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 1-40 01 01 10 «Программное обеспечение информационных технологий (программирование интернет-приложений)»

**ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

**К КУРСОВОМУ ПРОЕКТУ НА ТЕМУ:**

«Разработка компилятора SED-2021»

Выполнил студент Сятковская Екатерина Дмитриевна

(Ф.И.О.)

Руководитель проекта ассистент Мущук Артур Николаевич

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

Заведующий кафедрой к.т.н. доц. Пацей Наталья Владимировна

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

Консультанты ассистент Мущук Артур Николаевич

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

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

Нормоконтролер ассистент Мущук Артур Николаевич

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

Курсовой проект защищен с оценкой

Минск 2021

Содержание

[Введение 5](#_Toc91019194)

[Глава 1. Спецификация языка программирования 6](#_Toc91019195)

[1.1 Характеристика языка программирования 6](#_Toc91019196)

[1.2 Алфавит языка 6](#_Toc91019197)

[1.3 Применяемые сепараторы 6](#_Toc91019198)

[1.4 Применяемые кодировки 6](#_Toc91019199)

[1.5 Типы данных 7](#_Toc91019200)

[1.6 Преобразование типов данных 8](#_Toc91019201)

[1.7 Идентификаторы 8](#_Toc91019202)

[1.8 Литералы 8](#_Toc91019203)

[1.9 Объявления данных и область видимости 8](#_Toc91019204)

[1.10 Инициализация данных 8](#_Toc91019205)

[1.11 Инструкции языка 8](#_Toc91019206)

[1.12 Операции языка 9](#_Toc91019207)

[1.13 Выражения и их вычисления 9](#_Toc91019208)

[1.14 Программные конструкции языка 10](#_Toc91019209)

[1.15 Область видимости идентификаторов 10](#_Toc91019210)

[1.16 Семантические проверки 10](#_Toc91019211)

[1.17 Распределение оперативной памяти на этапе выполнения 11](#_Toc91019212)

[1.18 Стандартная библиотека и её состав 11](#_Toc91019213)

[1.19 Ввод и вывод данных 11](#_Toc91019214)

[1.20 Точка входа 11](#_Toc91019215)

[1.21 Препроцессор 11](#_Toc91019216)

[1.22 Соглашения о вызовах 11](#_Toc91019217)

[1.23 Объектный код 12](#_Toc91019218)

[1.24 Классификация сообщений транслятора 12](#_Toc91019219)

[1.25 Контрольный пример 12](#_Toc91019220)

[Глава 2. Структура транслятора 13](#_Toc91019221)

[2.1 Компоненты транслятора, их назначение и принципы взаимодействия 13](#_Toc91019222)

[2.2 Перечень входных параметров транслятора 14](#_Toc91019223)

[2.3 Перечень протоколов, формируемых транслятором и их содержимое 14](#_Toc91019224)

[Глава 3. Разработка лексического анализатора 15](#_Toc91019225)

[3.1 Структура лексического анализатора 15](#_Toc91019226)

[3.2 Контроль входных символов 15](#_Toc91019227)

[3.3 Удаление избыточных символов 16](#_Toc91019228)

[3.4 Перечень ключевых слов, сепараторов, символов операций соответствующих им лексем 16](#_Toc91019229)

[3.5 Основные структуры данных 17](#_Toc91019230)

[3.6 Принцип обработки ошибок 18](#_Toc91019231)

[3.7 Структура и перечень сообщений лексического анализатора 18](#_Toc91019232)

[3.8 Параметры лексического анализатора и режим его работы 19](#_Toc91019233)

[3.9 Алгоритм лексического анализа 19](#_Toc91019234)

[3.10 Контрольный пример 20](#_Toc91019235)

[Глава 4. Разработка синтаксического анализатора 21](#_Toc91019236)

[4.1 Структура синтаксического анализатора. 21](#_Toc91019237)

[4.2 Контекстно-свободная грамматика, описывающая синтаксис языка 21](#_Toc91019238)

[4.3 Построение конченого магазинного автомата 23](#_Toc91019239)

[4.4 Основные структуры данных 24](#_Toc91019240)

[4.5 Описание алгоритма синтаксического разбора 24](#_Toc91019241)

[4.6 Структура и перечень сообщений синтаксического анализатора 24](#_Toc91019242)

[4.7 Параметры синтаксического анализатора и режимы его работы 24](#_Toc91019243)

[4.8 Принцип обработки ошибок 24](#_Toc91019244)

[4.9 Контрольный пример 25](#_Toc91019245)

[Глава 5. Разработка семантического анализатора 26](#_Toc91019246)

[5.1 Структура семантического анализатора 26](#_Toc91019247)

[5.2 Функции семантического анализатора 26](#_Toc91019248)

[5.3 Структура и перечень сообщений семантического анализатора 26](#_Toc91019249)

[5.4 Принцип обработки ошибок 26](#_Toc91019250)

[5.5 Контрольный пример 26](#_Toc91019251)

[Глава 6. Преобразование выражений 27](#_Toc91019252)

[6.1 Выражения, допускаемые языком 27](#_Toc91019253)

[6.2 Польская запись и принцип ее построения 27](#_Toc91019254)

[6.3 Программная реализация обработки выражений 28](#_Toc91019255)

[6.4 Контрольный пример 28](#_Toc91019256)

[Глава 7. Генерация кода 29](#_Toc91019257)

[7.1 Структура генератора кода 29](#_Toc91019258)

[7.2 Представление типов данных в памяти 30](#_Toc91019259)

[7.3 Статическая библиотека 30](#_Toc91019260)

[Глава 8. Тестирование транслятора 31](#_Toc91019261)

[8.1 Тестирование фазы проверки на допустимость символов 31](#_Toc91019262)

[8.2 Тестирование лексического анализатора 31](#_Toc91019263)

[8.3 Тестирование синтаксического анализатора 32](#_Toc91019264)

[8.4 Тестирование семантического анализатора 32](#_Toc91019265)

[Заключение 33](#_Toc91019266)

[Список использованных источников 34](#_Toc91019267)

[ПРИЛОЖЕНИЕ А 35](#_Toc91019268)

[ПРИЛОЖЕНИЕ Б 37](#_Toc91019269)

[ПРИЛОЖЕНИЕ В 39](#_Toc91019270)

[ПРИЛОЖЕНИЕ Г 41](#_Toc91019271)

[ПРИЛОЖЕНИЕ Д 43](#_Toc91019272)

[ПРИЛОЖЕНИЕ Е 45](#_Toc91019273)

[ПРИЛОЖЕНИЕ Ж 48](#_Toc91019274)

[ПРИЛОЖЕНИЕ K 50](#_Toc91019275)

# Введение

Задачей данного курсового проекта является разработка компилятора для языка программирования SED-2021. Он предназначен для выполнения простейших операций и арифметических действий над строками и числами.

Транслятор – это комплекс отдельных программ, позволяющих преобразовывать исходный код на одном языке программирования в исходный код на другом языке программирования. Транслироваться исходный код языка программирования SED-2021 будет в исходный код на языке ассемблера. Язык ассемблера – это машинно-ориентированный язык, представляющий формат записи машинных команд, которые понятны 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 с помощью польской инверсии;

– генерация кода на язык ассемблера;

– 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

* 1. **Характеристика языка программирования**

Язык программирования SED-2021 является строго типизированным, не объектно-ориентированным, транслируемым, процедурным языком программирования высокого уровня.

* 1. **Алфавит языка**

Алфавит языка программирования – набор символов, которые могут использоваться при написании исходного кода.

Символы, разрешенные к использованию при написании кода: символы латинского алфавита верхнего и нижнего регистров, арабские цифры, знаки препинания, знаки арифметических и логических операций.

* 1. **Применяемые сепараторы**

Сепараторы необходимы для разделения операций языка. Сепараторы, используемые в языке программирования SED-2021, приведены в таблице 1.1.

Таблица 1.1 – Сепараторы

|  |  |
| --- | --- |
| Сепаратор | Назначение |
| ;  :  « » (пробел)  , | Разделение конструкций |
| <, >, ==, != | Условные операторы. |
| = + - \* % | Арифметические операции |
| {} | Программные конструкции условного оператора |
| () | Параметры функций, изменение приоритетов в выражениях |
| [] | Ограничивает программные блоки |

* 1. **Применяемые кодировки**

Для написания исходного кода на языке программирования SED-2021 используется кодировка Windows-1251 – набор символов и кодировка, являющаяся стандартной 8-битной кодировкой для русских версий Microsoft Windows до 10-й версии – представленная на рисунке 1.1.



