コンパイラ学習者のための Yacc 用言語サーバの開発

稲垣研究室 石部 鳳空

あらまし Yacc は大学のコンパイラの講義に採用されているパーサジェネレータである. 現在 Yacc コーディング時に使用できる静的解析に基づく支援ツールとして VSCode 拡張のものが存在するが、他の主要なエディタでは同等のコーディング支援機能は提供されていない. そこで本研究では、LSP に則った Yacc の言語サーバを開発した. 各エディタで Yacc ファイルを開いた時に、作成した言語サーバを実行するように設定したことで、異なるエディタで静的解析に基づくコーディング支援機能が容易に得られることが確認できた.

キーワード LAT_FX, Yacc, Bison, LSP, 言語サーバ, 静的解析

1. 背景

1.1. Yacc とは

Yacc は高専や大学のコンパイラの講義やコンパイラ 開発者、研究者などに広く使用されている一般的な LALR(1) パーサジェネレータである. Yacc ファイルに は終端記号と非終端記号を定義し、それらを参照しながら BNF で構文規則と対応する意味動作を記述する. 作成したファイルを Yacc で処理すると、内部でオートマトン (実体としては構文解析表である) とスタックを生成し、構文解析関数がこれらを操作して構文解析を実行するため、ユーザはこの関数を呼び出すだけで構文解析が実行できる.

コンパイラの授業の演習などで Yacc を採用している事例として、豊田高専専攻科情報科学専攻の「コンパイラ」^[1] や岡山大学工学部情報系学科「コンパイラ」^[2]、電気通信大学情報理工学域「言語処理系論」^[3] などがある。またプリンストン大学のコンパイラの講義の教科書である『最新コンパイラ構成技法』^[4] では構文解析器を開発するために ML-Yacc(Yacc のstandard ML 実装)を使用している。このような使用例から分かるように、Yacc は学生やコンパイラ学習者に記述されていることが分かる。

パーサジェネレータを使用する場面でのLex/Yaccアプローチ以外の主要な選択肢としてANTLRを用いることがある。多くの大学の講義でLex/Yaccアプローチを採用しているという事実があるが、実験の結果ANTLRを用いた方が学生の試験合格率や最終的な成績が高いことが示されている [17]. つまり、この実験ではYaccが他のツールと比較してユーザにとって実装の難易度が高いことが示唆されている。また学生のアンケートでの意見によると、ANTLRはLex/Yaccよりもシンプルで直感的で保守性が高いと評価されている。

1.2. Yacc の周辺ツールの現状

Yacc の周辺ツールについて, 理解の補助のための ツールとして Leon Aaron Kaplan による"yaccviso - a tool for visualizing yacc grammars"^[5] や楠目勝利らによる「コ ンパイラにおける構文解析過程の視覚化」[6] が提案さ れている. これらのツールは Yacc ファイルに記述し た文法を概観したり、Yacc が出力した構文解析器の 動作を可視化したりする際に利用される. また, コー ディング時の補助ツールとして,シンタックスハイラ イトを施すもの (vscode-yacc^[7]) や静的解析を行うもの (Yash^[8]) が存在する. しかし, これらは Visual Studio Code (VSCode) 拡張として提供されているソフトウェ アであり、VimやEmacs などの他の主要なエディタで は同等の支援機能が得られない. 特に静的解析によ るコーディング支援を提供するものは Yash のみであ るが、これは現時点(2024年2月8日時点)で73,000回 以上インストールされており、Yacc コーディング時 の静的解析による支援機能の提供への需要が存在する ことが分かる.



図 1: 言語サーバを利用したコーディング支援

2. 問題と解決法

セクション 1.2 で述べたように、現時点では Yacc コーディング時の静的解析によるコーディング支援機能の提供を行うソフトウェアは VSCode 上でのみ動作し、他の主要なエディタでは同等の支援機能が得られない。そこで本研究では、現在多くの主要なプログラミング言語の実装が存在する Language Server Protocol (LSP)[9]

