 | \* |
| s(i) | q2\* | - |
| (i) | q2\* | - |
| i) | q2\* | - |
| ) | q2\*i | - |
|  | q2\*i@1- |  |

# **Глава 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.

Структура генератора кода представлена на рисунке 7.1.



Рисунок 7.1 Структура генератора кода

Функция генератора кода представлена на рисунке 7.2.

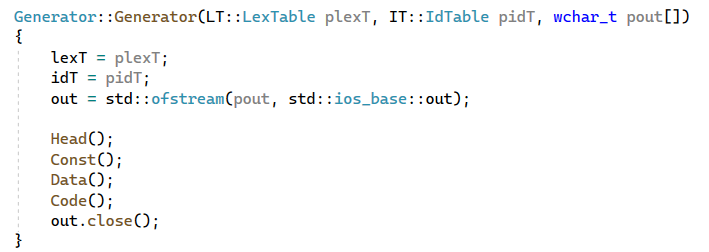


Рисунок 7.2 – Функция генератора кода

Функция Head отвечает за установку необходимого соглашения о вызовах, модели памяти, Подключение библиотек и объявление прототипов статической библиотеки.

Функция Const отвечает за создание сегмента констант, куда заносятся лексемы имеющихся типов данных. Циклом проходим по таблицы идентификаторов, если находим лексему, то смотрим ее тип данных и даем ей эквивалентный тип данных в ассемблере.

Функция Data отвечает за создание сегмента данных, куда заносятся идентификаторы имеющихся типов данных.

Функция Code отвечает за создание сегмента кода, инструкции, написанные на языке SMA-2023 генерируются в Assembler.

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

Элементы таблицы идентификаторов расположены в разных сегментах языка ассемблера – .data и .const. Идентификаторы языка SMA-2023 размещены в сегменте данных(.data). Литералы – в сегменте констант (.const). Соответствия между типами данных идентификаторов на языке SMA-2023 и на языке ассемблера приведены в таблице 7.1. Сгенерированный код предоставлен в приложении И.

Таблица 7.1 – Соответствия типов идентификаторов языка SMA-2023 и языка Ассемблера

|  |  |  |
| --- | --- | --- |
| Тип идентификатора на языке SMA-2023 | Тип идентификатора на языке ассемблера | Пояснение |
| int | SDWORD | Хранит целочисленный тип данных со знаком. 4 байта. |
| string | BYTE | Каждый символ строки хранится размером в 1 байт. |
| symbol | word | Символ хранится размером в 2 байт. |
| boolean | word | Переменная хранится размером в 2 байт. |

## **7.3 Статическая библиотека**

Функции из стандартной библиотеки содержатся в проекте StaticLibrary, в свойствах которого указан тип конфигурации «статическая библиотека». Подключение библиотеки происходит с помощью includelib на этапе генерации кода путем вывода в поток out. Таким же образом c помощью оператора EXTRN объявляются названия функций из библиотеки. Оператор EXTRN выполняет две функции. Во-первых, он сообщает ассемблеру, что указанное символическое имя является внешним для текущего ассемблирования. Вторая функция оператора EXTRN состоит в том, что он указывает ассемблеру тип соответствующего символического имени. Так как ассемблирование является очень формальной процедурой, то ассемблер должен знать, что представляет из себя каждый символ. Это позволяет ему генерировать правильные команды. Вышеописанное проиллюстрировано на рисунке 7.3.

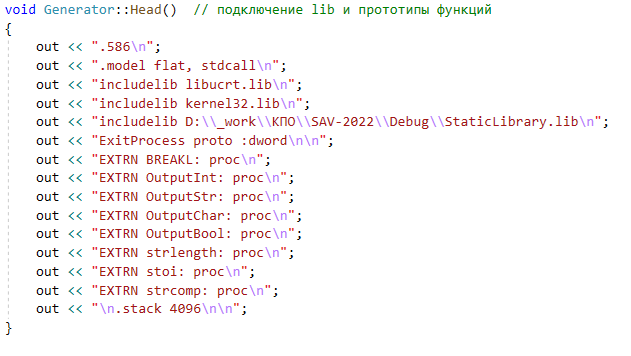


Рисунок 7.3 Фрагмент функции генерации кода

# **Глава 8. Тестирование транслятора**

## **8.1 Тестирование фазы проверки на допустимость символов**

В языке SMA-2023 не разрешается использовать запрещённые входным алфавитом символы. Результат использования запрещённого символа показан в таблице 8.1.