Рисунок 1.1 – Кодировка Windows-1251

* 1. **Типы данных**

В язык SED-2021 предусмотрены два типа данных: целочисленный и строковый, представленные в таблице 1.2. Пользовательские типы данных не поддерживаются.

Таблица 1.2 – Типы данных

|  |  |
| --- | --- |
| Тип | Описание |
| real | Фундаментальный тип данных, предназначенный для хранения целых чисел. В памяти занимает 2 байта. Автоматическая инициализация 0 (нулем). |
| char | Фундаментальный тип данных, предназначенный для хранения символьной информации. В памяти занимает 1 байт. Без явно указанной инициализации переменной, присваивается нулевое значение (пустая строка). |

* 1. **Преобразование типов данных**

Преобразование не поддерживается, все типы данных определены однозначно и не могут быть преобразованы в другие, так как язык SED-2021 является строго типизируемым.

* 1. **Идентификаторы**

Для именования функций, параметров и переменных используются идентификаторы. Не предусмотрены зарезервированные идентификаторы. Имя идентификатора составляется по следующим образом:

* состоит из символов латинского алфавита [a..z].
* максимальная длина идентификатора равна 9.
* цифры и знаки препинания запрещены.
  1. **Литералы**

В языке существует 2 вида литералов: литералы целого типа и строковые, описаны в таблице 1.3. Они осуществляют инициализацию переменных.

Таблица 1.3 – Литералы

|  |  |
| --- | --- |
| Тип | Описание |
| Литералы целого типа | Двоичное представление: 0zX, X - число в двоичной системе счисления. |
| Строковые литералы | Символ, помещённый в ' ' (код в Windows-1251: 39). |

* 1. **Объявления данных и область видимости**

В языке программирования SED-2021 переменная должны быть объявлена до ее использования. Областью видимости переменной является блок функции, в которой она определена. Допустимо объявление глобальных переменных. Область видимости схожа с областью видимости C++, то есть сверху вниз.

* 1. **Инициализация данных**

Язык SED-2021 поддерживает инициализацию данных и предусматривает инициализацию по умолчанию.

* 1. **Инструкции языка**

Инструкция для языка SED-2021 представлена в таблице 1.4.

Таблица 1.4 – Инструкция языка

|  |  |
| --- | --- |
| Инструкция | Запись на языке SED-2021 |
| Объявление переменной | create <тип данных> <идентификатор>; |
| Присваивание | <идентификатор> = <значение>| <идентификатор>; |
| Объявление внешней функции | <тип данных> def <идентификатор> (<тип данных> <идентификатор>, …) […]; |
| Точка входа | start  [  …  ]; |
| Возврат значения из подпрограммы | out <идентификатор> | <литерал>; |
| Вывод данных | write (<идентификатор> | <литерал>); |
| Условный оператор | if (<имя переменной, литерал><условный оператор><имя переменной, литерал>)  {…}  else  {…};  Блок else не обязателен. |

* 1. **Операции языка**

В языке SED-2021 предусмотрены следующие операции с данными. Приоритетность операции умножения выше приоритета операций сложения и вычитания. Для установки наивысшего приоритета используются круглые скобки. Операции языка представлены в таблице 1.5.

Таблица 1.5 — Операции языка

|  |  |
| --- | --- |
| Тип операции | Оператор |
| Арифметические | + – сложение  - – вычитание  \* – умножение  / – деление  = – присваивание  % - остаток от делений |
| Сравнение | > – больше  < – меньше  == – равно  != – не равно |

* 1. **Выражения и их вычисления**

В выражении должны участвовать операторы и операнды одного типа, а также функции, возвращающие значения того же типа. Круглые скобки в выражении используются для изменения приоритетов операций. Не допускается запись двух подряд арифметических операций. Также круглые скобки могут использоваться для передачи параметров функций. Квадратные скобки содержат блоки кода функций. В фигурных скобках содержатся блоки инструкций условного оператора.

* 1. **Программные конструкции языка**

Ключевые программные конструкции языка программирования SED-2021 представлены в таблице 1.6

Таблица 1.6 – программные конструкции

|  |  |
| --- | --- |
| Главная функция (точка входа в приложение) | start  [  …  out <имя переменной/литерал>;  ]; |
| Функция | <тип> def <идентификатор>(<тип> <идентификатор>)  [  …  out <выражение>;  ]; |

* 1. **Область видимости идентификаторов**

В языке SED-2021 переменные обязаны находится внутри программного блока функций. Все объявления и операции с переменными происходят внутри какого-либо блока. Каждая переменная или параметр функции получают постфикс – название функции, внутри которой они находятся. Объявление глобальных переменных не предусмотрено. Объявление пользовательских областей видимости не предусмотрено.

* 1. **Семантические проверки**

Таблица с перечнем семантических проверок, предусмотренных языком SED -2021, приведена в таблице 1.7.

Таблица 1.7 – Семантические проверки

| Номер | Правило |
| --- | --- |
| 1 | Вызов функции должен соответствовать её прототипу |
| 2 | Идентификатор должен быть объявлен до его использования |
| 3 | Операнды в арифметическом выражении не могут быть разных типов |
| 4 | Каждый идентификатор может быть объявлен только один раз |
| 5 | Проверка на превышение максимального размера строкового и целочисленного литералов |
| 6 | Соответствие типа возвращаемого значения с типом функции |
| 7 | Соответствие типов в выражениях |

Часть семантических проверок выполняется на этапе лексического анализа.

* 1. **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.

* 1. **Стандартная библиотека и её состав**

Функции стандартной библиотеки с описанием представлены в таблице 1.8. Стандартная библиотека написана на языке программирования C++.

Таблица 1.8 – Состав стандартной библиотеки

|  |  |  |  |
| --- | --- | --- | --- |
| Имя функции | Возвращаемое значение | Принимаемые параметры | Описание |
| StrConcat | char | char x – строка  char y – строка | Функция объединяет 2 строки |
| stringLen | real | char x - строка | Функция вычисляет длину строки x |
| WriteN | 0 | real x - число | Функция выводит на консоль число x |
| WriteW | 0 | char x - строка | Функция выводит на консоль строку x |

* 1. **Ввод и вывод данных**

Ввод данных не поддерживается языком программирования SED-2021.

write (<идентификатор или литерал>); – вывод в стандартный поток вывода.

В зависимости от типа параметра определяется функция: WriteW или WriteN, которые входят в состав стандартной библиотеки и описаны в таблице 1.8.

* 1. **Точка входа**

Точкой входа является функция start.

* 1. **Препроцессор**

Препроцессор в языке программирования SED-2021 не предусмотрен.

* 1. **Соглашения о вызовах**

В языке вызов функций происходит по соглашению о вызовах stdcall.

Особенности stdcall:

– все параметры функции передаются через стек;

– память освобождает вызываемый код;

– занесение в стек параметров идёт справа налево.

* 1. **Объектный код**

SED-2021 транслируется в язык ассемблера.

* 1. **Классификация сообщений транслятора**

В случае возникновения ошибки в коде программы на языке SED-2021 и выявления её транслятором в текущий файл протокола выводится сообщение. Их классификация сообщений приведена в таблице 1.9.

Таблица 1.9 - Классификация сообщений транслятора

|  |  |
| --- | --- |
| Интервал | Описание ошибок |
| 0-99 | Системные ошибки |
| 100-109 | Ошибки параметров |
| 110-115 | Ошибки открытия и чтения файлов |
| 116-129 | Ошибки лексического анализа |
| 600-699 | Ошибки синтаксического анализа |
| 700-799 | Ошибки семантического анализа |

* 1. **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 программирования SED-2021 в программу на языке ассемблера. Компонентами транслятора являются лексический, синтаксический и семантический анализаторы, а также генератор кода на языке ассемблера. Принцип взаимодействия представлен на рисунке 2.1.



Рисунок 2.1 – Структура транслятора

Лексический анализ – первая фаза трансляции. На вход лексический анализатор получает исходный код на языке программирования SED-2021, в котором сепараторами были разделены слова. Задачей лексического анализатора является нахождение лексических ошибок и формирование таблиц лексем и идентификаторов.