の仕様に則った言語サーバを Yacc 用に開発する. こ 30 の言語サーバを利用することで、ユーザは使い慣れた31 エディタでコーディング支援機能を使うことができ, 33 Yacc コーディング時の負担が軽減する. また, 新しい 34 エディタが現われた場合でも,エディタ開発者はLSP 35 に則ってクライアントを開発するだけで Yacc コーディ $\frac{36}{37}$ ング時の支援機能をユーザに提供することができる.38 なお、現在 Yacc よりも一般に使用されている Bison 39 $3.8.1^{[16]}$ の仕様を元に静的解析器を実装する. Bison は $\frac{40}{100}$ GNUによる Yacc 実装で、軽微な記述の差異や拡張さ42 れた機能はあるが、基本的なファイル記述は Yacc と 43 同様であり、Yacc との上位互換性のあるソフトウェ アのため、便宜上Yaccと表現している. Bisonファイ ルの記述例をソースコード1に示す. Bison ファイル は基本的にプロローグ, Bison 宣言, 文法規則, エピ ローグの4つのセクションからなり、プロローグでは 出力ファイルの最初に記述される C コード, Bison 宣 言ではトークンの定義と優先順位や結合規則などの設 定, 文法規則では構文解析器が認識する文法規則, エ ピローグでは出力ファイルの最後に記述されるCコー ドが書かれる.

ソースコード 1: Bison の記述例

```
1
        #include <stdio.h>
2
3
        #include <stdlib.h>
4
5
        extern int yylex();
6
        extern char *yytext;
7
        extern FILE *yyin;
8
        void yyerror(const char *s) {
            fprintf(stderr, "error: %s\n", s);
10
11
12
   %}
13
14
    %token NUMBER
15
    %token PLUS MINUS MULTIPLY DIVIDE
   %token LPAREN RPAREN
16
17
18
   /* definition of grammar rules */
19
20
   expr:
21
        expr PLUS term \{ \$\$ = \$1 + \$3; \}
22
        | expr MINUS term { $$ = $1 - $3; }
                          { $$ = $1; }
23
        | term
24
25
26
    term:
27
        term MULTIPLY factor \{ \$\$ = \$1 * \$3; \}
28
        l term DIVIDE factor
                                 { $$ = $1 / $3; }
29
        | factor
                                 { $$ = $1; }
```

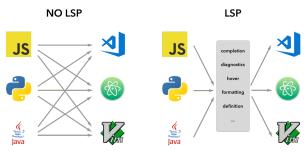


図 2: LSP 登場による言語とエディタの関係の変化 [11]

3. LSPとは

LSP は 2016 年に Microsoft が発表したプロトコルで、エディタと言語サーバ間の通信方法を規定している(図 2). 現在は多くの主要なプログラミング言語が公式、非公式を問わず言語サーバを持っており [10]、プログラミング言語のコーディング環境を整備する際に複数の拡張やプラグインをインストールする必要がないだけでなく、言語クライアントが存在するエディタならばどのエディタでも利用することができることが強みである.

近年、Go や TypeScript、Rust といった新しいプログラミング言語が注目を集め利用者が増えているが、エディタやIDE などの開発者がそれぞれのプログラミング言語用に補完や定義ジャンプ、ホバー時のヒント表示などの多くの支援機能を追加するには多大な労力を要する.従来は開発ツール毎に支援機能を開発する必要があり、M個の言語に対してN個のエディタが存在し、その実装の数はM×Nであった.しかしLSPの登場により、M個の言語サーバとN個のクライアントのみで同様の機能が実現できるようになり、実装の数はM+Nとなった.このようなことから、LSPは言語のユーザとエディタ開発者双方の負担を軽減することができると言える.また、LSPを用いることで言語の静的解析器を一度実装すれば良いため、新しい言語を

異なるエディタに素早く取り込めたり、特定のエディタの拡張やプラグインを大量に記述する必要がないため、言語の普及を早めたりする効果が期待できる. さらに、特定のエディタでのみ記述ができる言語は潜在的なユーザグループを排除しているという指摘^[12]もあるため、LSPを用いて多様なエディタに対応することで、Yacc ユーザやエディタ開発者の負担軽減だけでなく、言語の普及に貢献することもできると考える. なお、本研究ではLSPのバージョン 3.17 に準拠する.