Таблица 8.1 – Тестирование фазы проверки на допустимость символов

|  |  |
| --- | --- |
| Исходный код | Диагностическое сообщение |
| met№hod int counts(int k)  {  …  recive k;  } | ERROR CODE 200: { LEXICAL } Недопустимый символ в исходном файле (-in)  Строка 1 позиция 4 |

## **Тестирование лексического анализатора**

На этапе лексического анализа могут возникнуть ошибки, описанные в пункте 3.7. Результаты тестирования лексического анализатора показаны в таблице 8.2.

Таблица 8.2 – Тестирование лексического анализатора

|  |  |
| --- | --- |
| Исходный код | Диагностическое сообщение |
| var symbol g = 'f; | ERROR CODE 308: { SEMANTIC } Имеется незакрытый символьный литерал  Строка 1 позиция -1 |

## **8.3 Тестирование синтаксического анализатора**

На этапе синтаксического анализа могут возникнуть ошибки, описанные в пункте 4.6. Результаты тестирования синтаксического анализатора показаны в таблице 8.3.

Таблица 8.3 – Тестирование синтаксического анализатора

|  |  |
| --- | --- |
| Исходный код | Диагностическое сообщение |
| int method kek( int z)  {  var int n = z+5;  } | ERROR CODE 605: { SYNTACTIC } неверная структура метода  Строка 4 позиция -1 |

## **8.4 Тестирование семантического анализатора**

Итоги тестирования семантического анализатора приведены в пункте 5.5.**Приложения**

## **Приложение А**

boolean method EvenCheck(int k)

{

var boolean h;

k= k % 2;

condition (k ~ 0)

{

h = true;

}

otherwise

{

h = false;

}

recive h;

}

int method GetFact(int n)

{

var int k;

var int j;

condition (n < 0)

{

k = 0;

}

condition (n ~ 0)

{

k = 1;

}

otherwise

{

j=n-1;

k=n\*GetFact(j);

}

recive k;

}

procedure Count(string h)

{

write "счет от нуля до длины строки: ";

write h;

hortab;

var int length = strlength(h);

var int i=0;

cycle (i < length)

{

write i;

write ',';

i=i+1;

}

}

primary

{

var int number = GetFact(7); ? вычисляем факториал

write "факториал числа равен: ";

write number;

hortab ;

var boolean iseven = EvenCheck(number); ? Проверка на четность числа

write iseven;

hortab;

Count("hello");

hortab;

var int binary = ^10001010;

write binary;

hortab;

var int octal = 015;

var int result = strlength("hello")+((^1010+i012+45)/(2+(4\*5+1)/3))+ stoi("12")+(octal+binary)%2;

write result;

hortab;