Синтаксический анализ – это основная часть транслятора, предназначенная для распознавания синтаксических конструкций. Входным параметров для синтаксического анализа является таблица лексем. Синтаксический анализ распознаёт синтаксические конструкции, выявляет синтаксические ошибки при их наличии и формирует дерево разбора.

Семантический анализ, в свою очередь, является проверкой исходного кода программы на семантическую согласованность с определением конструкций языка, то есть 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 кода на языке программирования SED-2021 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 SED-2021

|  |  |  |
| --- | --- | --- |
| Входной параметр | Описание | Значение по умолчанию |
| -in:<имя\_файла> | Входной файл с расширением .txt, в котором содержится исходный код на SED-2021 | Не предусмотрено |
| -log:<имя\_файла> | Файл для записи полного протокола работы транслятора | <имя\_файла>.log |
| -out:<имя\_файла> | Файл для записи результата работы транслятора | <имя\_файла>.asm |

Обязательным является только параметр –in.

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

Таблица с перечнем протоколов, формируемых транслятором языка SED-2021 и их назначением представлена в таблице 2.2

Таблица 2.2 – Протоколы, формируемые транслятором языка SED-2021

|  |  |
| --- | --- |
| Формируемый протокол | Описание протокола |
| Файл журнала с параметром <log> | Содержит информацию о входных параметрах в приложение и о этапе проверки символов на допустимость. |
| SED2021.asm | Содержит сгенерированный код на языке Ассемблера. |
| LT.txt | Результат работы лексического анализатора. Содержит таблицу лексем. |
| ID.txt | Результат работы лексического анализатора. Содержит таблицу идентификаторов. |
| RULS.txt | Результат работы синтаксического анализа. Содержит правила разбора, а так же трассировку. |

При отсутствии соответствующего ключа, файлы будут формироваться из имени файла с исходным кодом расширением по умолчанию.

# 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 SED-2021. На выходе формируется таблица лексем и таблица идентификаторов. Структура лексического анализатора представлена на рисунке 3.1.



Рисунок 3.1 Структура лексического анализатора SED-2021

Функции лексического анализатора:

— удаление “пустых” символов и пробелов. К “пустым” символом относятся: пробелы, табуляции, переходы в новую строку;

— распознавание идентификаторов и ключевых слов;

— распознавание констант;

— распознавание разделителей и символов-операций;

**3.2 Контроль входных символов**

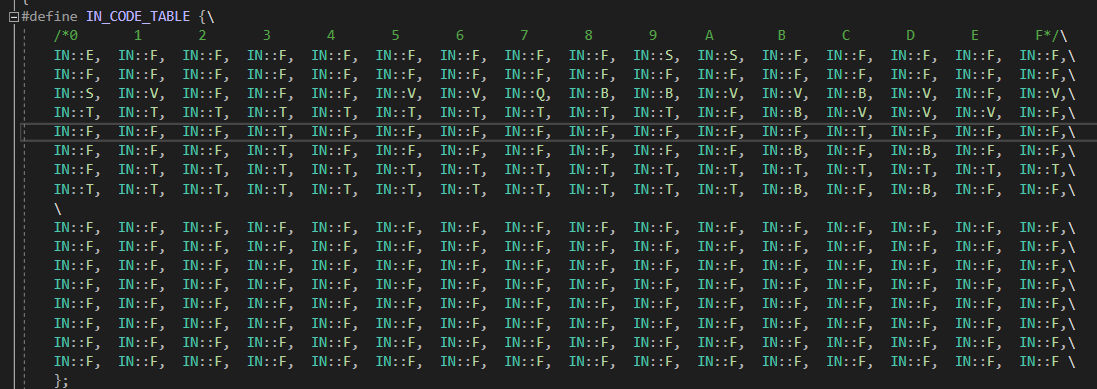
Таблица для контроля входных символов представлена на рисунке 3.2

Рисунок 3.2 - Таблица контроля входных символов

Принцип работы таблицы заключается в соответствии значения каждому элементу в шестнадцатеричной системе счисления значению в таблице ASCII.

Описание значения символов: T – разрешённый символ, F – запрещённый символ, S – сепаратор, V – арифметический символ, Q – символ ограничивающий литерал, B – блокообразующие символы, E – символ конца файла.

**3.3 Удаление избыточных символов**

Избыточными символами являются символы табуляции, переходы на новую строку и пробелы. Удаление избыточных символов происходит на этапе разбиения исходного текста на токены.

Описание алгоритма удаления избыточных символов:

1. Посимвольно считывается файл с исходным кодом программы.

2. Встреча пробела или знака табуляции является своего рода встречей символа-сепаратора.

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

## 3.4 Перечень ключевых слов, сепараторов, символов операций соответствующих им лексем

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 | create | c |
| real | n |
| char | w |
| def | p |
| write | r |
| StrConcat | x |
| stringLen | s |
| out | o |
| start | e |
| if | m |
| else | a |
| Иное | Идентификатор | i |
| Целочисленный литерал | l |
| Сепараторы | ; | ; |
| , | , |
| [ | [ |

Продолжение таблицы 3.1.

|  | ] | ] |
| --- | --- | --- |
| ( | ( |
| ) | ) |
| = | = |
| { | { |
| } | } |
| Операторы | -, +,\*,/,% | v |
|  | & | & |
|  | ! | ! |

Каждому выражению соответствует детерминированный конечный автомат, то есть автомат с конечным состоянием, по которому происходит разбор данного выражения. На каждый автомат в массиве подаётся фраза и с помощью регулярного выражения, соответствующего данному графу переходов, происходит разбор. В случае успешного разбора выражения оно записывается в таблицу лексем. Если выражение является идентификатором или литералом, информация также заносится в таблицу идентификаторов. Пример реализации таблицы лексем представлен в приложении Б.

Также в приложении В находятся конечные автоматы, соответствующие лексемам языка SED-2021.

## 3.5 Основные структуры данных

Основными структурами данных лексического анализатора являются таблица лексем и таблица идентификаторов. В таблице лексем содержится лексема, её номер, полученный при разборе и номер строки в исходном коде. В таблице идентификаторов содержится имя идентификатора, номер в таблице лексем, тип данных, смысловой тип идентификатора и его значение. Код структуры таблицы лексем на языке С++ представлен на рисунке 3.3, а код структуры таблицы идентификаторов на рисунке 3.4.

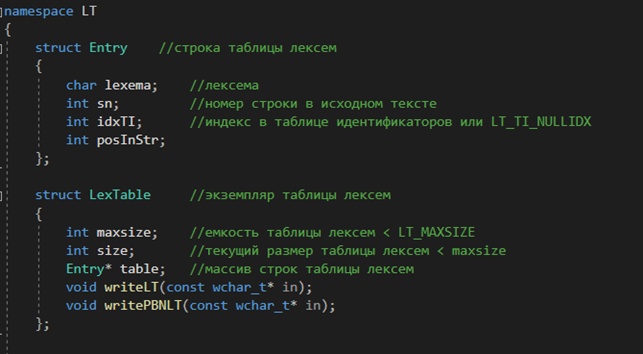


Рисунок 3.3 — Структуры таблиц лексем SED-2021

Kод структуры таблицы идентификаторов на языке С++ представлен на рисунке 3.4.