4. Yacc 用言語サーバの開発

4.1. 開発概要

本研究では Yacc 用言語サーバを開発と複数の言語 クライアントの設定を行う. また言語サーバについ て、LSPに則ってリクエストやレスポンスを処理した り、ドキュメントを読み書きしたりするサーバとして の機能だけでなく、静的解析に基づく支援機能をを提 供するために静的解析器も実装する. つまり、開発は 主に3つのプログラムからなり、それは言語サーバと 言語クライアント、静的解析器である.

表 1 に開発環境を示す。主要なエディタはすでに言語クライアントが実装済みであり、開発者は設定ファイルへの追記のみで対応できることが多いが、VSCode は言語クライアントのパッケージを用いて軽微な実装を行う必要があった。表 1 中の言語クライアントは VSCode の言語クライアントである。なお、今回開発した Yacc 言語サーバのソースコードは https://github.com/is-hoku/yacc-language-server で参照できる。

	言語サーバ	言語クライアント			
プログラミング言語	OCaml 4.14.0 TypeScript 5.3.3				
ビルドシステム	dune 3.7.2	tsc 5.3.3			
ランタイム	OCaml 4.14.0 Node.js v21.6.1				
OS	Arch Linux x86_64				
CPU	Intel i7-10710U				

表 1: 開発環境

4.2. 言語クライアントの実装・設定

VSCode で動作する言語クライアントを VSCode 拡張として実装した. そこでは Microsoft が開発している vscode-languageclient という npm パッケージを使用し、 VSCode の拡張機能としての設定や言語サーバとの接続方法などの設定を記述した. また、Vim と Emacs でも動作されるために、Yacc ファイルを開いた状態で

言語サーバと通信を行うように設定ファイルに記述を 追加した.

4.3. 言語サーバの通信

LSPではJSON-RPCを用いて通信をするが,通信方式は指定されていない.しかしLSPにおいては,同一のコンピュータでクライアントとサーバが動作することが通常であるため,今回は同一コンピュータ上で標準入出力による通信を確立することとした.実際にVSCodeの言語クライアントのパッケージでは標準入出力だけでなく,パイプ通信やソケット通信,IPC通信が選択できるが,一方でVimやEmacsの言語クライアントのプラグインの多くが標準入出力での通信のみをサポートしているという事実があり,これは標準入出力による通信がよく用いられることの証左である.

LSP の仕様に則って、ドキュメント (ファイル) はクライアントではなくサーバ側のハッシュ表で管理することとした。ドキュメントが開かれると、クライアントはサーバにドキュメントの内容を含むドキュメント情報を送信する。このリクエストを処理する際に、サーバがドキュメント情報をハッシュ表に格納する。これ以降のドキュメントの編集では、差分のみがサーバに送信されるため、差分をハッシュ表に格納されている情報に適用することで、サーバはクライアントが開いているドキュメントと同期するように実装した。

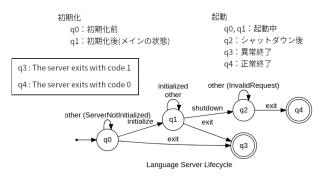


図 3: 言語サーバのライフサイクル

4.4. 言語サーバのライフサイクル

言語サーバのライフサイクルを図3に示す. 初期化と起動という2つの状態を保持し、これらの状態に応じてリクエストを受理してレスポンスを返却するか、状態を遷移させてエラーレスポンスを返却するかなどを判断する.

LSP の通信でやりとりされるメッセージには、request, response, notification の 3 種類がある. request は response を必要とし、notification は response を必要としないものである. クライアントとサーバのどちらからも、ど

のメッセージでも送受信できる. クライアントは支援機能に関する通信を行う前に, initialize request を送信する. これに対してサーバは response を返し, クライアントは initialized notification を送信する. この一連の通信でクライアントとサーバはそれぞれの capabilities を送り合う. これによって, 双方はそれぞれが対応している支援機能などについて知ることができる. これが完了すると, サーバはクライアントから送信された補完や定義ジャンプなどの支援機能に関わる request を受理するようになる.

今回は textDocument/didSave notification がクライアントから受信した時、サーバで静的解析を行い、その結果として得られた構文エラーなどの情報をtextDocument/publishDiagnostics notification を用いてクライアントに送信するように実装した. 現在、実装が完了し対応している LSP メソッドとその役割について表 2 に示す.