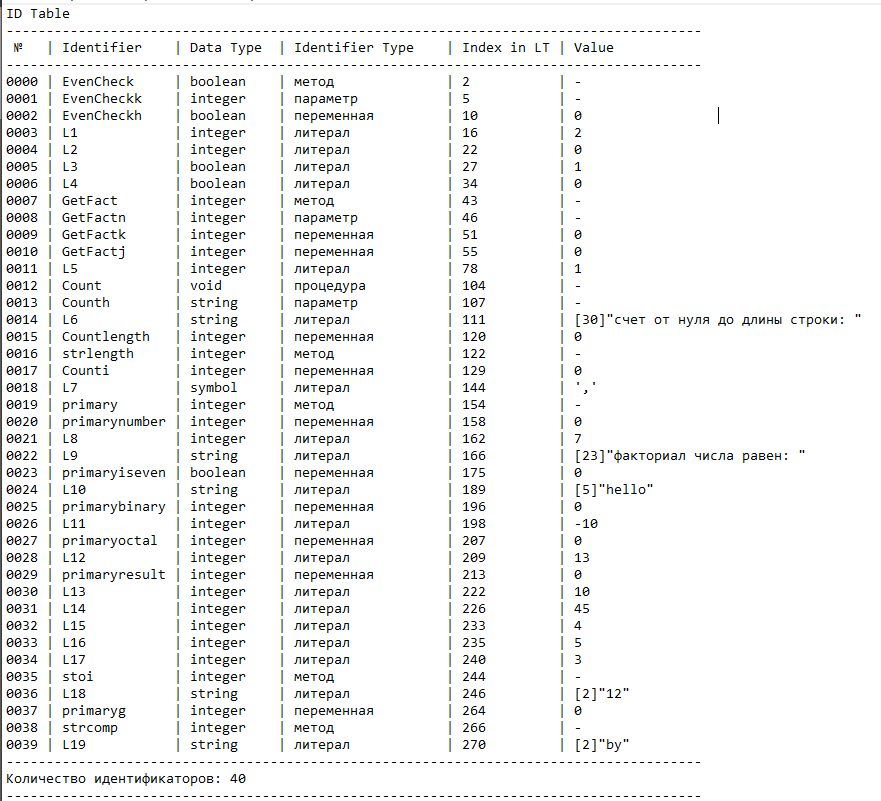
var int g = strcomp("hello","by");

write g;

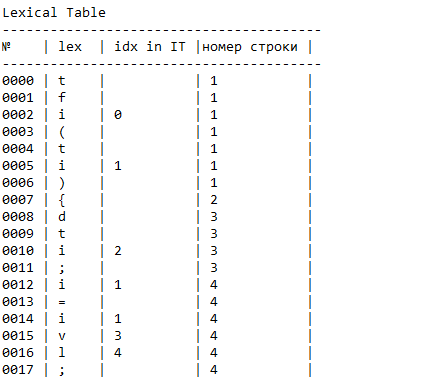
hortab;

}

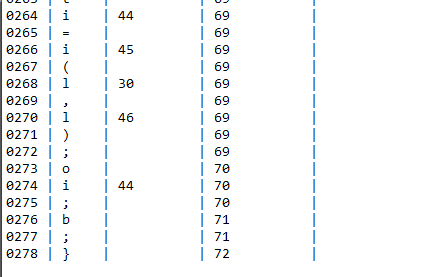
## **Приложение Б**

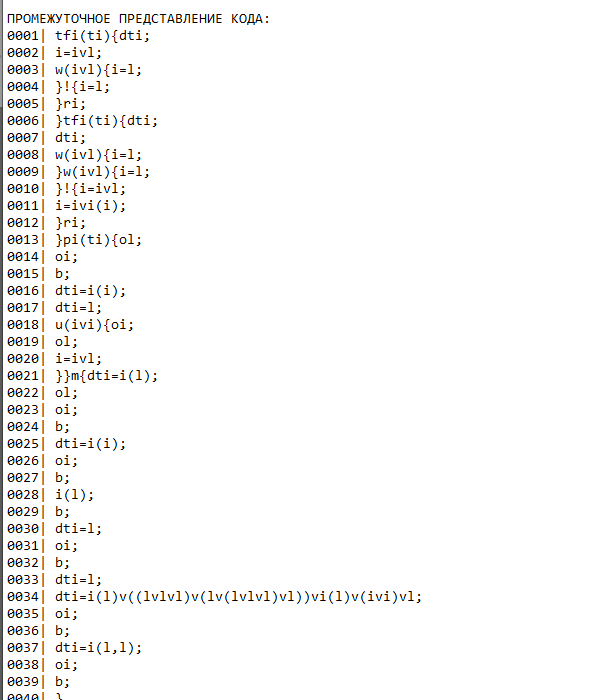


Начало таблицы лексем:

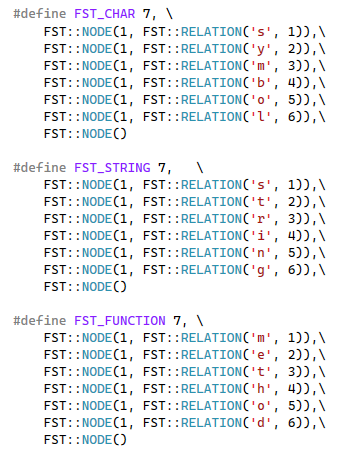
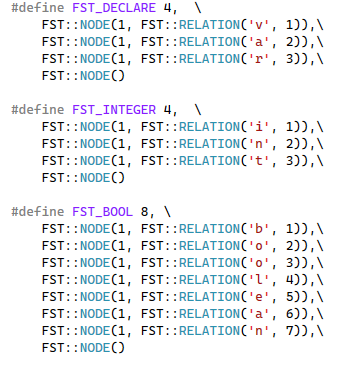


Окончание таблицы лексем:

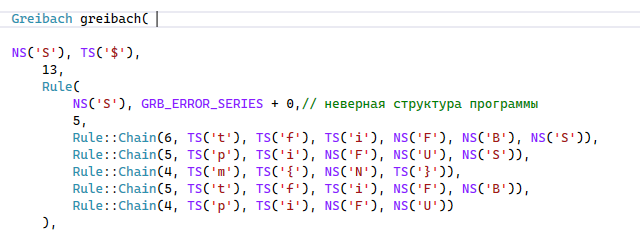




## **Приложение В**

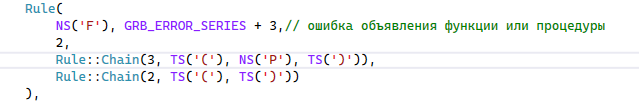


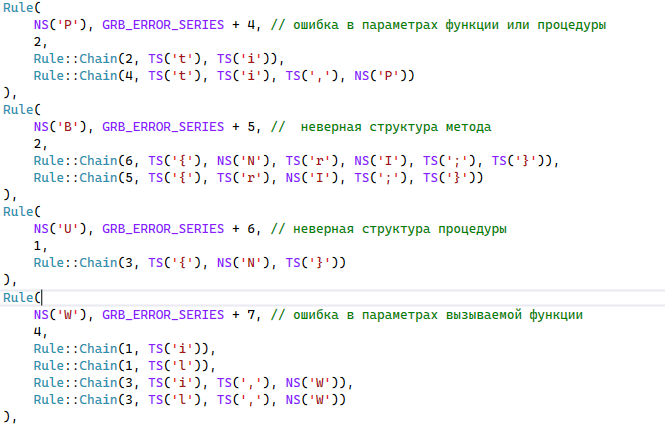
## **Приложение Г**



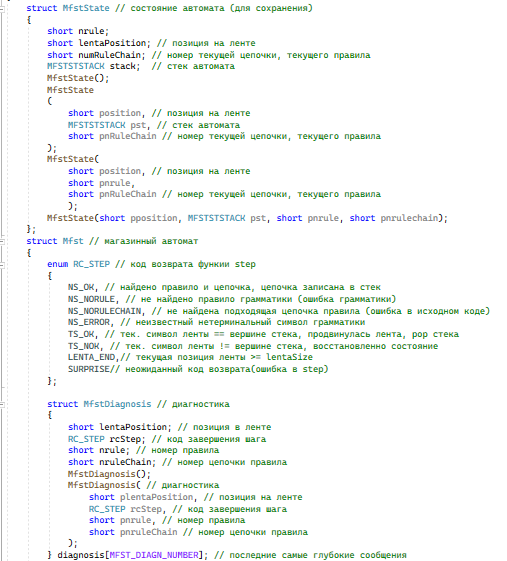
