Содержание

[Содержание 2](#_Toc185338842)

[Введение 4](#_Toc185338843)

[1. Спецификация языка программирования 5](#_Toc185338844)

[1.1 Характеристика языка программирования 5](#_Toc185338845)

[1.2 Определение алфавит языка программирования 5](#_Toc185338846)

[1.3 Применяемые сепараторы 5](#_Toc185338847)

[1.4 Применяемые кодировки 6](#_Toc185338848)

[1.5 Типы данных 6](#_Toc185338849)

[1.6 Преобразование типов данных 6](#_Toc185338850)

[1.7 Идентификаторы 6](#_Toc185338851)

[1.8 Литералы 6](#_Toc185338852)

[1.9 Объявление данных 7](#_Toc185338853)

[1.10 Инициализация данных 7](#_Toc185338854)

[1.11 Инструкции языка 7](#_Toc185338855)

[1.12 Операции языка 8](#_Toc185338856)

[1.13 Выражения и их вычисления 8](#_Toc185338857)

[1.14 Программные конструкции языка 9](#_Toc185338858)

[1.15 Область видимости идентификаторов 9](#_Toc185338859)

[1.16 Семантические проверки 10](#_Toc185338860)

[1.17 Распределение оперативной памяти на этапе выполнения 10](#_Toc185338861)

[1.18 Стандартная библиотека и её состав 10](#_Toc185338862)

[1.19 Ввод и вывод данных 11](#_Toc185338863)

[1.20 Точка входа 11](#_Toc185338864)

[1.21 Препроцессор 11](#_Toc185338865)

[1.22 Соглашения о вызовах 11](#_Toc185338866)

[1.23 Объектный код 11](#_Toc185338867)

[1.24 Классификация сообщений транслятора 11](#_Toc185338868)

[1.25 Контрольный пример 12](#_Toc185338869)

[2. Структура транслятора 13](#_Toc185338870)

[2.1 Компоненты транслятора их назначение и принципы взаимодействия 13](#_Toc185338871)

[2.2. Перечень входных параметров транслятора 13](#_Toc185338872)

[2.3 Протоколы, формируемые транслятором 14](#_Toc185338873)

[3. Разработка лексического анализатора 16](#_Toc185338874)

[3.1. Структура лексического анализатора 16](#_Toc185338875)

[3.2. Контроль входных символов 16](#_Toc185338876)

[3.3. Удаление избыточных символов 17](#_Toc185338877)

[3.4. Перечень ключевых слов 17](#_Toc185338878)

[3.5. Основные структуры данных 19](#_Toc185338879)

[3.6. Структура и перечень сообщений лексического анализатора 20](#_Toc185338880)

[3.7. Принцип обработки ошибок 20](#_Toc185338881)

[3.8. Параметры лексического анализатора 21](#_Toc185338882)

[3.9. Алгоритм лексического анализа 21](#_Toc185338883)

[3.10. Контрольный пример 22](#_Toc185338884)

[4. Разработка синтаксического анализатора 23](#_Toc185338885)

[4.1 Структура синтаксического анализатора 23](#_Toc185338886)

[4.2 Контекстно-свободная грамматика, описывающая синтаксис языка 23](#_Toc185338887)

[4.3. Построение конечного магазинного автомата 24](#_Toc185338888)

[4.4. Основные структуры данных 25](#_Toc185338889)

[4.5. Описание алгоритма синтаксического разбора 26](#_Toc185338890)

[4.6. Структура и перечень сообщений синтаксического анализатора 26](#_Toc185338891)

[4.7. Параметры синтаксического анализатора и режимы его работы 27](#_Toc185338892)

[4.8. Принцип обработки ошибок 27](#_Toc185338893)

[4.9. Контрольный пример 27](#_Toc185338894)

[5. Разработка семантического анализатора 28](#_Toc185338895)

[5.2. Функции семантического анализатора 28](#_Toc185338896)

Введение

Целью курсового проекта является разработка транслятора для собственного языка программирования: GMS-2024.

Написание транслятора будет осуществляться на языке C++, при этом код на языке GMS-2024 будет транслироваться в язык ассемблера.

Исходя из цели курсового проекта, были определены следующие задачи:

– разбработка спецификации языка программирования;

– разбратка структуры транслятора;

– разработка лексического и семантического анализаторов;

– разработка синтаксического анализатора;

– преобразование выражений;

– генерация кода на язык ассемблера;

– тестирование транслятора.

Решения каждой из поставленных задач будут приведены в соответствующих главах курсового проекта.

1. Спецификация языка программирования

1.1 Характеристика языка программирования

Язык программирования GMS-2024 – это язык высокого уровня. Он является процедурным, компилируемым, строго типизируемый.

1.2 Определение алфавит языка программирования

В основе алфавита языка GMS-2024 лежит таблица символов windows-1251

Таблица 1.1 – алфавит языка GMS-2024 в форме Бэкуса-Наура

|  |
| --- |
| <строчная буква латинского алфавита>::= a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z |
| <прописная буква латинского алфавита>::= A|B|C|D|E|F|G|H|I|J|K|L|M|N|O|P|Q|R|S|T|U|V|W|X|Y|Z |
| <цифра> ::= 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| <символ- сепаратор>::= ' '|,|(|)|{|}|;| "" |=| |
| <арифметическая операция>::= + | - | \* | / |
| <операции сравнения>::= > | < | <= | >= | == | != |

1.3 Применяемые сепараторы

Символы-сепараторы служат в качестве разделителей операций языка. Сепараторы, используемые в языке программирования GMS-2024, 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 |
| “…” | Двойные кавычки | Строковый литерал |
| ‘…’ | Одинарные кавычки | Допускается везде, кроме идентификаторов и ключевых слов |
| = | Знак «равно» | Присваивание значения |
| , | Запятая | Разделение параметров |
| / + - \* : | Знаки «косая черта», «плюс», «минус», «астерикс», «двоеточие», | Выражения |

Окончание таблицы 1.2

|  |  |  |
| --- | --- | --- |
| Сепаратор | Название | Область применения |
| < > >= <= == != | Знаки “меньше,” “больше”, “больше либо равно”, “меньше либо равно”, “равно”, “не равно” | Операции сравнения |

1.4 Применяемые кодировки

Для написания исходного кода на языке программирования GMS-2024 используется кодировка Windows-1251, которая представлена на рисунке 1.1.



