Национальный исследовательский университет ИТМО

Мегафакультет компьютерных технологий и управления

Факультет программной инженерии и компьютерной техники

Отчет по практическому заданию №1

по курсу «Языки программирования»

Вариант: 2

|  |  |  |
| --- | --- | --- |
| Выполнил студент группы P4119 | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Д.Л. Симоновский |
|  | (подпись) |  |

|  |  |  |
| --- | --- | --- |
| Руководитель | \_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Ю. Д. Кореньков |
|  | (подпись) |  |

22 октября 2025 г.

Санкт-Петербург

2025

Оглавление

[1. Цель работы 2](#_Toc210574064)

[2. План работы 2](#_Toc210574065)

[3. Инструментальные средства 2](#_Toc210574066)

[4. Ход работы 2](#_Toc210574067)

[4.1. Описание алгоритма 2](#_Toc210574068)

[4.2. Реализация модуля на SystemVerilog 3](#_Toc210574069)

[4.3. Симуляционное тестирование 8](#_Toc210574070)

[4.4. Характеристики разработанного устройства 10](#_Toc210574071)

[4.5. Тестирование проекта на ПЛИС 11](#_Toc210574072)

[5. Вывод 12](#_Toc210574073)

# Цель работы

Использовать средство синтаксического анализа по выбору, реализовать модуль для разбора текста в соответствии с языком по варианту. Реализовать построение по исходному файлу с текстом синтаксического дерева с узлами, соответствующими элементам синтаксической модели языка. Вывести полученное дерево в файл в формате, поддерживающем просмотр графического представления.

# План работы

1. Изучить выбранное средство синтаксического анализа.
2. Изучить синтаксис разбираемого по варианту языка и записать спецификацию для средства синтаксического анализа.
3. Реализовать модуль, использующий средство синтаксического анализа для разбора языка по варианту.
4. Реализовать тестовую программу для демонстрации работоспособности созданного модуля.

# Ход работы

## Разработка синтаксического анализатора

Для разработки синтаксического анализатора была выбрана библиотека tree-sitter, на ней уже есть готовые анализаторы для СИ, C++ и других языков, что поможет при проектировании своего анализатора.

Выполним реализацию общей части синтаксической модели для всех вариантов:

|  |
| --- |
| identifier: "[a-zA-Z\_][a-zA-Z\_0-9]\*"; // идентификатор  str: "\"[^\"\\]\*(?:\\.[^\"\\]\*)\*\""; // строка, окруженная двойными кавычками  char: "'[^']'"; // одиночный символ в одинарных кавычках  hex: "0[xX][0-9A-Fa-f]+"; // шестнадцатеричный литерал  bits: "0[bB][01]+"; // битовый литерал  dec: "[0-9]+"; // десятичный литерал  bool: 'true'|'false'; // булевский литерал  list<item>: (item (',' item)\*)?; // список элементов, разделённых запятыми |

В tree-sitter эти элементы будут выглядеть следующим образом:

Листинг 3.1. Базовые элементы грамматики

|  |
| --- |
| identifier: \_ => /[a-zA-Z\_][a-zA-Z\_0-9]\*/,  str: \_ => /\"[^\"\\]\*(?:\\.[^\"\\]\*)\*\"/,  char: \_ => /\'[^\']\'/,  hex: \_ => /0[xX][0-9A-Fa-f]+/,  bits: \_ => /0[bB][01]+/,  dec: \_ => /[0-9]+/,  \_true: \_ => token('true'),  \_false: \_ => token('false'),  bool: $ => choice($.\_true, $.\_false), |

Как можно заметить, в tree-sitter грамматика пишется на js, однако после этого компилируется в быстрый СИ-код. В приведенном фрагменте отсутствует описание списка элементов т.к. он будет реализовываться для каждого конкретного случая отдельно.

Далее будет приведена реализация блоков для варианта 2. Разработка выполнялась снизу вверх, сначала expr, затем statement, а потом уже верхнеуровневое описание sourceItem.