## 

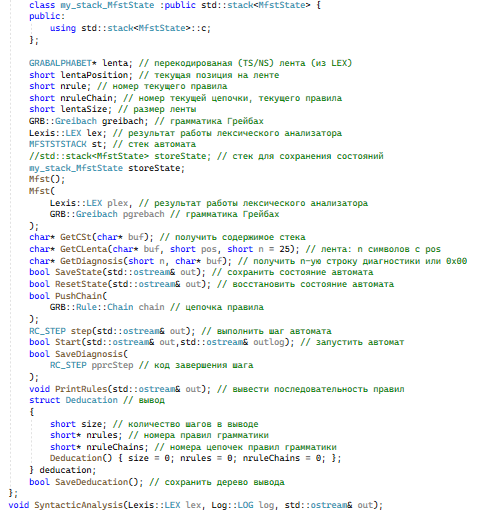
## 





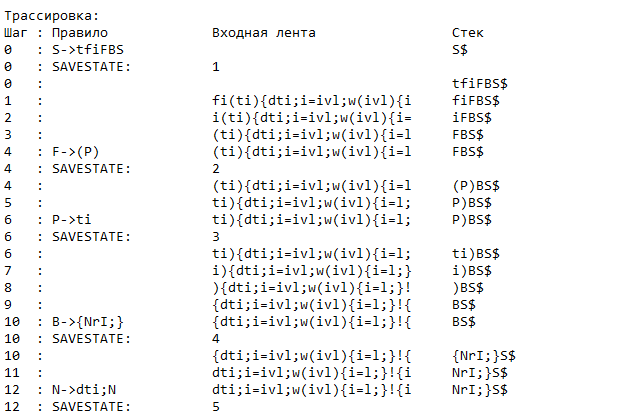
## **Приложение Д**



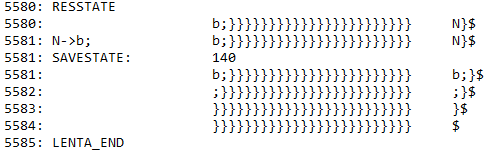


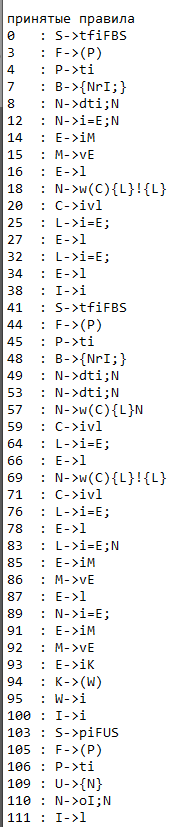
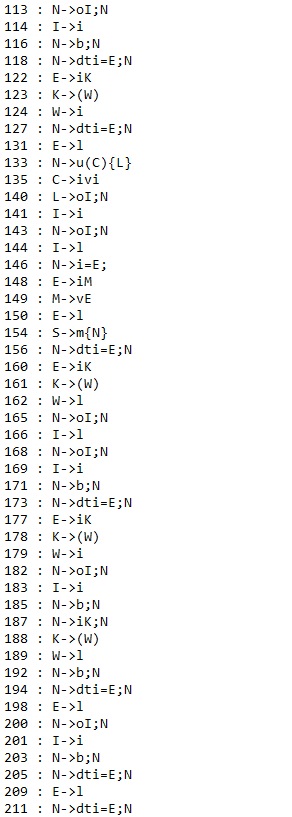
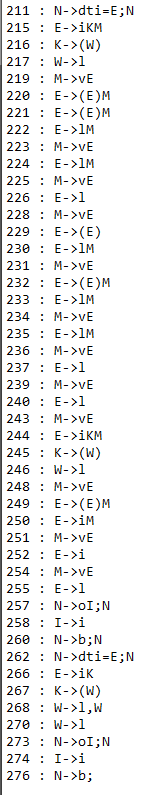
## **Приложение Е**