4.5. Yacc の静的解析器

Bison 実装の字句解析器と構文解析器の部分を ocamllex, Menhir を用いて OCaml への移植を行った. これによって Bison の仕様に則った字句解析器と構文解析器を得られた. また, 構文解析の結果の出力として抽象構文木を得るように実装をすることで, 静的解析が可能となった.

完全な入力を想定している通常のコンパイラと異な り、言語サーバは多くの場面で不完全なソースコード が入力として渡される. 抽象構文木を走査して型検 査や変数の定義と使用の検査などを行うため, どのよ うな入力であっても構文木を得ることができなければ 十分なコーディング支援を提供することが困難であ る. そのため, 不完全な入力に対してエラー回復を行 い抽象構文木を得る必要がある. エラー回復戦略と しては、特殊な error 記号を使用した局所的なエラー 回復や、Burke-Fisher エラー修復などが考えられるが、 前者はエラー回復後の解析の精度に問題があり、後者 は実装が複雑になることや, 計算量が大きいこと, 副 作用のある意味動作においてエラー修復動作が予期し ない影響を及ぼすという問題がある. まず error トー クンを使用した局所的なエラー回復について考える. この方法では、終端記号とみなされる error トークン を含んだ生成規則を定義するだけで実現でき、実装に かかる労力は小さい. しかしこの方法でのLR 構文解 析器の動作が、error トークンをシフトできる状態に 達するまでスタックをポップし, その後 error トーク ンをシフトしてから, エラーでない動作を持つ先読み

トークンまで必要ならば入力トークンを破棄して構文 解析を再開するというものであり, エラー回復後の解 析の精度に問題がある. 例えば, 適切に error トーク ンを含んだ生成規則を定義できなければ, エラー回復 が別のエラーを引き起こしたり、 スタックのポップに よって意味動作が実行不能となったりすることが考 えられる. 次に Burke-Fisher エラー修復について考え る. この手法はエラーに遭遇した時点の K トークン 前から, それ以降に現れる全ての時点において, 可能 な1トークンの挿入もしくは除去, 置換を全て試して 修復をするものであり, error トークンを用いた手法 のように生成規則を変更したり、 構文解析表が修正さ れたりすることが無い. しかし, これを実現するため にKトークン前のスタックの状態を保持する旧スタッ クが必要となり、N種類のトークンを持つ言語での最 悪計算量は K+K·N+K·N である. このようなこと から、Burke-Fisher エラー修復は実装の労力と計算量 が大きい点で不利であり, エラー修復中のシフトや還 元動作で意味動作が何度も実行されることから, 副作 用を持つ意味動作を記述した場合に問題がある.

Merlinの報告書^[15]によると、エラー回復の厳密さは解析対象の言語機能に依存するとされ、スコープ構造と前方参照を許す言語であるかどうかを考慮すべは持たないが、生成規則の記述中で前方参照を許したりなるため、厳密さを考えずにスタックをポップしたり入力トークンを破棄したりすると、意味解析段階でエラー回復手法では問題があり、より素朴なエラー回復が表り正常に解析を続行できる状態になるまでで、そことが求められた。そことが求められた。そことがでは Menhir の Inspection API を利用することとや、た、Inspection API は LR 構文解析器のオートマトンやスタックの状態にアクセスすることができるため、比較的汎用なエラー回復が実現できる.

静的解析の流れを図4に示す. 図中では今回実装した部分を青色の文字で表している.

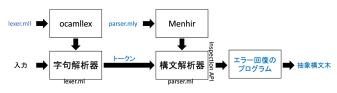


図 4: 静的解析の流れ

エラー回復戦略を説明する. エラーとなるトークン

表 2: 実装済みの LSP メソッドとその役割

種類	メソッド名	役割
	initialize	初期化.クライアントの capabilities をサーバで保持し, サーバの capabilities を生成して返す.
request	shutdown	サーバの状態を遷移させ、これ以降 exit notification 以外の要求を拒否するようにする.
	initialized	クライアントが初期化が完了したことの通知.
	textDocument/didOpen	ドキュメントが開かれたことの通知. ドキュメント情報をハッシュ表で保持する.
notification	textDocument/didChange	ドキュメントが変更されたことの通知. ドキュメントの変更をハッシュ表に反映させる.
	textDocument/didSave	ドキュメントの変更が保存されたことの通知.これをトリガーとして静的解析を行う.
	textDocument/publishDiagnostics	静的解析によるコード診断結果の通知. サーバからクライアントに送信する.
	exit	サーバプロセスを終了させる.