Рисунок 1.1 – таблица символов кодировки windows-1251

1.5 Типы данных

В языке GMS-2024 реализованы 2 типа данных: целочисленный (2 байта) и символьный. Описание типов данных, предусмотренных в данном языке представлено в таблице 1.3.

Таблица 1.3 – Типы данных языка GMS-2024

|  |  |
| --- | --- |
| Типы данных | Описание типа данных |
| Целочисленный int | Фундаментальный тип данных. Предусмотрен для объявления целочисленных данных (2 байта). Инициализация по умолчанию: значение 0. Максимально допустимое значение 215-1. Минимально допустимым является -215. |
| Символьный char | Фундаментальный тип данных. Используется для работы с символом, который в памяти занимает 1 байт. Инициализация по умолчанию: символ конца строки “\0”. |
| Строковой str | Строковой тип данных. Каждый символ занимает 1 байт и принадлежит кодировке windows1251. Значение по умолчанию отсутствует. Максимальная длина 255, хранится в стеке как массив. Является ссылочным типом. |

Пользовательские типы данных не поддерживаются.

1.6 Преобразование типов данных

В языке программирования GMS-2024 преобразование типов данных не поддерживается. Все типы данных определенны однозначно и не могут быть преобразованы в другие.

1.7 Идентификаторы

В имени идентификатора допускаются только символы латинского алфавита и знак «\_» и цифры. Имя идентификатора не может совпадать с ключевыми словами языка и именами функций стандартной библиотеки и не должно начинаться с цифры. Формальное описание идентификатора можно представить в следующем виде:

<идентификатор> ::= <буква> {<буква> | <цифра> | '\_'}, где:

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

| u | v | w | x | y | z | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |

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

1.8 Литералы

В языке GMS-2024 предусмотрены 3 вида литералов: целочисленные, символьные и строковые. Краткое описание литералов приведено в таблице 1.4.

Таблица 1.4 – Описание литералов

|  |  |
| --- | --- |
| Литерал | Пояснение |
| Целочисленный | Целочисленные литералы, по умолчанию инициализируются 0. Представление: шестнадцатеричное (0x05), восьмеричное (05), двоичное (0b101), десятичное (5). |
| Символьный | Используются одиночные символы кодировки windows-1251, заключенные в одинарные кавычки: ‘5’. |
| Строковой | Допускается использование только внутри вызова функции потокового вывода. Заключены в двойные кавычки: “строка” |

Числовые литералы являются константами времени компиляции и объявляются только один раз.

1.9 Объявление данных

В языке GMS-2024 требуется объявить идентификатор перед его использованием. Все переменные должны находится внутри программного блока. Допускается объявление переменных с одинаковыми именами в разных программных блоках. Область видимости реализована по принципу «сверху вниз».

1.10 Инициализация данных

При объявлении переменной возможна ее инициализация литералом. Однако при компиляции в язык ассемблера, все переменные в сегменте данных будут иметь значение по умолчанию.

Таблица 1.5 – способы и присвоения значения переменной

|  |  |  |
| --- | --- | --- |
| Конструкция | Описание | Пример |
| <тип данных>  <идентификатор> =  <литерал>; | Объявление идентификатора и его инициализация. | int x = 10;  char c = ‘x’ |
| <тип данных>  <идентификатор>; | Объявление идентификатора и его автоматическая инициализация значением по умолчанию. | int x; |

Соответствие типов проверяется на этапе синтаксического анализа.

1.11 Инструкции языка

Все возможные инструкции языка программирования GMS-2024 представлены в общем виде в таблице 1.6.

Таблица 1.6 – Инструкции языка программирования GMS-2024

|  |  |
| --- | --- |
| Инструкция | Запись на языке GMS-2024 |
| Объявление переменной | <тип данных> <идентификатор>; |
| Объявление функции | <тип данных> <идентификатор> (<тип данных> <идентификатор>, …) {<блок кода>} |
| Вызов функции | <идентификатор функции> ( <параметры функции> | <пусто>); |
| Присваивание | <идентификатор> = <литерал> | <идентификатор> | <выражение>; |
| Блок инструкций | {  …  } |
| Возврат из подпрограммы | return <идентификатор> | <литерал> | <выражение>; |
| Условная инструкция | if(<условие>){<блок кода>}; |
| Вывод данных | print (<идентификатор> | <литерал> | <выражение>);  write <идентификатор> | <литерал> | <выражение>; |
| Ветвление | if (<условное выражение>) {<тело блока истины>} else [if (<условное выражение>)] {<тело блока лжи>} |

1.12 Операции языка

Язык программирования GMS-2024 может выполнять операции, представленные в таблице 1.7.

Таблица 1.7 – Операции языка программирования GMS-2024

|  |  |  |
| --- | --- | --- |
| Операция | Примечание | Пример |
| ( | Приоритет операций | (a + b) \* c; |
| ) |
| + | Суммирование | a + b; |
| - | Вычитание | a – b; |
| \* | Умножение | a \* b; |
| / | Деление | a / b; |
| == | Сравнение на равенство значений | a == b; |
| != | Сравнение на неравенство значений | a != b; |
| = | Присваивание | sum = 15;  chr = ‘T’; |
| <, > | Знаки «меньше», «больше» для условной инструкции | x < 3; |

1.13 Выражения и их вычисления

Круглые скобки в выражении используются для изменения приоритета операций. Два арифметических оператора не могут идти подряд. В выражении может быть вызов функции, если возвращаемый тип этой функции совместим с операциями, выполняемыми в данном выражении. Выражения не могут использоваться как аргументы функций, но могут быть использованы для присвоения значения переменной или в качестве логических операндов.

При генерации кода, для вычисления выражений используется польская запись.

1.14 Конструкции языка

Ключевые программные конструкции языка GMS-2024 представлены в таблице 1.8.

Таблица 1.8 – Программные конструкции языка GMS-2024

|  |  |
| --- | --- |
| Конструкция | Запись на языке GMS-2024 |
| Главная функция (точка входа) | main  {  …  [return <идентификатор> | <литерал>; ]  }; |
| Функция | <тип> <идентификатор> ([(<тип> <идентификатор>, …])  {  …  [return <идентификатор> | <литерал>; ]  }; |
| Условная конструкция | if(<условие>)  {  …  }  else [if…]  {  …  } |

В данной таблице представлены различные конструкции программирования и их реализация. Главная функция «main» является основной точкой входа в программу. Внешние функции определяются с помощью указания типа и названия, со списком параметров, который может отсутствовать, внутри круглых скобок. Условные конструкции позволяют выполнять различные кодовые блоки в зависимости от условий истинности или ложности.