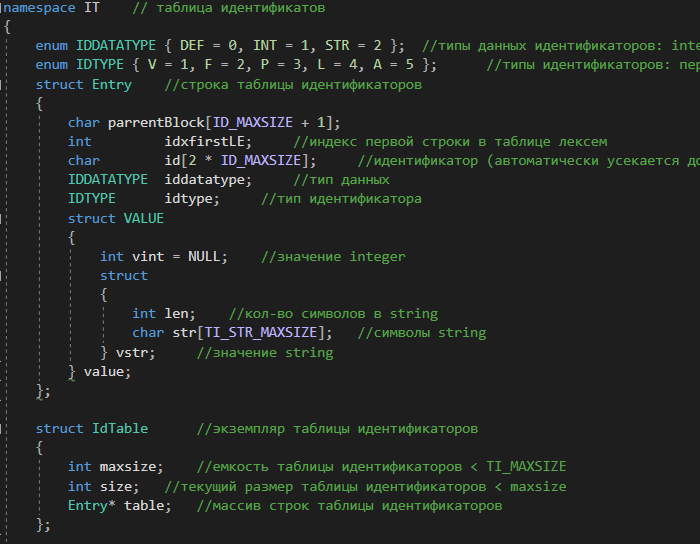
**

Рисунок 3.4 — Структуры таблиц идентификаторов SED-2021

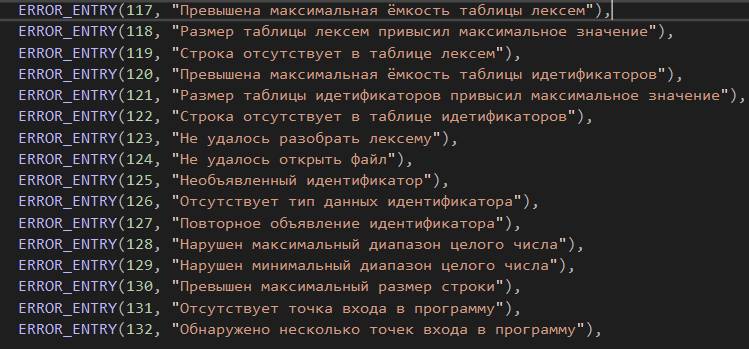
**3.6 Принцип обработки ошибок**

При возникновении ошибки типа предупреждение транслятор продолжает свою работу, а предупреждения записываются в специальную структуру с номером ошибки и диагностическим сообщением.

Когда возникает критическая ошибка – работа транслятора прекращается.

**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 режим его работы

Входным параметром лексического анализатора является исходный текст программы, написанный на языке SED-2021, а также 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 распознаёт и разбирает цепочки исходного текста программы. Этот разбор 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.

Пример. Регулярное выражение для ключевого слова def.

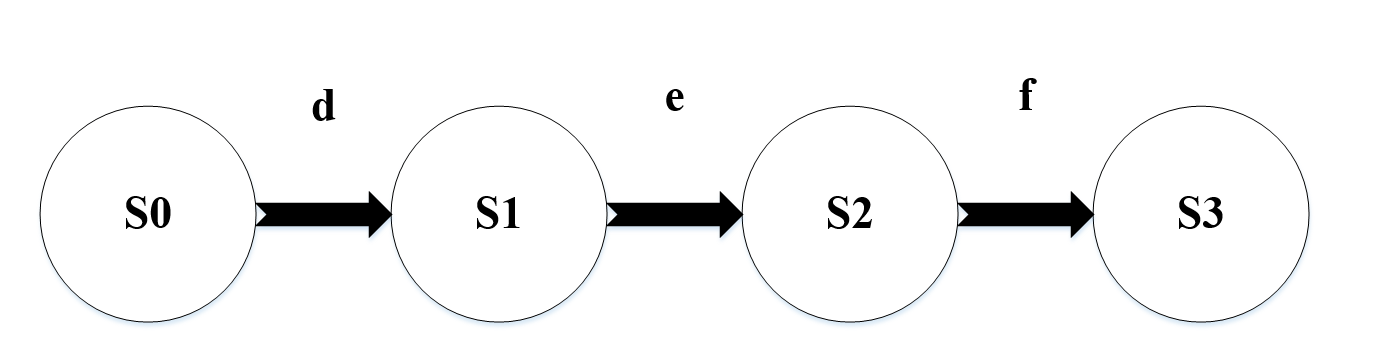
Граф конечного автомата для этой лексемы представлен на рисунке 3.6. S0 – начальное состояние, S3 – конечное состояние автомата.

Рисунок 3.6 — Граф переходов для цепочки “def”

В виде кода представлен на рисунке 3.7.

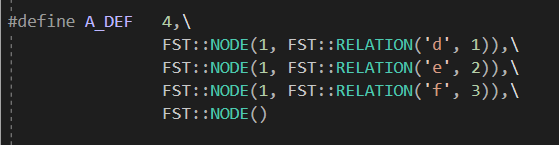


Рисунок 3.7 — Граф переходов для цепочки “def”

## 3.10 Контрольный пример

Результат работы лексического анализатора – вывод в файл таблицы лексем и идентификаторов – представлен 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 SED-2021 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.

Правила языка SED-2021 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 SED-2021

|  |  |  |
| --- | --- | --- |
| Нетерминал | Цепочки правил | Описание |
| S | e[NoE;];  tpi(F)[N];S  tpi(F)[N]; | Порождает правила, описывающее общую структуру программы |
| N | cti;N  i=E;N  rE;N  oE;N  i(W);N  s(W);N  x(W);N  mQ{N}a{N};N  mQ{N};N  cti;  rE;  s(W);  i=E;  oE;  x(W);  i(W);  mQ{N};  mQ{N}a{N}; | Порождает правила, описывающие конструкции языка |
| E | i  l  iM  lM  i(W)  (E)  (E)M  i(W)M  s(W)M  x(W)M  s(W)  x(W) | Порождает правила, описывающие выражения |
| F | ti  ti, F | Описывают параметры функции при объявлении |
| W | i  l  i, W  l, W | Порождает правила, описывающие принимаемые параметры функции |
| M | vE  vEM | Порождает правила, описывающие знаки арифметических операций |

Продолжение таблицы 4.1.

|  |  |  |
| --- | --- | --- |
| Q | (E<E)  (E>E)  (E&E)  (E!E) | Порождает правила условия в условных конструкциях. |

## 4.3 Построение конченого 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. |
|  | 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 SED-2021. 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 4.2.

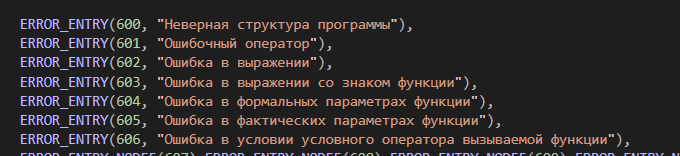


Рисунок 4.2 – Перечень сообщений синтаксического анализатора

## 4.7 Параметры синтаксического анализатора и режимы его работы

Входным параметром синтаксического анализатора является таблица лексем, полученная на этапе лексического анализа, поток 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 данного этапа обработки.

## 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.Все ошибки записываются в общую структуру 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

* 1. **Структура семантического анализатора**

Семантический анализ языка SED-2021 выполняется после выполнения лексического и синтаксического анализа. Несмотря на это, некоторые семантические проверки выполняются на этапе лексического анализа. На вход семантического анализатора подаются таблица лексем и таблица идентификаторов.

* 1. **Функции семантического анализатора**

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

* 1. **Структура и перечень сообщений семантического анализатора**

Сообщения, формируемые семантическим анализатором, представлены на рисунке 5.1.

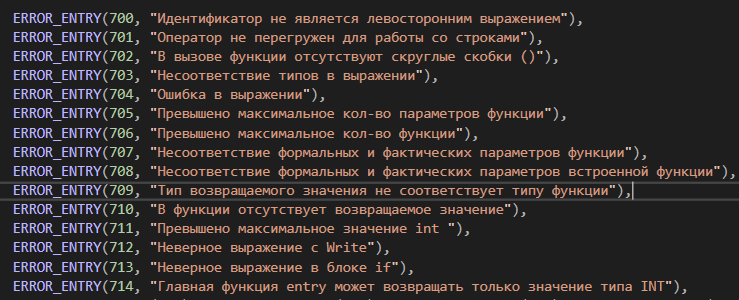


Рисунок 5.1 – Перечень сообщений семантического анализатора

* 1. **Принцип обработки ошибок**

Принцип обработки ошибок идентичен принципу обработки ошибок на этапе лексического анализа (раздел 3.6).

* 1. **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.

**Глава 6. Преобразование выражений**

**6.1 Выражения, допускаемые языком**

В языке SED-2021 допускаются вычисления выражений целочисленного типа данных, а также конкатенация строк, с поддержкой вызова функций внутри выражений. Приоритет операций представлен на таблице 6.1.

Таблица 6.1. Приоритеты операций

|  |  |
| --- | --- |
| Приоритет | Операция |
| 0 | ( ) |
| 1 | , |
| 2 | + - |
| 3 | / \* |
| 4 | < > |
| 5 | == != |

**6.2 Польская запись и принцип ее построения**

Выражения в языке SED-2021 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:

1. исходная строка: выражение;

2. результирующая строка: польская запись;

3. стек: пустой;

4. результирующая строка: польская запись;

5. исходная строка просматривается слева направо;

6. операнды переносятся в результирующую строку в порядке их следования;

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

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

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

10. отрывающая скобка помещается в стек;

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

12. 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;

13. 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 к польской записи представлена в приложении Ж.

