特別研究報告

題目

抽象構文木に基づく静的解析ツールに向けた
多言語拡張可能なフレームワークの提案

指導教員
肥後 芳樹 教授

報告者
石村 涼介

令和 8 年 2 月 9 日

大阪大学 基礎工学部 情報科学科

令和 7 年度 特別研究報告

抽象構文木に基づく静的解析ツールに向けた多言語拡張可能なフレームワークの提案

石村 涼介

内容梗概

抽象構文木（AST: Abstract Syntax Tree）は，プログラムのソースコードを構文規則に従って木構造として表現したデータ構造であり，文や式の階層的な構造を表現する．静的解析ツールの多くは，この AST を用いて解析処理を行う．しかし，AST は各プログラミング言語の文法に基づいて構築されるため，その構造は言語ごとに大きく異なる．

そのため，AST ベースの静的解析ツールの多くは単一のプログラミング言語を対象としており，多言語への対応には大きな実装コストがかかる．これは，解析処理の大半が対象言語に依存して実装されており，言語間で再利用可能な部分が限定的であるためである．

一方，コンパイラの分野では，LLVM のように，言語非依存な中間表現を用いることで多言語に対応している．しかし，コンパイラ基盤は主に効率的なコード生成や最適化を目的として設計されており，静的解析とは目的や要求される情報の粒度が異なる．特に，静的解析に必要となる構文構造や意味情報が中間表現への変換過程で抽象化されるため，そのまま静的解析に利用することは容易ではない．

本研究では，静的解析処理を言語依存部分と言語非依存部分に分離することで，多言語に拡張可能な静的解析の基盤となるフレームワークを提案する．本研究の目的は，既存の AST ベースの静的解析ツールを，低コストで多言語対応可能にする基盤を構築することである．具体的には，言語非依存なデータ構造として拡張 CST(E-CST: Extended Code Structure Tree) を導入し，フレームワークを設計した．E-CST は，言語固有の構文構造を抽象化しつつ，静的解析に必要な情報を保持する木構造である．

実際に，既存の静的解析ツールである REPFINDER と PYREF に対してフレームワークを適用することで，単一言語を対象とする静的解析ツールを低コストで多言語に拡張できる可能性を示した．

主な用語

静的解析
抽象構文木 (AST: Abstract Syntax Tree)

中間表現

目 次

1 まえがき **5**

2 背景 **7**

2.1 静的解析 7
2.2 抽象構文木 (AST) 7
2.3 RefDiff 8
2.4 CST 8

3 提案フレームワーク **11**

3.1 E-CST 11
3.1.1 保持する情報 11
3.1.2 提供する操作 12
3.1.3 E-CST の例 13
3.2 処理の流れ 13

4 提案フレームワークの適用 **15**

4.1 適用方法 15
4.2 REPFINDER への適用 15
4.2.1 REPFINDER の概要 15
4.2.2 REPFINDER への適用例 16
4.3 PYREF への適用 17
4.3.1 PYREF の概要 17
4.3.2 PYREF への適用例 17

5 関連研究 **19**

5.1 LiSA 19
5.2 GAST 19
5.3 UNICOEN 20

6 考察 **22**

7 今後の課題 **23**

7.1 E-CST パーサの整備 23
7.2 他の言語非依存な中間表現との連携 23

8　あとがき **24**

謝辞 **25**

参考文献 **26**

1　まえがき

静的解析とは，プログラムを実行せずにソースコードを解析する手法である．リンタ (Linter) はその代表例であり，主に構文規則やコーディング規約に基づく検査を行うことで，コード品質の向上や，コードレビューのコスト削減に寄与している．例えば，Python を対象とした flake8[1] や mypy[2]などのリンタが開発現場において広く利用されている．

一方，近年の研究では，単純な構文規則やコーディング規約の検査にとどまらず，コード要素間の関係やリビジョン間の変化といった，より高度な構造情報を用いる静的解析ツールが提案されている．例えば，PYREF[3] は，Python プロジェクトにおいて，リビジョン間のリファクタリングを検出する静的解析ツールである．REPFINDER[6] は，Java プロジェクトにおいて，ライブラリ更新時に欠落した API の代替となる API を探索する静的解析ツールである．

静的解析では，解析の目的に応じて用いられるデータ構造が異なる．例えば，抽象構文木 (AST: Abstract Syntax Tree) は，プログラムの構文構造を木構造として表現したデータ構造である．AST は構文規則やコーディング規約の検査，型解析など，多くの静的解析で用いられる．一方，制御フローグラフ (CFG: Control Flow Graph) は，文や基本ブロック間の実行順序をグラフとして表現したデータ構造である．CFG はデータフロー解析や抽象解釈など，制御構造を考慮した解析に用いられる．

このように，多種多様な静的解析ツールが提案・利用されているが，多くの静的解析ツールは単一のプログラミング言語を対象としており，多言語への拡張には大きなコストを要するという課題がある．これは，解析処理の大半が対象言語に依存して実装されており，言語間で再利用可能な部分が限定的であるためである．したがって，新たに別のプログラミング言語に対応する際には，解析器の新規実装が必要となり，大きなコストを要する．

一方，コンパイラの分野では，多くのプログラミング言語に対応可能なコンパイラ基盤として LLVM[7] が知られている．LLVM では，コンパイル対象となるソースコードを，プログラミング言語に依存しない中間表現である LLVM IR へ変換する．その後の最適化や機械語生成などの処理は LLVM IR 上で行われるため，言語間で共通の処理を再利用できる．また，新たなプログラミング言語に対応する際には，当該言語から LLVM IR への変換部分のみを新規実装すればよいため，低コストである．