1.15 Область видимости идентификаторов

В языке GMS-2024 все переменные, объявленные внутри блока кода, являются локальными, т.е. имеют функциональную область видимости. Они обязаны находится внутри программного блока функций. Объявление глобальных переменных не предусмотрено. Объявление пользовательских областей видимости не предусмотрено.

1.16 Семантические проверки

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

Таблица 1.8 – Семантические проверки

|  |  |
| --- | --- |
| № | Правило |
| 1 | Наличие точки входа main |
| 2 | Отсутствие переопределения идентификатора |
| 3 | Усекание длины имени идентификатора до 25 символов |
| 4 | Объявление идентификатора перед его использованием |
| 5 | Не допускается совпадения названий идентификаторов с ключевыми словами |
| 6 | Присвоение идентификатору значения, соответствующего его типу |
| 7 | Соответствие типов формальных и фактических параметров при вызове функций |
| 8 | Соответствие числа параметров при объявлении функции и ее вызове |
| 9 | Корректное использование операций сравнения и арифметических операций |
| 10 | Использование строго инициализированной переменной |
| 11 | Соответствие типов при различном их использовании в выражениях |
| 12 | Возвращаемое значение инструкцией return должно совпадать типу возвращаемого функцией значения, указанного при ее объявлении. |
| 13 | Деление на ноль |

Если семантическая проверка не проходит, то в лог журнал записывается соответствующая ошибка.

1.17 Распределение оперативной памяти на этапе выполнения

Все переменные размещаются в глобальной области памяти. Литералы заносятся в сегмент данных только для чтения (.const), а локальные переменные – в сегмент данных (.data). Эти секции используют статически выделенную память, доступную на протяжении всего времени выполнения программы. Локальная область видимости в исходном коде реализована с помощью правил префиксного именования переменных из разных блоков, что обуславливает их локальность в исходном коде.

1.18 Стандартная библиотека и её состав

Для использования функций стандартной библиотеки, не нужно явно подключать её. Все функции стандартной библиотеки доступны по умолчанию и не подлежат переопределению. Функции стандартной библиотеки с описанием представлены в таблице 1.9.

Таблица 1.9 – Состав стандартной библиотеки

|  |  |
| --- | --- |
| Функция | Назначение |
| int strcmp(str str1, str str2) | Принимает две строки и возвращает результат сравнения двух строк:  0, если 2 строки равны; 1, если первая строка лексикографически больше второй, иначе -1. |
| void print(int v) | Принимает значение любого типа и выводит его в консоль, автоматически добавляя символ новой строки в конце. |
| void print(char ch) |
| void print(str v) |

1.19 Ввод и вывод данных

В языке GMS-2024 не реализованы средства ввода данных. Для вывода данных в стандартный поток вывода предусмотрена функция print из стандартной библиотеки и оператор write.

1.20 Точка входа

В языке GMS-2024 каждая программа должна содержать главную функцию “main” - точку входа, с которой начнется последовательное выполнение программы. В программе может быть только одна точка входа.

1.21 Препроцессор

Препроцессор в языке программирования GMS-2024 не предусмотрен.

1.22 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.

1.23 Объектный код

GMS-2024 транслируется в язык ассемблера MASM.

1.24 Классификация сообщений транслятора

В случае возникновения ошибки в коде программы на языке GMS-2024 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 |
| 0-99 | Системные ошибки |
| 100-115 | Ошибки открытия и чтения файлов |
| 116-139 | Параметров |
| 140-200 | Ошибки лексического анализа |
| 130-139 | Ошибки таблиц лексем и таблиц идентификаторов |
| 600-699 | Ошибки синтаксического анализа |
| 700-900 | Ошибки семантического анализа |

Компилятор может обрабатывать до 1000 различных ошибок.

1.25 Контрольный пример

Контрольный пример представлен в Приложении А.

2. Структура транслятора

2.1 Компоненты транслятора их назначение и принципы взаимодействия

Язык GMS-2024 транслируется в язык ассемблера. Предварительно должен быть обработан исходный код, составлены таблицы идентификаторов и лексем, которые будут использованы для генерации кода.

Лексический анализатор разбивает исходный текст программы на лексемы, выделяя ключевые слова, идентификаторы, числа и другие элементы. Результатом его работы является последовательность лексем, которая также содержит ссылку на элементы таблицы идентификаторов.

Синтаксический анализатор проверяет последовательность лексем на соответствие грамматике языка GMS-2024. Он строит синтаксическое дерево, которое отображает структуру программы.

Семантический анализатор выполняет проверку логической корректности программы. Он контролирует типы данных, проверяет корректность операций и использования идентификаторов.

Генератор кода преобразует промежуточное представление исходного кода в текст программы на языке MASM. На этом этапе используются данные из таблицы идентификаторов для имен переменных и меток.

Принципы взаимодействия компонентов основываются на последовательной обработке данных. Исходный текст программы проходит через все этапы, начиная с лексического анализа и заканчивая генерацией кода. Каждый этап использует результаты предыдущего, обеспечивая корректность и точность трансляции.

2.2. Перечень входных параметров транслятора

Все доступные входные параметры транслятора, управляющие его работой, представлены в таблице 2.1

Таблица 2.1 – входные параметры транслятора языка GMS-2024

|  |  |  |
| --- | --- | --- |
| Параметр | Описание параметра | Значение по умолчанию |
| -in:<путь к файлу> | Задает путь к файлу с исходным кодом на языке GMS-2024, имеющим расширение .txt | Не предусмотрено |
| -log:<путь к файлу> | Файл журнала для вывода протоколов работы программы. | <путь -in файла>.log |
| -out:<путь к файлу> | Выходной файл с результатом предварительной обработки исходного кода | <путь -in файла>.out |

Окончание таблицы 2.1

|  |  |  |
| --- | --- | --- |
| Параметр | Описание параметра | Значение по умолчанию |
| -lt:<путь к файлу> | Указывает выходной путь таблицы лексем. Все директории из пути должны существовать, а файл должен иметь расширение .html | <путь -in файла>.LT.html |
| -it:<путь к файлу> | Указывает выходной путь таблицы идентификаторов. Все директории из пути должны существовать, а файл должен иметь расширение .html | <путь -in файла>.IT.html |
| -fo:<путь к файлу> | Указывает выходной путь исполняемого файла. | <путь -in файла>.exe |
| -stack:<число> | Задает размер стека в байтах | 1024 |
| /cst | Включает вывод на консоль дерева разбора. | Вывод отключен |
| /lex | Включает сохранение таблиц идентификаторов и лексем | Сохранение отключено |

2.3 Протоколы, формируемые транслятором

Во время работы транслятор генерирует различные протоколы, представляющие результаты промежуточных этапов. Все файлы создаются в каталоге, в котором расположен исполняемый файл, если иное не было задано входными параметрами. В таблице 2.2 приведена информация про формируемые протоколы.