**6.4 Контрольный пример**

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.

# 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.



Рисунок 7.1 - Структура генератора кода

Генератор кода последовательно проходит таблицу лексем, при необходимости обращаясь к таблице идентификаторов. В зависимости от пройденных лексем выполняется генерация кода ассемблера.

Обобщенная блок-схема алгоритма генерации кода языка ассемблера изображена на рисунке 7.2.

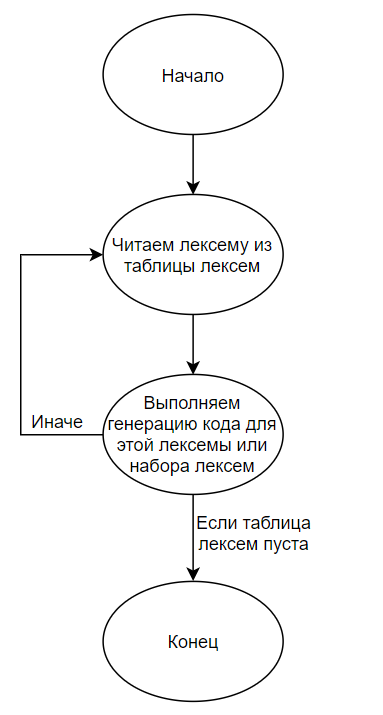


Рисунок 7.2 - Блок-схема алгоритма генерации кода языка ассемблер

## 7.2 Представление типов данных в памяти

Элементы таблицы идентификаторов расположены в разных сегментах языка ассемблера – .data и .const. Идентификаторы языка SED-2021 размещены в сегменте данных(.data). Литералы – в сегменте констант (.const). Соответствия между типами данных идентификаторов на языке SED-2021 и на языке ассемблера приведены в таблице 7.1. Сгенерированный код предоставлен в приложении К.

Таблица 7.1 – Соответствия типов идентификаторов языка SED-2021 и языка Ассемблера

|  |  |  |
| --- | --- | --- |
| Тип идентификатора на языке SED-2021 | Тип идентификатора на языке ассемблера | Пояснение |
| real | SDWORD | Хранит целочисленный тип данных со знаком. |
| char | DWORD | Хранит указатель на начало строки. |
| L(0-9) | BYTE  SDWORD | Литералы: строковые,  целочисленные |

## 7.3 Статическая библиотека

Функции из стандартной библиотеки содержатся в проекте SED2021Lib, в свойствах которого указан тип конфигурации «статическая библиотека».

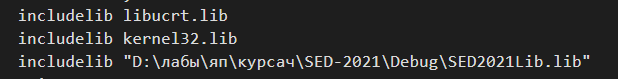


Рисунок 7.3 - Фрагмент генерации кода

Статическая библиотека содержит функции, написанные на языке C++.

# Глава 8. Тестирование транслятора

**8.1 Тестирование фазы проверки на допустимость символов**

В языке SED-2021 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 |
| real def test ”real a” | Ошибка 111: Недопустимый символ в исходном коде (-in)  Строка 1 Позиция: 14 |

* 1. **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 |
| start  [  ы  ]; | Ошибка 111: Недопустимый символ в исходном коде (-in)  Строка 3 Позиция: 1 |
| start  [  x = 5;  ]; | Ошибка 126: Использование необъявленного идентификатора  Строка 2 Позиция: 2 |
| start  [  real create hello;  ]; | Ошибка 126: Использование необъявленного идентификатора  Строка 2 Позиция: 6 |
| hello  [  ]; | Ошибка 124: Не объявлена точка входа в программу ( start ) |
| start  [  start[];  ]; | Ошибка 125: Определено несколько точек входа в программу (start)  Строка 2 Позиция: 1 |
| start  [  Create real nameoverflow;  ]; | Ошибка 121: Превышена длинна идентификатора  Строка 2 Позиция: 5 |

**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 |
| real def test (a, real b) | Ошибка 127: Неверное объявление параметров Строка 0 Позиция: 6 |
| real def eq(real a)  [  out 2;  ]; | Ошибка 600: Неверная структура программы  Строка 4 Позиция 1  Ошибка 602: Ошибка в выражении  Строка 3 Позиция 5  Ошибка 602: Ошибка в выражении  Строка 3 Позиция 5 |
| if(1 3)  {  …  } | Ошибка 602: Ошибка в выражении  Строка 2 Позиция 6  Ошибка 602: Ошибка в выражении  Строка 2 Позиция 6  Ошибка 602: Ошибка в выражении  Строка 2 Позиция 6 |

**8.4 Тестирование семантического анализатора**

Итоги тестирования семантического анализатора приведены в таблице 8.4.

Таблица 8.4 –Тестирование семантического анализатора

|  |  |
| --- | --- |
| Исходный код | Диагностическое сообщение |
| start  [  create char qa;  qa = ‘123’ + 5;  ]; | Ошибка 703: Несоответствие типов в выражении  Строка 8 Позиция: 0 |
| start  [  create char qa;  qa = ‘123’ + 5;  out ‘bb’;  ]; | Ошибка 714: Главная функция start может возвращать только значение типа INT  Строка 5 Позиция: 0 |

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

# Заключение

В ходе выполнения курсовой работы был разработан транслятор для языка программирования SED-2021. Таким образом, были выполнены основные задачи данной курсовой работы:

* Сформулирована спецификация языка SED-2021;
* Разработаны конечные автоматы и алгоритмы для реализация лексического анализатора;
* Разработана контекстно-свободная, приведённая к нормальной форме Грейбах, грамматика для описания синтаксически верных конструкций языка;
* Разработан семантический анализатор, осуществляющий проверку смысла используемых инструкций;
* Разработан транслятор с языка программирования SED-2021 на язык низкого уровня Assembler;
* Проведено тестирование всех вышеперечисленных компонентов.

Окончательная версия языка SED-2021 включает:

1. 2 типа данных;
2. поддержка операции вывода;
3. 2 библиотечные функции
4. возможность вызова функций стандартной библиотеки;
5. наличие 5 арифметических операторов для вычисления выражений;
6. 4 оператора сравнения для целочисленного типа.
7. структурированная система для обработки ошибок пользователя.
8. условный оператор;

# Список использованных источников

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

2. Смелов, В.В. Курс лекций по предмету языки программирования – 2016

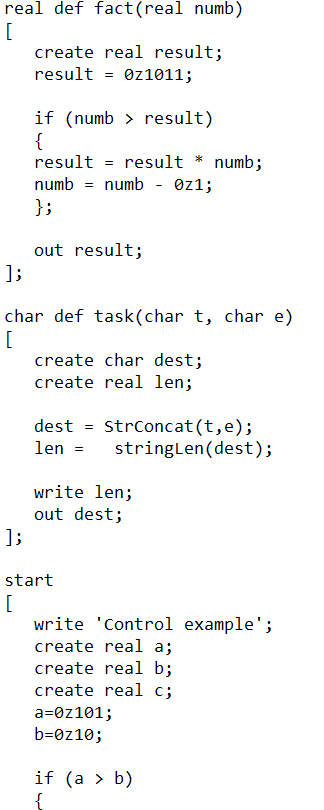
3. Прата, С. Язык программирования С++. Лекции и упражнения / С. Прата. – М., 2006 — 1104 c.

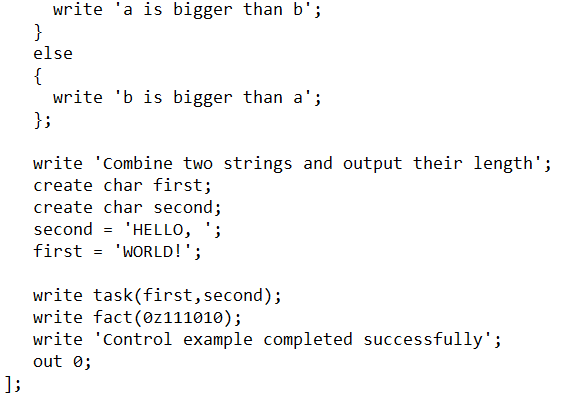
4. Страуструп, Б. Принципы и практика использования C++ / Б. Страуструп – 2009 – 1238 с.