Ниже приведен блок синтаксической модели для expr:

|  |
| --- |
| expr: { // присваивание через ':='  |binary: expr binOp expr; // где binOp - символ бинарного оператора  |unary: unOp expr; // где unOp - символ унарного оператора  |braces: '(' expr ')';  |call: expr '(' list<expr> ')';  |indexer: expr '[' list<expr> ']';  |place: identifier;  |literal: bool|str|char|hex|bits|dec;  }; |

Аналогичным образом выглядит этот же блок в tree-sitter:

Листинг 3.2. Блок expr в tree-sitter

|  |
| --- |
| expression: $ => choice(    field('unary', $.unary\_expression),    field('binary', $.binary\_expression),    field('braces', $.braces\_expression),    field('call', $.call\_expression),    field('indexer', $.indexer),    field('place', $.identifier),    field('literal', $.literal),  ), |

Каждый из полей ссылается на другое правило. Самые простые – последние два. Place напрямую ссылается на идентификатор, объявленный раньше, literal же является простым выбором между объявленными ранее элементами:

Листинг 3.3. Поле literal в tree-sitter

|  |
| --- |
| literal: $ => choice($.bool, $.str, $.char, $.hex, $.bits, $.dec), |

Далее идут очень схожие конструкции: call и indexer:

Листинг 3.4. Поля indexer и call

|  |
| --- |
| indexer: $ => prec(PREC.SUBSCRIPT, seq(    field('expr', $.expression),    seq('[', optional(field('listExpr', $.list\_expr)), ']')  )),  call\_expression: $ => prec(PREC.CALL, seq(    field('expr', $.expression),    seq('(', optional(field('listExpr', $.list\_expr)), ')')  )),  list\_expr: $ => seq(    field('expr', $.expression),    repeat(seq(',', field('expr', $.expression)))  ), |

PREC задает приоритет операций вызова, это enum с значениями приоритета, взятый из репозитория для парсинга языка СИ, он выглядит следующим образом:

Листинг 3.5. Enum приоритетов операций

|  |
| --- |
| const PREC = {    PAREN\_DECLARATOR: -10,    ASSIGNMENT: -2,    CONDITIONAL: -1,    DEFAULT: 0,    LOGICAL\_OR: 1,    LOGICAL\_AND: 2,    INCLUSIVE\_OR: 3,    EXCLUSIVE\_OR: 4,    BITWISE\_AND: 5,    EQUAL: 6,    RELATIONAL: 7,    OFFSETOF: 8,    SHIFT: 9,    ADD: 10,    MULTIPLY: 11,    CAST: 12,    SIZEOF: 13,    UNARY: 14,    CALL: 15,    FIELD: 16,    SUBSCRIPT: 17,  }; |

Как можно заметить, тут также заданы приоритеты и для многих других бинарных операций. Их использование приведено ниже:

Листинг 3.6. Обработка бинарных операций в tree-sitter

|  |
| --- |
| binary\_expression: $ => {        const table = [          [':=', PREC.ASSIGNMENT],          ['+', PREC.ADD],          ['-', PREC.ADD],          ['\*', PREC.MULTIPLY],          ['/', PREC.MULTIPLY],          ['%', PREC.MULTIPLY],          ['||', PREC.LOGICAL\_OR],          ['&&', PREC.LOGICAL\_AND],          ['|', PREC.INCLUSIVE\_OR],          ['^', PREC.EXCLUSIVE\_OR],          ['&', PREC.BITWISE\_AND],          ['=', PREC.EQUAL],          ['!=', PREC.EQUAL],          ['>', PREC.RELATIONAL],          ['>=', PREC.RELATIONAL],          ['<=', PREC.RELATIONAL],          ['<', PREC.RELATIONAL],          ['<<', PREC.SHIFT],          ['>>', PREC.SHIFT],        ];        return choice(...table.map(([operator, precedence]) => {          return prec.right(precedence, seq(            field('expr', $.expression),            // @ts-ignore            field('binOp', alias(operator, $.bin\_op)),            field('expr', $.expression),          ));        }));      }, |

Данный блок кода также был взят из кода для анализатора для языка СИ и адаптирован под существующую грамматику. Данный код создает правило «любой из» с заданными приоритетами.

Оставшиеся конструкции приведены ниже:

Листинг 3.7. Грамматика для unary и braces

|  |
| --- |
| unary\_expression: $ => prec.left(PREC.UNARY, seq(    field('unOp', alias(choice('!', '~'), $.un\_op)),    field('expr', $.expression),  )),  braces\_expression: $ => seq(    '(',    field('expr', $.expression),    ')'  ), |

Таким образом были реализованы все блоки Expressions, далее будет рассмотрен следующий уровень, а именно Statement, основной «строительный блок» функции.