Таблица 2.2 – протоколы, формируемые транслятором языка GMS-2024

|  |  |
| --- | --- |
| Протокол | Описание протокола |
| Файл журнала, задающийся параметром “-log” | Файл с журналом работы транслятора, содержащий основную информацию о результате работы транслятора: информацию про значения параметров, сведения об результатах работы синтаксического, семантического анализаторов, возникших ошибках. |
| Выходной файл предварительной обработки исходного кода, задающийся параметром “-out” | Файл содержит результат обработки исходного кода: текст, который поступит на лексический анализ. |

Окончание таблицы 2.2

|  |  |
| --- | --- |
| Протокол | Описание протокола |
| Выходной файл таблицы идентификаторов, задающийся параметром “-it” | Файл содержит представление таблицы идентификаторов, которая была получена в результате выполнения лексического анализа. |
| Выходной файл таблицы лексем, задающийся параметром “-lt” | Файл содержит представление таблицы лексем, которая была получена в результате выполнения лексического анализа. |
| Выходной файл с расширением “.asm” | Результат выполнения программы - исходный код на языке ассемблера. |

3. Разработка лексического анализатора

3.1. Структура лексического анализатора

Лексический анализатор — это программа, выполняющая разбор исходного текста и преобразование его в последовательность лексем. Структура лексического анализатора схематично представлена на рисунке 3.1.

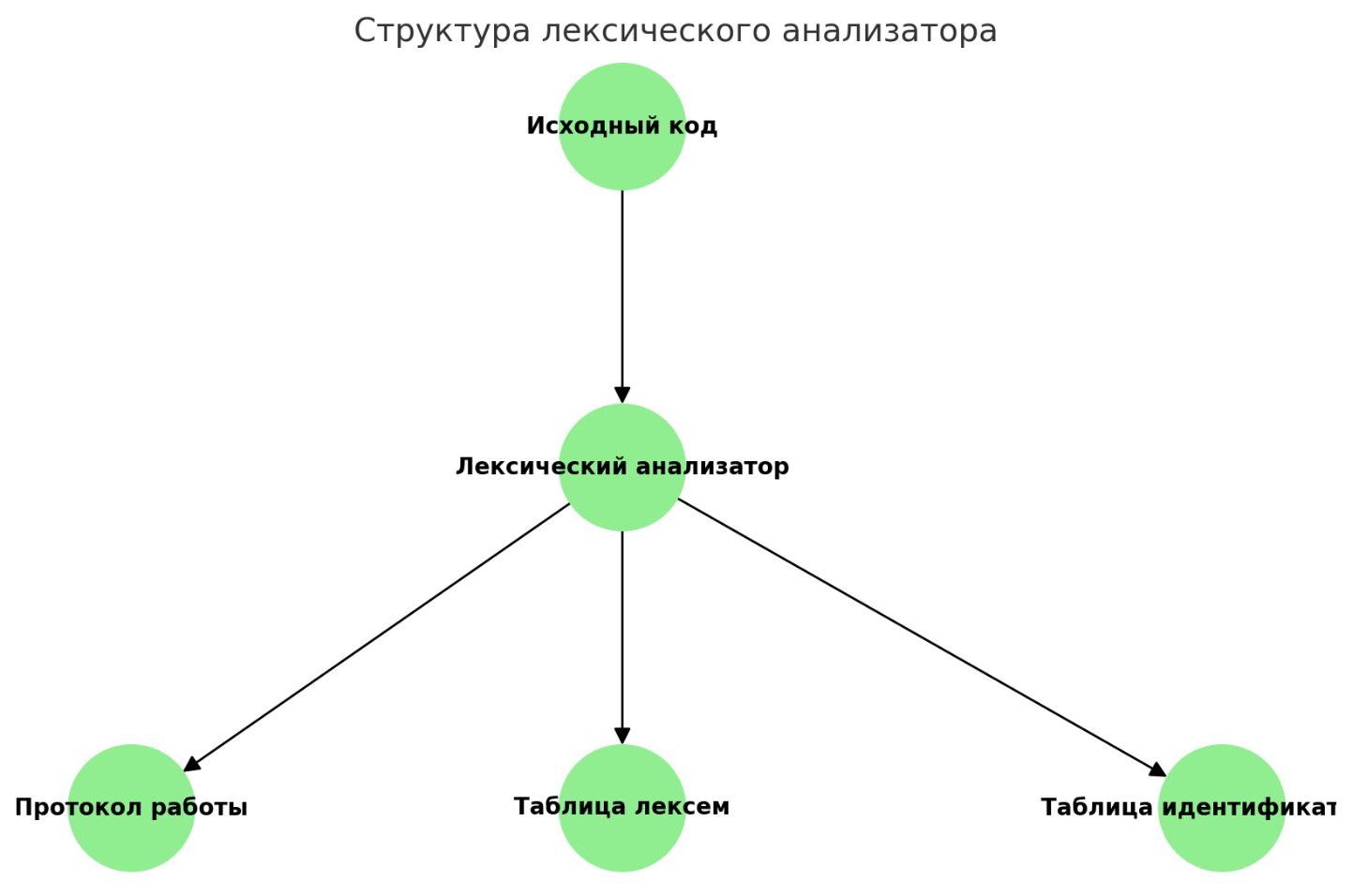


Рисунок 3.1 – структура лексического анализатора

Лексический анализатор принимает на вход исходный текст программы в текстовом формате. На выходе формируются таблицы распознанных лексем и идентификаторов.