Начало разбора

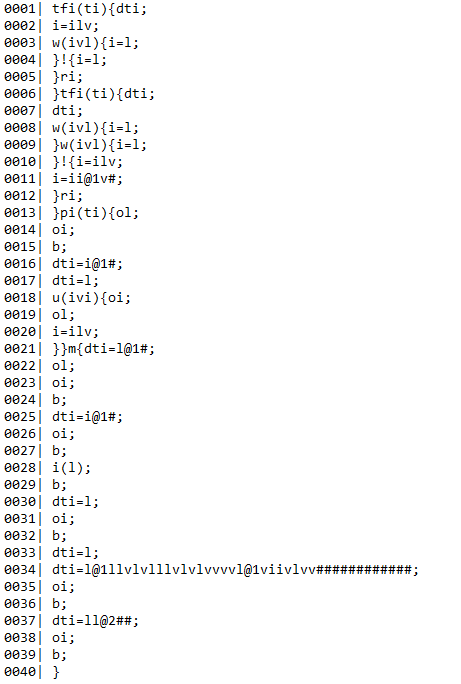


Конец разбора



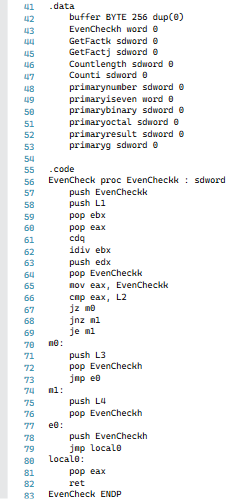
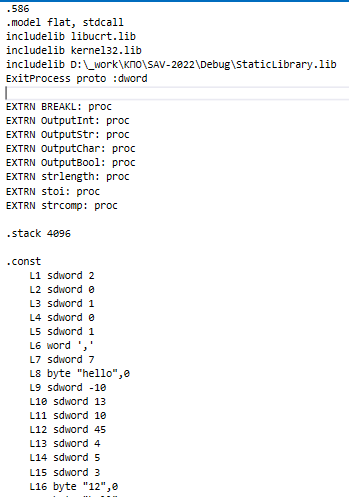
  

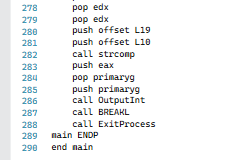
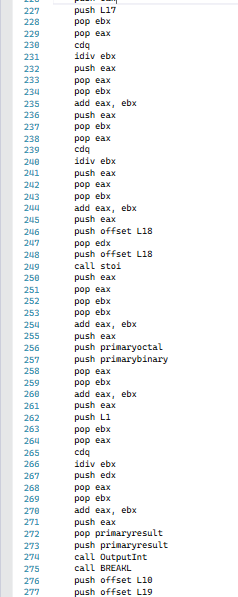
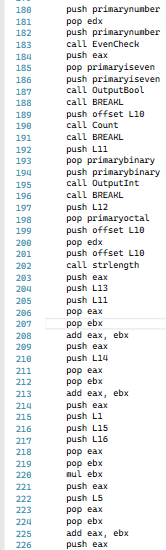
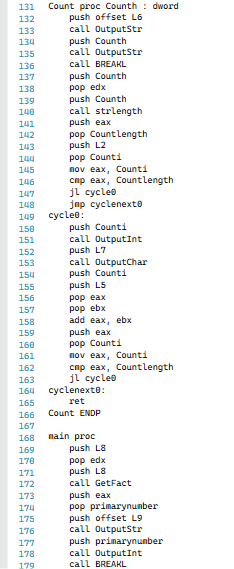
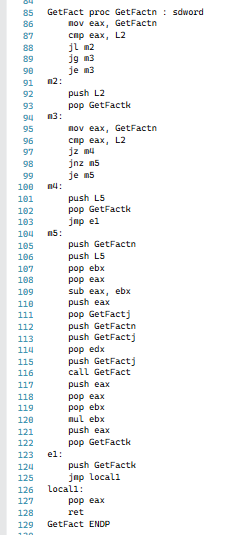
## **Приложение Ж**



**Приложение З**

## **Приложение И**





# **Литература**

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. Ирвин К. Р. Язык ассемблера для процессоров Intel / К. Р. Ирвин. – M.: Вильямс, 2005. – 912с.
4. Герберт, Ш. Справочник программиста по C/C++ / Шилдт Герберт. - 3-е изд. – Москва : Вильямс, 2003. - 429 с.