5. [Руководство по программированию на языке С++ (metanit.com)](https://metanit.com/cpp/tutorial/)

# ПРИЛОЖЕНИЕ А

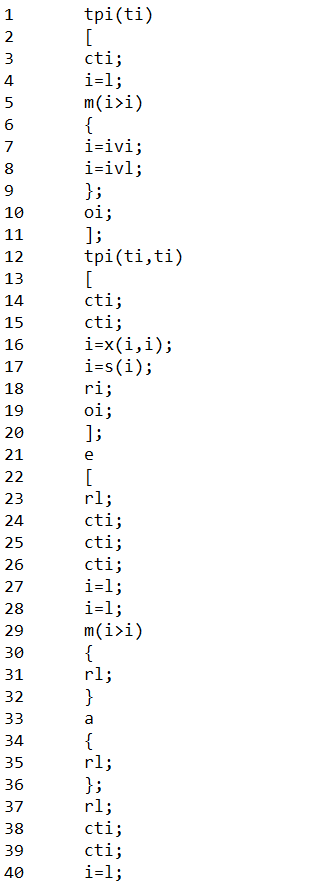
Исходный код на языке SED-2021





# ПРИЛОЖЕНИЕ Б

Таблица лексем на выходе лексического анализатора



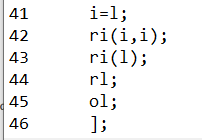
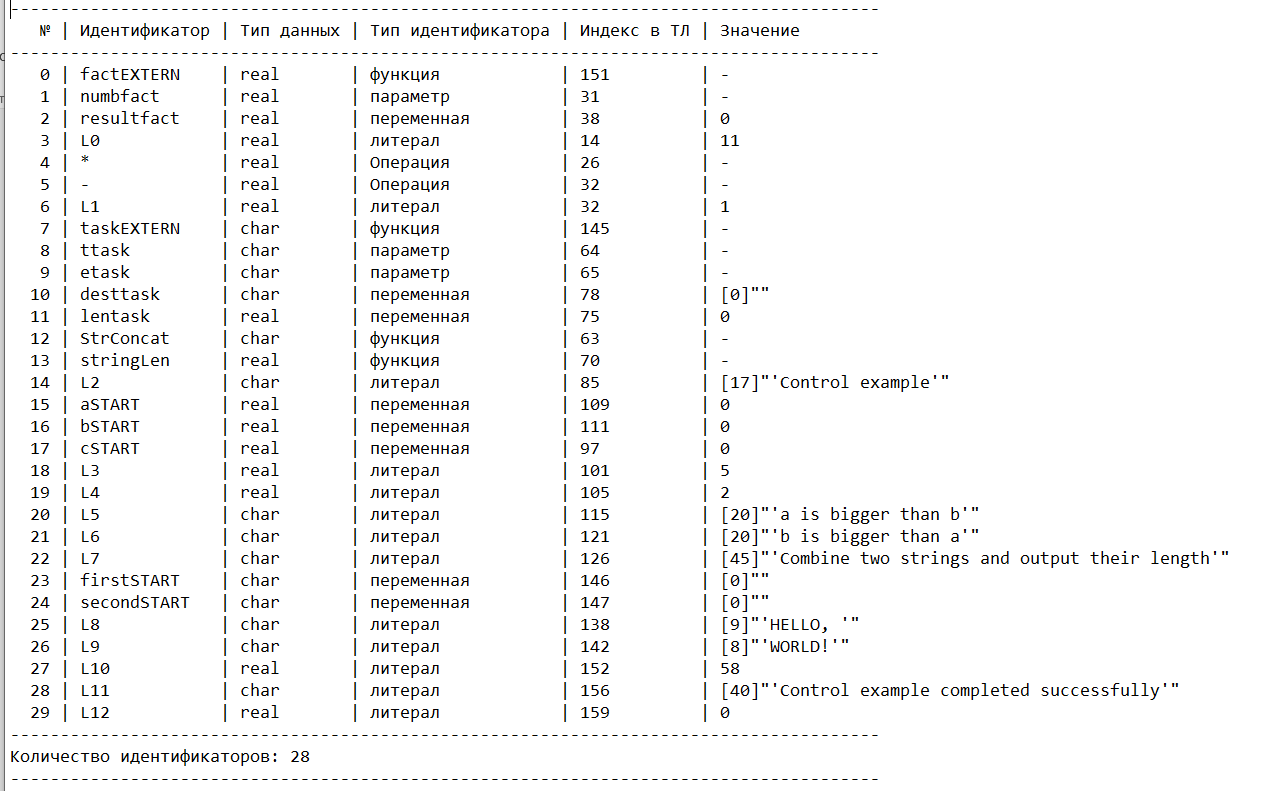
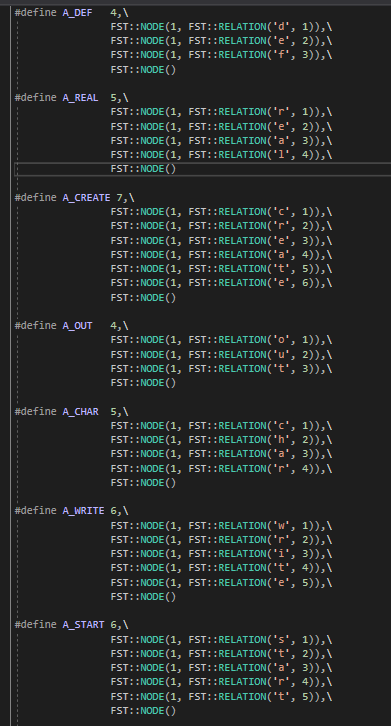


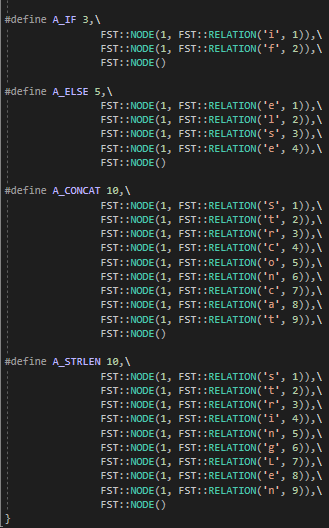
Таблица идентификаторов



# ПРИЛОЖЕНИЕ В

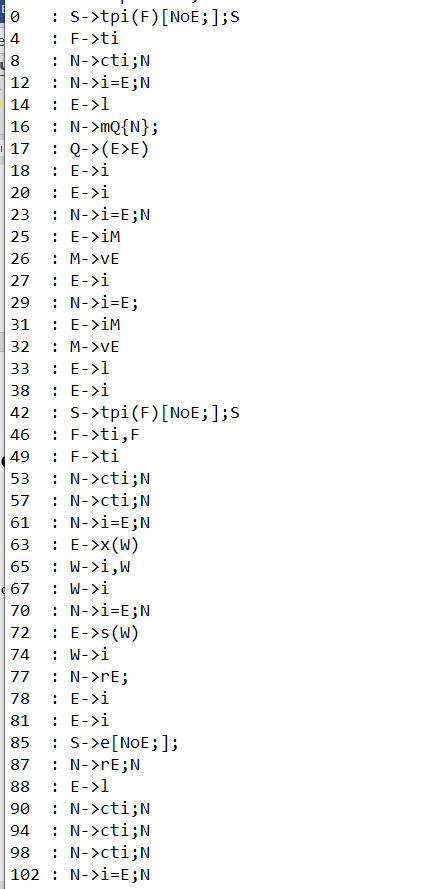
Структуры конечных автоматов для лексического распознавателя