3.2. Контроль входных символов

Для контроля правильности написания исходного текста используется таблица допустимых символов (рисунок 3.2). каждый элемент в ней соответствует коду символа в кодировке windows-1251 и представляет специальный “флаг” символа, определяющий, является ли символ допустимым и предопределяет обработку отдельных символов.

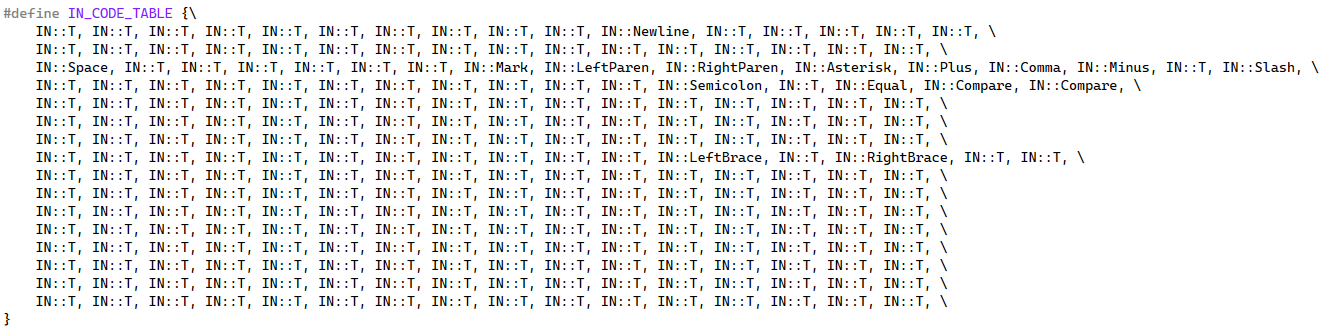


Рисунок 3.2 – таблица допустимых символов

3.3. Удаление избыточных символов

Для облегчения и шаблонизации лексического анализа используется предварительная обработка исходного кода, во время которой удаляются “избыточные” символы, а исходный код подготавливается к лексическому анализу.

“Избыточными” символами являются пробельные символы, стоящие там, где их присутствие не обязательно для разделения конструкций языка. Например, логические и арифметические операторы не обязаны быть разделены пробельными символами, а ключевые слова – да.

Алгоритм обработки исходного кода следующий: комментарии в исходном коде игнорируются, пробельные символы удаляются там, где их присутствие необязательно, а также там, где они дублируются, за исключением строковых литералов.

3.4. Перечень ключевых слов

Таблица 3.1 – ключевые слова, операторы и сепараторы языка GMS-2024

|  |  |  |
| --- | --- | --- |
| Токен языка | Лексема | Назначение |
| main | m | Определяет точку входа в программе |
| if | ? | Оператор условного перехода |
| else | : | Оператор перехода при невыполнении условия логического оператора if |
| return | r | Оператор выхода из функции и возврата ею некоторого значения |
| int | t | Указывает тип данных идентификатора как целочисленный |
| str | t | Указывает тип данных идентификатора как строковой |
| char | t | Указывает тип данных идентификатора как один символ |
| ( ) | ( ) | Круглые скобки используются для изменения приоритета операций, для вызова функций, а также в них заключено условие логического оператора if. |

Окончание таблицы 3.1

|  |  |  |
| --- | --- | --- |
| Токен языка | Лексема | Назначение |
| + - \* / | v | Арифметическая операция между целочисленными типами |
| < > <= >= == != | c | Оператор логического сравнения |
| { } | { } | Указывают начало и конец некоторого блока кода |

Для распознания цепочек из таблицы 3.1 используются конечный автомат, который использует регулярные выражения. Пример регулярного выражения для идентификаторов: [a-zA-Z\_][a-zA-Z0-9\_]. Его реализация на языке c++ представлена в листинге 3.1.

Листинг 3.1 – фрагмент кода на c++, реализующий конечный автомат

|  |
| --- |
| #define nums(state) \  R('0', state), R('1', state), R('2', state),\  R('3', state), R('4', state), R('5', state),\  R('6', state), R('7', state), R('8', state), R('9', state)  #define relationsForState(state) \  R('a', state), R('b', state), R('c', state), R('d', state), \  R('e', state), R('f', state), R('g', state), R('h', state), \  R('i', state), R('j', state), R('k', state), R('l', state), \  R('m', state), R('n', state), R('o', state), R('p', state), \  R('q', state), R('r', state), R('s', state), R('t', state), \  R('u', state), R('v', state), R('w', state), R('x', state), \  R('y', state), R('z', state), R('A', state), R('B', state), \  R('C', state), R('D', state), R('E', state), R('F', state), \  R('G', state), R('H', state), R('I', state), R('J', state), \  R('K', state), R('L', state), R('M', state), R('N', state), \  R('O', state), R('P', state), R('Q', state), R('R', state), \  R('S', state), R('T', state), R('U', state), R('V', state), \  R('W', state), R('X', state), R('Y', state), R('Z', state), R('\_', state)  FST::FST\* CreateIdentifierFST(const char\* str)  {  return new FST::FST(str,  2,  FST::NODE(NUM\_LETTERS \* 2 + 20,  relationsForState(0), nums(0),  relationsForState(1), nums(1)),  FST::NODE()  );  } |

3.5. Основные структуры данных

Для построения таблиц лексем и идентификаторов используются специальные структуры данных на языке c++.

Таблица идентификаторов представлена структурой ID\_Table (листинг 3.3), которая содержит объекты структуры Entry (листинг 3.2), представляющие собой запись про конкретный идентификатор. Она содержит ссылку на соответствующую ссылку в таблице лексем, символьное представление названия идентификатора, его тип и тип данных, значение идентификатора, а также область видимости – ссылку на другую запись в таблице идентификаторов, которая является функцией или блоком кода.

Листинг 3.2 – структура Entry

|  |
| --- |
| enum IDDATATYPE { INT = 1, STR, CHAR};  enum IDTYPE { V = 1024, F, P, L };  struct Entry // строка таблицы идентификаторов  {  int idxfirstLE; // индекс первой строки в таблице лексем  char id[ID\_MAXSIZE]; // идентификатор (автоматически усекается до ID\_MAXSIZE)  IDDATATYPE iddatatype; // тип данных  IDTYPE idtype; // тип идентификатора (переменная, функция, параметр, локальная переменная)  Entry\* scope;  union  {  int vint; // значение integer  struct  {  int len; // кол-во символов в string  char str[TI\_STR\_MAXSIZE - 1]; // символы string  } vstr[TI\_STR\_MAXSIZE]; // значение string  } value; // значение идентификатора  }; |

Таблица идентификаторов содержит текущий и максимальный размер, а также саму таблицу, которая представлена массивом объектов Entry.