に遭遇したら、その時点のスタックに還元できる生成規則があれば還元を行い、そうでなければ FIRST 集合を計算してシフトできるトークンを取得し、ダミーの値を入れてシフトする。もしシフトできるトークンが一つも無ければ、スタックからトークンをポップし還元できる生成規則を探すところから再開する。

ここで文法を図5として、この戦略でのエラー回復の動作例を図6に示す.入力として%token NUMBER %が与えられた時,%token NUMBER までは通常の構文解析過程でスタックにシフトされるが,次の入力である%は文法のどこにも出現しない文字であり入力が不正だと分かる.ここで構文解析器は解析を実行することができずエラー状態を報告し,これをエラー回復機構が受け取りエラー回復を試みる.まず還元できる生成規則を探すが,スタックにはpercent_symbolidが積まれているため,idをtoken_declに還元できることが分かる.一度還元を実行すると,通常の構文解析過程に戻る.そこでは,もう不正な入力は無いため,生成規則に従って還元を順番に実行していき,最終的にsymbol_declarationを得ることができる.

図 5: 文法 (Bison 文法の一部抜粋)

	スタック	入	、カ	動作	処理
1		%token	NUMBER %	シフト	構文解析器
2	percent_symbol		NUMBER %	シフト	構文解析器
3	percent_symbol	id	%	エラー	エラー回復機構
4	percent_symbol	token_decl		還元	エラー回復機構
5	percent_symbol token_decl_1		還元	構文解析器	
6	percent_symbol token_decls		還元	構文解析器	
7	symbol_declara	tion		還元	構文解析器

図 6: エラー回復の動作例

2で示したように Yacc はセクション毎に記述が分かれており、セクションによって使用する文字や文字列が異なる. このエラー回復戦略はヒューリスティックな解決法ではあるが、このようなセクション構造のために字句解析段階で一部の構文解析も行っているという事情により、良い結果が得られたと推察する. この戦略では不正な入力文字を破棄してしまう点や、エラーの発見が先延ばしになるような状況で最善の回復ができない点などで問題はあるものの、ユーザの入力途中の記述が渡されることが多いということから、他の方法に比べて有用であると考える.

5. 実行例

コード診断を VSCode, Vim, Emacs で行っている様子を それぞれ図 7, 図 8, 図 9 に示す. ここでは%という入力が 不正であることと,予期しない位置に EOF があることを 構文解析の過程で検出し,LSP の textDocument/didSave メソッドによってエディタに報告している.

図 7: VSCode でのコード診断実行例

6. 考察

VSCode と Vim, Emacs の言語クライアントを使用し、Yacc ファイルが開かれた時に今回作成した言語サーバを実行するように設定したことで、異なるエディタで同様の支援機能が容易に得られることが確認できた。また、Bison の仕様や実装を元に字句解析器と構文解析器を作成し、構文解析器の出力として抽象構文木を構築することで、静的解析を可能にした。今回は

```
#include <stdio.h>
           #include <stdlib.h>
           extern file *yyin;
           void yyerror(const char *s) {↓
fprintf(stderr, "エラー: %s\n", s);↓
      %token NUMBER.
NORMAL & main src/static_analyzer/test/testcase/test2.y
```

図 8: Vim でのコード診断実行例

図 9: Emacs でのコード診断実行例

不完全な入力に対して可能な限りエラー回復を試みて 抽象構文木を構築し、その過程で見つかった構文エ ラーなどを LSP の textDocument/diagnostic メソッドに よってクライアントに送信することで、 開発した言語 サーバが機能することを示した.

7. 今後の課題

今回記述したパーサジェネレータのファイルは Bison のものの移植のため, error 記号によるエラー回復を 前提としており、静的解析を行うことは想定されてい ない. そのため, 文法に対応する意味動作に副作用が ある場合が多いが, エラー回復のパターンを増やした り, より安全にエラー回復を行ったりするためには, 意味動作を副作用の無いものに変更する必要がある. また, 意味解析器を実装し, コード補完や定義ジャン プなどサポートする機能を増やして利便性を高めてい く. その後, 支援機能に関する計算を並列化するなど して効率化を図りたい.