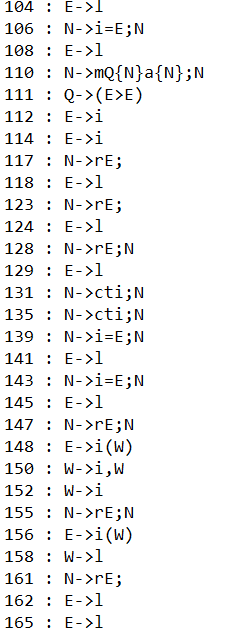




# ПРИЛОЖЕНИЕ Г

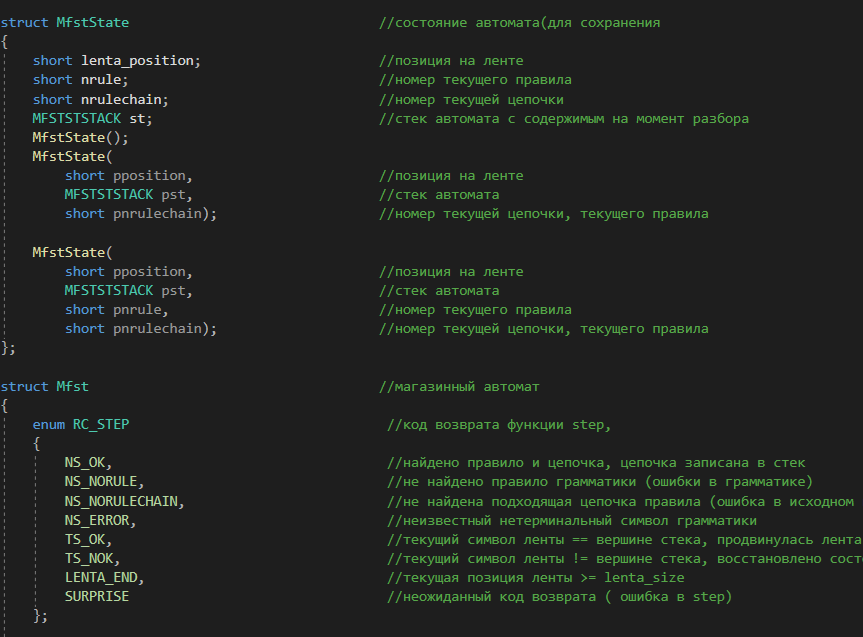
Правила синтаксического разбора

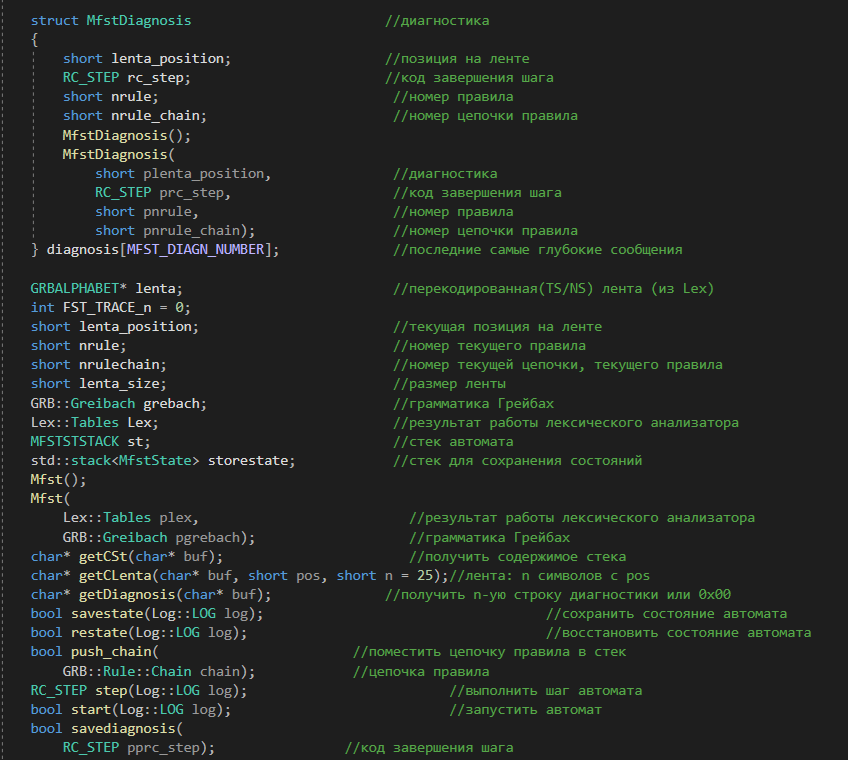


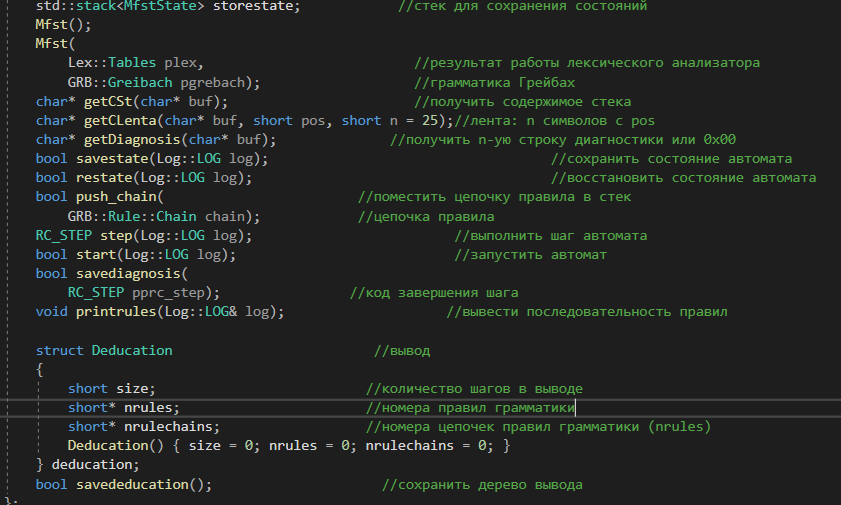


# ПРИЛОЖЕНИЕ Д

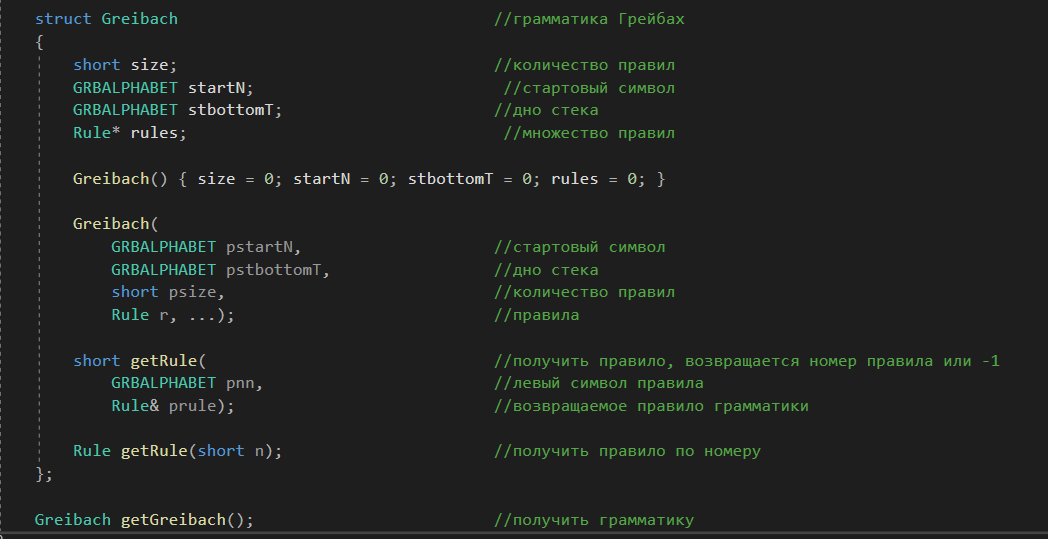
Структуры автомата синтаксического анализатора