Листинг 3.3 – структура ID\_Table

|  |
| --- |
| struct ID\_Table // экземпляр таблицы идентификаторов  {  int maxsize; // емкость таблицы идентификаторов < TI\_MAXSIZE  int size; // текущий размер таблицы идентификаторов < maxsize  Entry\* table; // массив строк таблицы идентификаторов  }; |

3.6. Структура и перечень сообщений лексического анализатора

После обнаружения ошибки лексический анализатор немедленно прекращает работу, выводит номер строки на консоль, опционально, дополнительную, поясняющую информацию, на консоль и записывает её в протокол.

Всевозможные ошибки перечислены в таблице 3.2

Таблица 3.2 – номера ошибок при лексическом анализе и их пояснения

|  |  |
| --- | --- |
| Номер ошибки | Текст ошибки |
| 140 | Может быть только одна точка входа main |
| 141 | Строковые и символьные литералы не могут быть пустыми |
| 142 | Идентификатор использован, но не объявлен |
| 143 | Идентификатору целочисленного типа присваивается значение строкового |
| 144 | Идентификатору строкового типа присваивается значение целочисленного типа |
| 145 | Превышен максимальный размер входного текста |
| 146 | Переопределение функции |
| 147 | Переопределение параметра функции |
| 148 | Переопределение переменной |
| 149 | Некорректный литерал |
| 150 | Идентификатор не может начинаться с цифры |

3.7. Принцип обработки ошибок

Все ошибки являются критическим, при возникновении первой ошибки, выполнение лексического анализа останавливается, выводится соответствующее сообщение об ошибке на консоль, программа завершает свое выполнение.

3.8. Параметры лексического анализатора

Некоторые входные параметры транслятора управляют сохранением таблиц идентификаторов и лексем. Они представлены в таблице 3.3.

Таблица 3.3 – параметры лексического анализатора

|  |  |  |
| --- | --- | --- |
| Параметр | Описание параметра | Значение по умолчанию |
| -lt:<путь к файлу> | Указывает выходной путь таблицы лексем. Все директории из пути должны существовать, а файл должен иметь расширение .html | <путь -in файла>.LT.html |
| -it:<путь к файлу> | Указывает выходной путь таблицы идентификаторов. Все директории из пути должны существовать, а файл должен иметь расширение .html | <путь -in файла>.IT.html |
| /lex | Включает сохранение таблиц идентификаторов и лексем | Сохранение отключено |

3.9. Алгоритм лексического анализа

Алгоритм лексического анализатора:

1. Входной текст считывается посимвольно. Символы сохраняются во временном буфере.
2. При встрече символа-сепаратора содержимое буфера разбирается конечным автоматам. При успешном распознании строки, конечный автомат возвращает лексему, соответствующую разобранному выражению. Если лексема была распознана, она сохраняется в таблицу лексем.
3. Если лексема равна лексеме типа, то в функции разбора строки устанавливается флаг, соответствующий типу данных. Иначе, если это лексема идентификатора – идет проверка из пункта 4.
4. Если перед лексемой идентификатора была сохранена лексема типа, то эта запись считается объявлением идентификатора, иначе выполняется пункт 5. В случае объявления, происходит проверка, нет ли идентификатора с таким именем и такой областью видимости для предотвращения переопределения. В случае успеха, идентификатор и лексема сохраняются в соответствующие таблицы и выполнение продолжается с пункта 1. В случае переопределения будет выброшено соответствующее исключение.
5. Если перед лексемой идентификатора не была сохранена лексема типа, то эта запись интерпретируется как использование ранее объявленного идентификатора. В таком случае выполняется его поиск в таблице идентификаторов и в случае отсутствия – выбрасывается соответствующее исключение. Иначе лексема сохраняется в таблице лексем.
6. Если лексема была распознана как литерал, будут осуществлены проверки на его корректность.
7. Если встречен символ ‘(‘, то будет осуществлена проверка предыдущих сохраненных лексем: если предыдущая лексема – это идентификатор, а предыдущая ей – тип, то предыдущий идентификатор будет расценен как функция или ее вызов. Это повлечет изменения в таблице идентификаторов для соответствующего идентификатора.
8. Если весь входной текст еще не разобран, то алгоритм возвращается к пункту 1, иначе завершается.

3.10. Контрольный пример

Результат лексического анализа в виде таблиц лексем и идентификатора представлены в приложении Б.

4. Разработка синтаксического анализатора

4.1 Структура синтаксического анализатора

Синтаксический анализ выполняется после лексического анализа и предназначен для распознавания заранее определенных синтаксических конструкций. Синтаксический анализатор принимает на вход таблицу лексем и генерирует дерево разбора или выявленные ошибки в синтаксисе.

4.2 Контекстно-свободная грамматика, описывающая синтаксис языка

В языке GMS-2024 синтаксический анализатор использует контекстно-свободную грамматику G = < T, N, P, S >, где:

– T – множество терминальных символов

– N – множество нетерминальных символов

– P – множество правил языка

– S – начальный символ грамматики (нетерминал)

Эта грамматика имеет нормальную форму Грейбах, так как она не содержит леворекурсивных правил.