Ниже приведена синтаксическая модель, данная во втором варианте, которую необходимо реализовать:

|  |
| --- |
| statement: {  |if: 'if' expr 'then' statement ('else' statement)?;  |block: 'begin' statement\* 'end' ';';  |while: 'while' expr 'do' statement;  |do: 'repeat' statement ('while'|'until') expr ';';  |break: 'break' ';';  |expression: expr ';';  }; |

В грамматике tree-sitter он будет иметь следующий вид:

Листинг 3.8. Блок statement в грамматике tree-sitter.

|  |
| --- |
| statement: $ => choice(    field('if', $.if\_statement),    field('block', $.block\_content),    field('while', $.while\_content),    field('do', $.do\_content),    field('break', $.break\_content),    field('expression', $.expression\_content),  ), |

Здесь также удобно пройтись снизу вверх, разобрав каждый блок. Ниже приведены блоки break и expression, как самые простые:

Листинг 3.9. Конструкции break и expression

|  |
| --- |
| break\_content: $ => seq('break', ';'),  expression\_content: $ => field('expr', seq($.expression, ';')), |

Конструкции циклов while и do, а также block не сложные, они приведены ниже:

Листинг 3.10. Конструкции циклов и block

|  |
| --- |
| do\_content: $ => seq(    'repeat',    field('statement', $.statement),    choice('while', 'until'),    field('expr', $.expression),    ';'  ),  while\_content: $ => seq(    'while',    field('expr', $.expression),    'do',    field('statement', $.statement),  ),  block\_content: $ => seq(    'begin',    repeat(field('statement', $.statement)),    'end',    ';'  ), |

Конструкция if не так проста, как кажется из-за блока else, который является необязательным. Здесь необходимо правильно решить проблему «висящего else». В данной работе это было решено по примеру с СИ грамматики:

Листинг 3.11. Конструкция if-else

|  |
| --- |
| if\_statement: $ => prec.right(seq(    'if',    field('expr', $.expression),    'then',    field('statement', $.statement),    optional(seq('else', field('statement', $.statement))),  )), |

Таким образом был полностью обработан блок statement, для завершения грамматики осталось лишь разработать объявление функций:

|  |
| --- |
| source: sourceItem\*;  typeRef: {  |builtin: 'bool'|'byte'|'int'|'uint'|'long'|'ulong'|'char'|'string';  |custom: identifier;  |array: 'array' '[' (',')\* ']' 'of' typeRef;  };  funcSignature: identifier '(' list<argDef> ')' (':' typeRef)? {  argDef: identifier (':' typeRef)?;  };  sourceItem: {  |funcDef: 'method' funcSignature (body|';') {  body: ('var' (list<identifier> (':' typeRef)? ';')\*)? statement.block;  }; |

Здесь нет никаких особенностей, все пишется ровно так, как и написано, лишь с особенностями перевода на tree-sitter, код приведен ниже:

Листинг 3.12. Конструкции объявления функций и параметров

|  |
| --- |
| source: $ => repeat(    field('sourceItem', $.sourceItem),  ),  sourceItem: $ => field('funcDef', $.funcDef),  funcDef: $ => seq(    'method',    field('funcSignature', $.funcSignature),    choice(      field('body', $.body),      ';',    )  ),  body: $ => seq(    optional(seq(      'var',      repeat(seq(        optional(field('list\_identifier', $.list\_identifier)),        optional(seq(          ':',          field('typeRef', $.typeRef)        )),        ';'      ))    )),    field('statment\_block', $.block\_content)  ),  list\_identifier: $ => seq(    field('identifier', $.identifier),    repeat(seq(',', field('identifier', $.identifier)))  ),  funcSignature: $ => seq(    field('identifier', $.identifier),    '(',    optional(field('list\_argDef', $.list\_argDef)),    ')',    optional(seq(      ':',      field('typeRef', $.typeRef),    ))  ),  list\_argDef: $ => seq(    field('argDef', $.argDef),    repeat(seq(',', field('argDef', $.argDef)))  ),  argDef: $ => seq(    field('identifier', $.identifier),    optional(seq(      ':',      field('typeRef', $.typeRef)    ))  ),  typeRef: $ => choice(    alias(choice(      'bool',      'byte',      'int',      'uint',      'long',      'ulong',      'char',      'string',    ), 'builtin'),    field('custom', $.identifier),    field('array', $.array)  ),  array: $ => seq(    'array',    '[',    repeat(','),    ']',    'of',    $.typeRef,  ), |

Приведенная грамматика генерирует практически точное дерево для задания, за несколькими отличиями:

1. В названиях полей запрещены точки и скобки, из-за этого несколько названий требуется переименовать.
2. В tree-sitter нельзя сделать полностью пустой узел, который бы отобразился в дереве, из-за этого возникает проблема c отображением пустого списка.