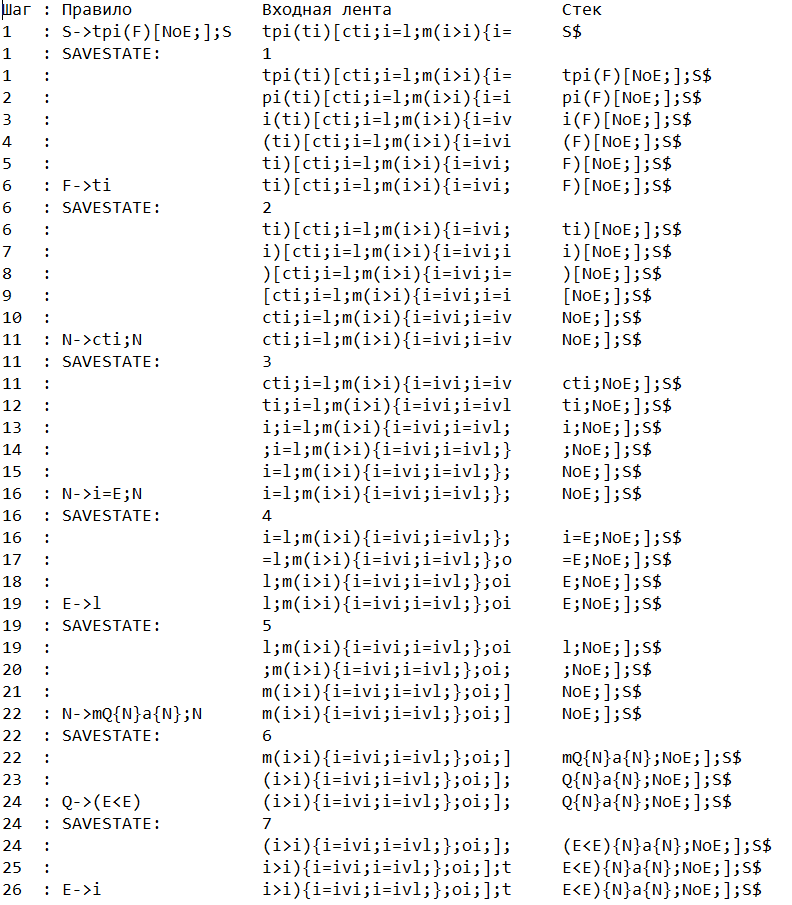
Структура грамматики Грейбах



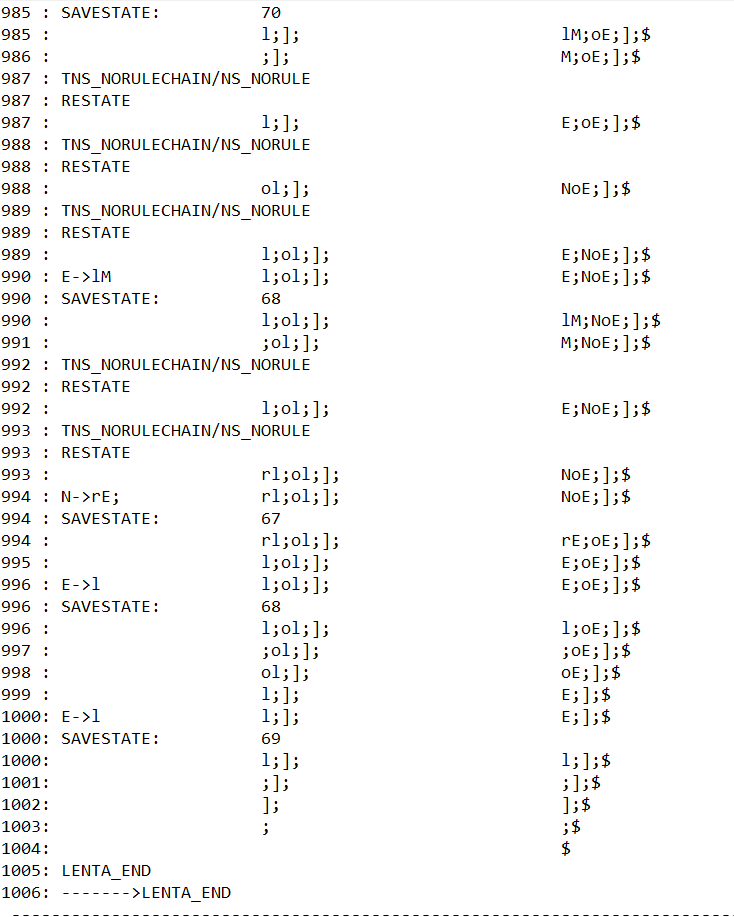
# ПРИЛОЖЕНИЕ Е

Пример трассировки синтаксического анализатора

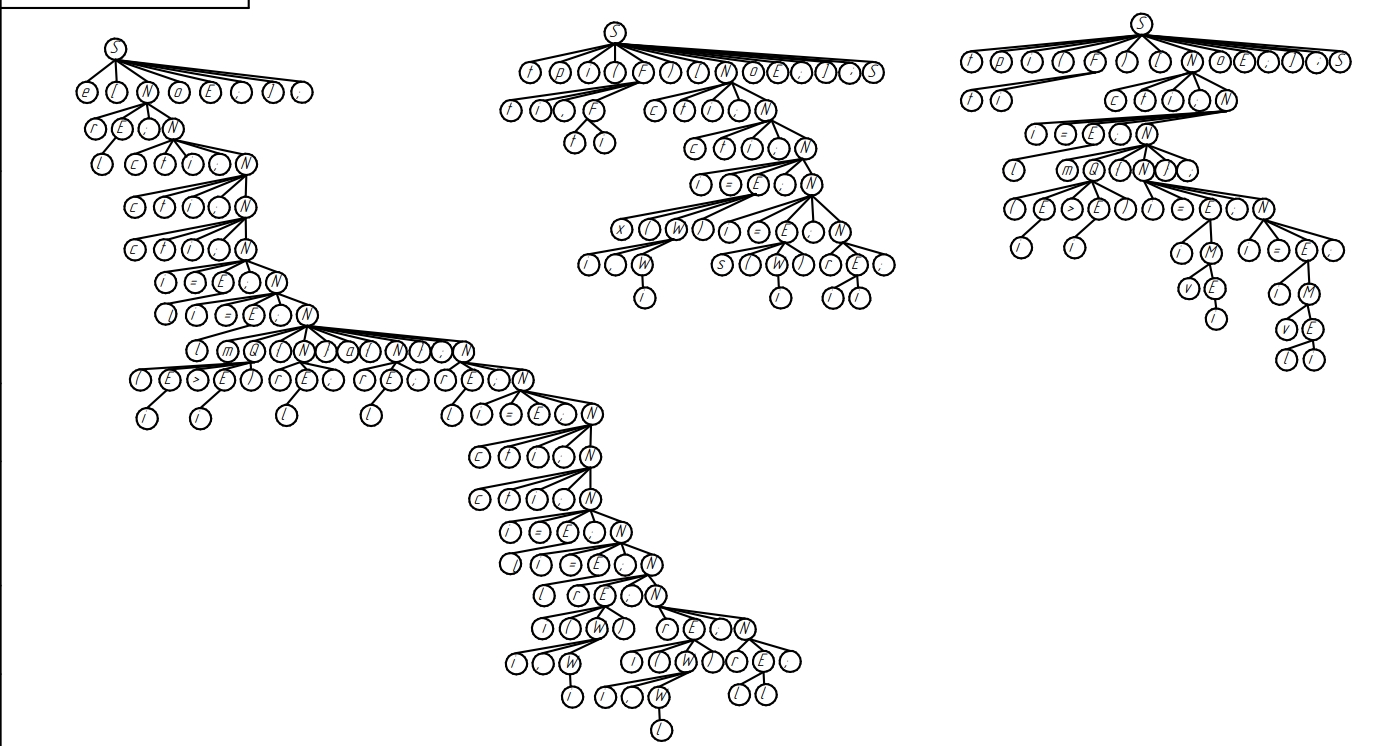
НАЧАЛО разбора



КОНЕЦ разбора

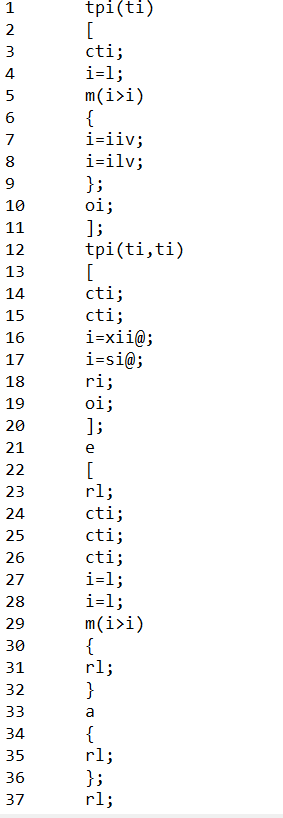


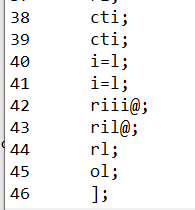
Дерево разбора



# ПРИЛОЖЕНИЕ Ж

Таблица лексем после преобразования к польской нотации





# ПРИЛОЖЕНИЕ K

Сгенерированный код