Таблица 4.1 – синтаксические правила языка GMS-2024

|  |  |  |
| --- | --- | --- |
| Нетерминал | Цепочка (правило) | описание |
| S | ti(F){NrE;}S | Стартовый символ, порождающий всю структуру исходного кода |
| ti(F){rE;}S |
| ti(F){rE;}S |
| ti(F){rE;}S |
| m{NrE;} |
| m{N} |
| N | ti=E;N | Операторы программы |
| ti;N |
| i=E;N |
| rE;N |
| p(E);N |
| p(EcE);N |
| wE;N |
| wEcE;N |
| ?Y;N |
| i(W);N |
| i();N |
| ti=E; |
| ti; |
| i=E; |
| rE; |

Окончание таблицы 4.1

|  |  |  |
| --- | --- | --- |
| Нетерминал | Цепочка (правило) | описание |
| N | p(E); |  |
| p(EcE); |
| wE; |
| wEcE; |
| ?Y; |
| i(W); |
| i(); |
| E | l | Выражения |
| i |
| iM |
| lM |
| i(W)M |
| i(W) |
| i()M |
| i() |
| (E)M |
| (E) |
| A | l | Параметры стандартной функции strcmp |
| i |
| i(W) |
| i() |
| B | s(A,A) | Вызов стандартной функции strcmp |
| Y | (EcE){N}X | Условный оператор |
| (EcE){N} |
| (BcE){N}X |
| (BcE){N} |
| X | :{N} | Альтернативный блок |
| :?Y |
| W | i,W | Подвыражения |
| l,W |
| i |
| l |
| F | ti,F | Параметры функции |
| ti |
| M | vE M | Арифметические выражения |
| vE |

4.3. Построение конечного магазинного автомата

Формальное описание автомата с магазинной памятью можно представить в виде

– M = <Q, V, Z, δ, q0, z0, F>, где:

* Q – множество состояний управляющего устройства;

– V – алфавит входных символов;

– Z – специальный алфавит магазинных символов;

– δ – функция переходов автомата, представляющая собой множество правил, описанных в таблице 4.1;

– q0 ∈ Q – начальное состояние автомата;

– z0 ∈ Z – начальное состояние магазина (маркер дна);

– F⊆ Q – множество конечных состояний.

4.4. Основные структуры данных

Описание основных структур, описывающих контекстно-свободную грамматику, представлено в таблице 4.2.

Таблица 4.2 – описание основных структур синтаксического анализатора

|  |  |
| --- | --- |
| Структура | Описание |
| Greibach | Представляет полную грамматику Грейбах, состоящую из множества правил.  Поля:  – size: Количество правил в грамматике (short).  – startN: Стартовый символ грамматики (нетерминал).  – stbottomT: Специальный символ, обозначающий дно стека.  – rules: Указатель на массив правил (Rule\*). Методы:  – getRule(GRBALPHABET pnn, Rule& prule): Возвращает правило по указанному нетерминалу и сохраняет в prule.  – getRule(short n): Возвращает правило по его индексу. |
| Rule | Представляет правило в грамматике Грейбах.  Поля:   * nn: Левый символ правила (нетерминал, < 0). * iderror: Идентификатор диагностического сообщения (int). * size: Количество правых частей правила (short). * chains: Массив цепочек правых частей правила (Chain\*).   Методы:   * getCChain(char\* b): Возвращает правую часть правила как строку. * T(char t): Преобразует символ в терминал. * N(char n): Преобразует символ в нетерминал. * isT(GRBALPHABET s): Проверяет, является ли символ терминалом. * isN(GRBALPHABET s): Проверяет, является ли символ нетерминалом. * alphabet\_to\_char(GRBALPHABET s): Преобразует GRBALPHABET в символ char. |

Окончание таблицы 4.2

|  |  |
| --- | --- |
| Структура | Описание |
| Rule::Chain | Описывает цепочку символов (правую часть правила). Поля:   * size: Длина цепочки (short). * nt: Массив символов цепочки (GRBALPHABET\*), где   терминалы > 0, а нетерминалы < 0. |

Реализация структур данных семантического анализа на языке C++ представлена в приложении В.

4.5. Описание алгоритма синтаксического разбора

1. В стек (магазин) помещается стартовый символ и символ дна стека “$”.
2. Из таблицы лексем формируется входная лента.
3. Выбирается цепочка, которая соответствует текущему нетерминальному символу. Эта цепочка записывается в магазин.
4. Если терминалы из вершины стека и ленты совпадают – данный терминал удаляется из ленты и с вершины стека. Иначе происходит возврат к последнему сохраненному состоянию, после чего происходит переход к пункту 3, выбирая следующую подходящую цепочку.
5. Если в магазине встречается нетерминал, то происходит сохранение текущего состояния и переход к пункту 3.
6. Если мы достигаем дна стека и лента пуста – синтаксический анализ выполнен успешно. Если достигнуто дно стека, но лента не пуста – генерируется исключение.

4.6. Структура и перечень сообщений синтаксического анализатора

Возможные ошибки при синтаксическом анализе представлены с таблице 4.3.

Таблица 4.3 – Перечень сообщений синтаксического анализатора

|  |  |
| --- | --- |
| Код ошибки | Текст ошибки |
| 601 | Ошибочный оператор |
| 602 | Ошибка в выражении |
| 603 | Ошибка в параметрах функции или ее использовании |
| 604 | Ошибка в параметрах вызываемой функции |
| 605 | Ошибка в условии |
| 606 | Ошибка в условном операторе |
| 607 | Ошибка в операторе else |
| 608 | Ошибка в арифметическом выражении |
| 609 | Ошибка в параметрах функции strcmp |
| 610 | Ошибка в функции strcmp |

4.7. Параметры синтаксического анализатора и режимы его работы

Синтаксический анализатор принимает на вход таблицы лексем и идентификаторов, а также единственный параметр, задающийся из командной строки, /cst который включает вывод дерева разбора в консоль.

4.8. Принцип обработки ошибок

Ниже представлен принцип обработки ошибок синтаксического анализа.

1. Синтаксический анализатор последовательно проверяет правила грамматики и их цепочки, чтобы найти соответствие для конструкции, заданной в таблице лексем.
2. Если подходящая цепочка отсутствует, генерируется ошибка, отражающая данную ситуацию с пояснениями.
3. Все выявленные ошибки сохраняются в общей структуре, предназначенной для их регистрации.
4. По завершении процесса анализа в протокол и на консоль выводятся диагностические сообщения о найденных ошибках.

Всего в структуре может быть сохранено до 3 ошибок. В случае, когда их больше, старые ошибки перезаписываются новыми. При возникновении ошибки анализ продолжается для выявления других возможных ошибок.

4.9. Контрольный пример

**Результат работы протокола синтаксического разбора и дерево разбора представлены в приложении Г**

5. Разработка семантического анализатора

Семантический анализ — третья фаза работы транслятора, предназначенная для проверки исходного текста на соответствие семантическим правилам языка.

На вход семантический анализатор принимает информацию из таблиц лексем и идентификаторов. Результаты работы семантического анализатора записываются в файл протокола.