このように問題を解決したり機能を追加したりする ことで, 言語サーバを十分実用に耐え得るものにして いき, ユーザの獲得や継続的な開発を実現したい.

謝辞 8.

本研究の遂行にあたり、指導教員として多大なご指 導を賜りました稲垣宏教授(豊田工業高等専門学校情 報工学科),ご多忙のなか相談に応じてくださりご助 言を頂いた内山慎太郎氏(豊橋技術科学大学大学院工 学研究科 情報・知能工学専攻 博士後期課程 応用数理 ネットワーク研究室)に感謝申し上げます.

- **文献** 高専 Web シラバス, "コンパイラ" [1] https://syllabus.kosen-k.go.jp/Pages/PublicSyllabus? school_id=23&department_id=25&subject_code= 95018&year=2017&lang=ja, 2023 年 10 月 18 日閲覧 Nobuya WATANABE, "コンパイラ(Con
- [2] Nobuya WATANABE, pilers)", http://www.arc.cs.okayama-u.ac.jp/~nobuya/ lecture/compiler/, 2023 年 10 月 18 日閲覧
- [3] 電気通信大学 シラバス Web 公開システム, "シラバス 参照", http://kyoumu.office.uec.ac.jp/syllabus/2023/31/ 31_21124118.html, 2023年10月18日閲覧
- [4] Andrew W. Appel, Modern compiler implementation in ML, Cambridge University Press, New York, Cambridge, 2008. (アンドリュー・W・エイペル 神林靖・ 滝本宗宏 (訳), 最新コンパイラ構成技法, 翔泳社, 東京, 2020.)
- Francisco Ortin, Jose Quiroga, Oscar Rodriguez-Prieto, Miguel Garcia, "An empirical evaluation of Lex/Yacc and ANTLR parser generation tools", 2022.
- [6] Leon Aaron Kaplan, "yaccviso a tool for visualizing yacc grammars", 2006.
- [7] 楠目 勝利, 佐々 政孝. "コンパイラにおける構文解析 過程の視覚化",全国大会講演論文集,第55回,ソフト ウェア科学・工学, pp448-449, 1997.
- "VSCode-YACC" Visual Studio Marketplace, https://marketplace.visualstudio.com/items?itemName= carlubian.yacc, 2023 年 10 月 19 日閲覧
- Visual Studio Marketplace, "Yash" //marketplace.visualstudio.com/items?itemName= daohong-emilio.yash, 2024年2月8日閲覧
- [10] Official page for Language Server Protocol, "Language Server Protocol Specification - 3.17" https://microsoft.github.io/language-server-protocol/ specifications/lsp/3.17/specification/, 2023 年 10 月 19 日閲覧
- [11] Bison 3.8.1, https://www.gnu.org/software/bison/
- manual/bison.html Official page f Γ121 page for Language Server Protocol, "Implementations" https://microsoft.github.io/ language-server-protocol/implementors/servers/, 2023 年10月19日閲覧
- Language Server Extension Guide | Visual Studio Code Extension API, https://code.visualstudio.com/api/ language-extensions/language-server-extension-guide, 2024年2月8日閲覧
- [14] Bünder, H. "Decoupling Language and Editor The Impact of the Language Server Protocol on Textual Domain-Specific Languages." International Conference on Model-Driven Engineering and Software Development, no.10.5220/0007556301290140, pp.129-140, Prague, Czech Republic, 2019.
- [15] npm, "vscode-languageclient", https://www.npmjs.com/ package/vscode-languageclient, 2023年10月19日閲覧
- [16] Menhir Reference Manual (version 20231231), "Inspection API", https://gallium.inria.fr/~fpottier/menhir/ manual.html#sec64, 2024年1月28日閲覧
- Frédéric Bour, Thomas Refis, and Gabriel Scherer. 2018. Merlin: a language server for OCaml (experience report). Proc. ACM Program. Lang. 2, ICFP, Article 103 (September 2018), 15 pages.