Обе эти проблемы необходимо решать через тестовую программу, которая будет принимать грамматику и выводить дерево разбора, вместе с ошибками, возникшими в ходе разбора переданного файла.

## Разработка тестовой программы

Tree-sitter был выбран по причине того, что результатом его разбора является дерево, что ускоряет процесс построения итогового результата. Тестовую программу было решено делать на python из-за его удобства при написании скриптов.

Первый этап – получение аргументов командной строки, для этого используется следующая функция:

Листинг 3.13. Функция для получения данных из командной строки

|  |
| --- |
| def parse\_cli():      p = argparse.ArgumentParser(description="Сборка и запуск tree-sitter парсера")      p.add\_argument(          "--lib",          dest="lib\_path",          help="Путь к уже скомпилированной tree-sitter библиотеке (.so/.dll/.dylib). "               "В этом случае grammar\_dir и lang\_name не нужны.",      )      p.add\_argument("grammar\_dir", nargs="?", help="Путь к папке с грамматикой (library)")      p.add\_argument("lang\_name", nargs="?", help="Имя парсера, напр. 'foo' для tree\_sitter\_foo")      p.add\_argument("file\_path", help="Путь к файлу для парсинга")      p.add\_argument("out\_file\_path", help="Путь к выходному файлу")      a = p.parse\_args()      if a.lib\_path:          grammar\_dir = None          stem = Path(a.lib\_path).stem  # e.g. libfoo -> libfoo, foo -> foo          lang\_name = stem[3:] if stem.startswith("lib") else stem          lib\_path = a.lib\_path      else:          if not a.grammar\_dir or not a.lang\_name:              p.error("Нужно указать либо --lib <путь\_к\_библиотеке>, либо grammar\_dir и lang\_name.")          grammar\_dir = a.grammar\_dir          lang\_name = a.lang\_name          lib\_path = None      return grammar\_dir, lang\_name, a.file\_path, a.out\_file\_path, lib\_path |

Функция, приведенная выше может принимать грамматику в двух формах:

1. В виде пути к папке с грамматикой и её названием;
2. В виде непосредственно пути к файлу грамматики.

Такой подход позволяет упростить отладку (не требуется перекомпиляция т.к. происходит в автоматическом режиме), но и соответствовать заданию (тестовый скрипт должен принимать на вход файл грамматики).

В случае, если библиотеку не передали, требуется её скомпилировать, делается это следующей функцией:

Листинг 3.14. Функция для компиляции лексического анализатора

|  |
| --- |
| def build\_parser(grammar\_dir: str, lang\_name: str) -> str:      """Собрать shared-библиотеку в <grammar\_dir>/build/<lang\_name>.(dll|so|dylib)."""      out\_dir = os.path.join(grammar\_dir, "build")      os.makedirs(out\_dir, exist\_ok=True)      out\_path = os.path.join(out\_dir, f"{lang\_name}.dll")      subprocess.run(          ["tree-sitter", "generate", "--abi", str(tree\_sitter.LANGUAGE\_VERSION)],          cwd=grammar\_dir,          check=True,      )      subprocess.run(          ["tree-sitter", "build", "-o", out\_path], cwd=grammar\_dir, check=True      )      return out\_path |

Следующим этапом является открытие скомпилированной библиотеки и получение корня дерева для её обхода:

Листинг 3.15. Функция для получения корня дерева

|  |
| --- |
| def load\_and\_parse(lib\_path: str, lang\_name: str, file\_path: str):      """Загрузить язык из DLL, распарсить файл, вернуть корневой узел."""      cdll = CDLL(os.path.abspath(lib\_path))      func\_name = f"tree\_sitter\_{lang\_name}"      if not hasattr(cdll, func\_name) and hasattr(cdll, func\_name + "\_language"):          func\_name = func\_name + "\_language"      func = getattr(cdll, func\_name)      func.restype = c\_void\_p      ptr = func()      PyCapsule\_New = PYFUNCTYPE(py\_object, c\_void\_p, c\_char\_p, c\_void\_p)(          ("PyCapsule\_New", pythonapi)      )      capsule = PyCapsule\_New(ptr, b"tree\_sitter.Language", None)      lang = Language(capsule)      parser = Parser(lang)      with open(file\_path, "rb") as f:          data = f.read()      tree = parser.parse(data, encoding="utf8")      return tree.root\_node |

Последняя функция разбирает дерево и выводит его в удобочитаемой форме.