5.2. Функции семантического анализатора

Полный перечень семантических проверок приведен в таблице 5.1.

Таблица 5.1 – Перечень семантических проверок

|  |  |  |
| --- | --- | --- |
| № | Проверка | Фаза |
| 1 | Наличие точки входа main | Лексический |
| 2 | Отсутствие переопределения идентификатора | Лексический |
| 3 | Усекание длины имени идентификатора до 25 символов | Лексический |
| 4 | Объявление идентификатора перед его использованием | Лексический |
| 5 | Не допускается совпадения названий идентификаторов с ключевыми словами | Лексический |
| 6 | Присвоение идентификатору значения, соответствующего его типу | Лексический и семантический |
| 7 | Соответствие типов формальных и фактических параметров при вызове функций | Семантический |
| 8 | Соответствие числа параметров при объявлении функции и ее вызове | Семантический |
| 9 | Корректное использование операций сравнения и арифметических операций | Семантический |
| 10 | Использование строго инициализированной переменной | Семантический |
| 11 | Соответствие типов при различном их использовании в выражениях | Семантический |
| 12 | Возвращаемое значение инструкцией return должно совпадать типу возвращаемого функцией значения, указанного при ее объявлении. | Семантический |
| 13 | Деление на ноль | Время выполнения программы |

Семантический анализатор реализован множеством функций, которые осуществляют проверку определенных правил. Например, функция void detect\_cycles\_in\_references(const IT::ID\_Table& ID\_Table, const LT::LexTable& LEX\_Table) проверяет, есть ли «цикличные» вызовы функций: когда функции вызывают друг друга или сами себя.

5.3. Структура и перечень сообщений семантического анализатора

При работе семантического анализатора могут возникать различные ошибки. Например, при неправильном использовании идентификаторов и их значений. В таком случае семантический анализатор выводит в протокол и на консоль соответствующие данные.

Таблица 5.2 – перечень кодов ошибок семантического анализатора и их сообщения

|  |  |
| --- | --- |
| Код ошибки | Текст ошибки |
| 700 | Ошибка 700. Строка {}: переменная {} не может быть использована в выражении, так как она не инициализирована. |
| 702 | ошибка 702. Строка {}: операции сравнения не могут быть применены к данным разных типов. |
| 703 | Ошибка 703. Строка {}: операции сравнения не могут быть применены данным типа char. |
| 704 | Ошибка 704. Строка {}: оператор write не может участвовать в выражениях. |
| 705 | Ошибка 705. Строка {}: функция print не может участвовать в выражениях. |
| 706 | Ошибка 706. Строка {}: результат сравнения используется вне контекста сравнения. |
| 707 | Ошибка 707. Строка {}: выражение содержит тип(ы), которые не поддерживают операцию {}. |
| 708 | Ошибка 708. Строка {}: тип {} не совместим с типом {}. |
| 709 | Ошибка 709. Строка {}: аргумент функции {} имеет тип {}, но ожидается {}. |
| 710 | Ошибка 710. Строка {}: Присваивание значения функции непосредственно переменной недопустимо. |
| 711 | Ошибка 711. Строка {}: арифметические операции не могут быть применены к данным разных типов или сравнениям. |
| 712 | Ошибка 712. Строка {}, число передаваемых аргументов в функцию {} не соответствует ее прототипу. |
| 713 | Ошибка 713. Строка {}: Функция {} ссылается на {} и образует цикл. |

Примечание – в таблице 5.2 символы “{}” заменяются на номер строки, имена переменных, функций, операторов и другие динамические значения.

****5.4 Принцип обработки ошибок****

Ошибки, возникающие в процессе семантического анализа, фиксируются в протокол и дублируются на консоль. При этом выполнение семантического анализатора продолжается. После завершения анализа, если возникли ошибки, работа транслятора будет завершена.

5.5 Контрольный пример

В таблице 3.3 представлен пример семантической проверки. Полный перечень контрольного примера приведен в приложении Д.

Таблица 3.3 – пример теста контрольного примера

|  |  |
| --- | --- |
| Исходный код | Сообщение |
| main  {  str s = "test";  int z = 12 + s;  } | Ошибка 707. Строка 4: выражение содержит тип(ы), которые не поддерживают операцию + |

6.1 Вычисление выражений

6.1. Выражения, допускаемые языком

В языке GMS-2024 допускаются вычисления выражений целочисленного типа данных с поддержкой вызова функций внутри выражений. Приоритет операций представлен на таблице 6.1.

Таблица 6.1. – Приоритеты операций

|  |  |
| --- | --- |
| Операция | Значение приоритета |
| ( ) | 3 |
| \* | 2 |
| / | 2 |
| + | 1 |
| - | 1 |

Сначала вычисляются операторы внутри скобок, затем умножение или деление, после сложение или вычистание.

6.2 Польская запись и принцип её построения

Польская запись является способом представления выражений, который заключается в помещении операндов и операций в стек, для облегчения их вычислений.

Алгоритм построения польской записи:

* Если встречается операнд, он добавляется в результирующую строку преобразования.
* Если встречается вызов функции, он помещается в стек.
* Если встречается операция, то:
  + Если стек не пуст, приоритет текущей операции сравнивается с приоритетом операции на вершине стека. Пока приоритет операции из вершины стека выше или равен приоритету текущей операции, операция извлекается из стека и переносится в результирующую строку.
  + Текущая операция помещается в стек. Если стек изначально был пуст, операция сразу добавляется в стек.
* Если встречается открывающая скобка, она помещается в стек.
* Если встречается закрывающая скобка, операции извлекаются из стека и добавляются в результирующую строку до тех пор, пока не встретится открывающая скобка. Открывающая скобка также извлекается из стека.
* Если выражение полностью обработано, но стек не пуст, все оставшиеся операции извлекаются из стека и добавляются в результирующую строку.

Программная реализация польской запись представлена в приложении Ж.

6.4 Контрольный пример

Пример преобразования выражения приведен в таблице 6.2. В приложении З приведены все результаты преобразования выражений из таблицы лексем

Таблица 6.2 – пример преобразования в польскую запись

|  |  |
| --- | --- |
| Исходный код | Польская запись |
| (-1\*(-7))+5 | \_\_INT\_L4\_\_INT\_L5\*\_\_INT\_L6+ |