|  |
| --- |
| def write\_tree\_to\_file(node, out\_path: str | Path, \*, ascii: bool = False) -> None:      """      Пишет вывод дерева в файл out\_path.      Ветвление оформлено как в `tree`. Логика печати полностью соответствует исходной.      """      VBAR, SPACE, TEE, ELBOW = ("|   ", "    ", "|-- ", "`-- ") if ascii else ("│   ", "    ", "├── ", "└── ")      replace\_names = {          "list\_argDef": "list<argDef>",          "list\_identifier": "list<identifier>",          "statment\_block": "statment.block",          "listExpr": "list<expr>",      }      def \_prefix(anc\_has\_next: list[bool], is\_last: bool) -> str:          if not anc\_has\_next:              return ""            base = "".join(VBAR if has\_next else SPACE for has\_next in anc\_has\_next)          return base + (ELBOW if is\_last else TEE)      def \_child\_anc(anc\_has\_next: list[bool], is\_last: bool) -> list[bool]:          return [\*anc\_has\_next, not is\_last]      def \_walk(n, indent: int, anc\_has\_next: list[bool], is\_last: bool, out):          fname = field\_name\_of(n)          is\_token = n.type in (              "identifier",              "bin\_op",              "str",              "char",              "hex",              "bits",              "dec",              "un\_op",              "builtin",          )          if fname in replace\_names:              fname = replace\_names[fname]          if is\_token:              out.write(f"{\_prefix(anc\_has\_next, is\_last)}{fname}\n")          if n.is\_missing:              print(f"Error: missing element \"{n.type}\" in end point {n.end\_point}", file=sys.stderr)          if n.is\_error:              print(f"Error: incorrect in end point {n.end\_point}", file=sys.stderr)            if len(n.children) == 0:              text = n.text.decode("utf-8")              if is\_token:                  child\_pref = \_prefix(\_child\_anc(anc\_has\_next, is\_last), True)                  out.write(f'{child\_pref}"{text}"\n')              else:                  out.write(f'{\_prefix(anc\_has\_next, is\_last)}"{text}"\n')          else:              out.write(f"{\_prefix(anc\_has\_next, is\_last)}{fname}\n")          child\_anc = \_child\_anc(anc\_has\_next, is\_last)          inject\_cfg = {              "funcSignature": ("list\_argDef", "("),              "call\_expression": ("listExpr", "("),              "indexer": ("listExpr", "["),          }          need\_field, left\_token = inject\_cfg.get(n.type, (None, None))          present\_fields = {n.field\_name\_for\_child(i) for i in range(n.child\_count)} - {None}          should\_inject = bool(need\_field) and (need\_field not in present\_fields)          injected = False          for i, ch in enumerate(n.children):              \_walk(ch, indent + 1, child\_anc, i == len(n.children) - 1, out)              if (should\_inject and not injected and ch.type == left\_token):                  display = replace\_names.get(need\_field, need\_field)                  out.write(f"{\_prefix(child\_anc, False)}{display} (empty)\n")                  injected = True        out\_path = Path(out\_path)      out\_path.parent.mkdir(parents=True, exist\_ok=True)      with out\_path.open("w", encoding="utf-8") as f:          \_walk(node, 0, [], True, f) |

Самый важный тут момент, это функция *\_walk*, которая непосредственно «рисует» дерево в файл. Грубо говоря здесь приведен обычный обход дерева, который лишь исправляет «косяки» лексического анализатора.

Первым делом он обрабатывает имя очередного ребенка и, при необходимости, его обновляет (таким образом получается добавить различные стрелочки и точки в названия). Потом обрабатываются токены, те элементы, которые изначально не выводятся, чтоб вместо:

был добавлен промежуточный

Последующая обработка – проверяет, не является ли данный элемент ошибкой или пропуском и при необходимости выводит информацию об этом. После этого проверяет, а не лист ли это, чтоб вывести его значение.

Последняя обработка занимается вставкой пустого списка, который опускается при лексическом разборе, если его нет.

В итоге, функция верхнего уровня для разбора выглядит следующим образом:

Листинг 3.16. Функция верхнего уровня для разбора входного файла

|  |
| --- |
| def main():      grammar\_dir, lang\_name, file\_path, out\_file\_path, lib\_path = parse\_cli()      if lib\_path:          root = load\_and\_parse(lib\_path, lang\_name, file\_path)      else:          built\_lib = build\_parser(grammar\_dir, lang\_name)          root = load\_and\_parse(built\_lib, lang\_name, file\_path)      write\_tree\_to\_file(root, out\_file\_path) |

После этого можно переходить непосредственно к тестированию.

## Тестирование разработанной грамматики