LLVM のようなコンパイラ基盤と同様に，静的解析においても基盤を構築することで，多言語への拡張を低コストで実現できると考えられる．LiSA[4] は，CFG に基づく静的解析基盤であり，現時点では Java，Python，Go を対象言語としている．LiSA では，解析対象と

[1] https://flake8.pycqa.org
[2] https://mypy-lang.org

なるソースコードを，プログラミング言語に依存しない中間表現である LiSA CFG へ変換する．以降の解析処理は LiSA CFG 上で行われるため，LLVM と同様に，言語間で共通の処理を再利用できる．

このように，CFG に基づく静的解析ツールを対象とした基盤は存在するが，AST に基づく静的解析ツールを対象とした基盤は発展途上である．そこで本研究では，AST に基づく静的解析基盤として，拡張 CST(E-CST: Extended Code Structure Tree) を提案する．E-CST は，RefDiff 2.0[10] で提案された CST(Code Structure Tree) を拡張した言語非依存なデータ構造であり，RefDiff 2.0 以外の静的解析ツールにも適用可能である．

実際に，既存の静的解析ツールである REPFINDER と PYREF に対して E-CST を適用することで，単一言語を対象とする静的解析ツールを低コストで多言語拡張できる可能性を示す．

本論文の構成は以下の通りである．第 2 章では研究の背景について詳細に説明する．第 3 章では提案フレームワークについて述べる．第 4 章では提案フレームワークの適用について述べる．第 5 章では関連研究について述べる．第 6 章では考察を述べる．第 7 章では今後の課題を述べる．

2　背景

本章では，研究の背景として，静的解析，抽象構文木，RefDiff，CST について述べる．

2.1　静的解析

静的解析とは，プログラムを実行せずにソースコードを解析する手法である．リンタ (Linter) はその代表例であり，主に構文規則やコーディング規約に基づく検査を行うことで，コード品質の向上や，コードレビューのコスト削減に寄与している．例えば，Error Prone[9] は Google が開発した Java 言語向けのリンタであり，コンパイル時に AST を解析することで，実行時にエラーを引き起こす可能性のあるバグパターンを検出する．FindBugs[5] は Java 言語を対象とし，バグパターンとの照合を通じてプログラム内の潜在的な欠陥を検出するリンタである．

静的解析を実現するためには，解析対象となるソースコードを何らかのデータ構造に変換する必要がある．代表的なデータ構造として，AST や CFG が挙げられる．AST は，プログラムの構文構造を木構造で表現したもの [1] であり，構文規則の検査やリファクタリング検出などに広く用いられている．一方，CFG は基本ブロックをノードとし，制御の流れをエッジとする有向グラフ [2] であり，データフロー解析や抽象解釈などに用いられる．

多くの静的解析ツールは，特定のプログラミング言語を対象として実装されている．これは，データ構造の構築や解析処理が対象言語の構文や仕様に強く依存しており，複数言語を同一の枠組みで扱うことが容易ではないためである．その結果，既存の静的解析ツールを多言語に拡張する場合，大きなコストが発生するという課題が指摘されている [13]．

このような背景から，言語に依存しない共通のデータ構造を用いることで，静的解析ツールの多言語対応を低コストで実現する試み [4][8][13] が注目されている．

2.2　抽象構文木 (AST)

AST は，プログラムの構文構造を木構造で表現したもの [1] であり，構文規則の検査やリファクタリング検出などに広く用いられている．AST は，プログラムの意味に直接関係しない要素を省略し，構文の本質的な構造のみを表現する点に特徴がある．例えば，括弧や区切り記号といった情報は AST 上では省略されることが多い．

AST の構造や仕様はプログラミング言語ごとに異なり，また同一言語であっても，処理系やツールごとに異なる AST が定義される場合がある．そのため，AST の定義に応じて，対応するパーサが実装されることが多い．Java 言語の場合，Eclipse の JDT(Java Development Tools)[3]が提供する AST が広く知られており，専用のパーサも JDT に含まれている．Python

[3]https://www.eclipse.org/jdt/

言語においては，標準ライブラリとして ast モジュール[4]が提供されており，ソースコードから AST を生成することが可能である．

2.3 RefDiff

RefDiff[11] とは，Java プロジェクトにおいて，異なるリビジョン間のソースコードを比較し，リファクタリングを自動的に検出する静的解析ツールである．例えば，メソッド M がクラス X からクラス Y へ移動された場合，RefDiff はその変更を Move Method として検出する．

このようなリファクタリング検出においては，条件文や式の詳細な構文構造よりも，コード要素の包含関係や名前，位置関係といった構造的情報が重要となる．RefDiff では，Java ソースコードから AST を構築し，構築した AST からクラス，メソッド，フィールドなどの情報を抽出することで解析を行う．

当初，RefDiff は Java 言語のみを対象としていたが，RefDiff 2.0 では多言語に拡張され，Java，JavaScript，C の 3 言語に対応可能となった．RefDiff 2.0 では，多言語対応を実現するために，CST と呼ばれる言語非依存なデータ構造が導入された．

2.4 CST

CST とは，RefDiff 2.0 において導入された言語非依存なデータ構造であり，Java，JavaScript，C のいずれかの AST をもとに構築される．CST は AST と同様に木構造でプログラムを表現するが，すべての要素を表現するのではなく，クラスやメソッド，関数といった要素のみに着目する．一方，条件文や式などの詳細な要素は抽象化され，リファクタリング検出に必要な情報に限定した表現となっている．

図 1，図 2 のソースコードから得られる CST の例を図 3 に示す．CST は 1 リビジョンにつき 1 つ生成されるため，図 3 では同一リビジョンに属する 2 つのソースコードから 1 つの CST が生成されている．CST のルートノードは，各ソースコードをトークン単位に区切った状態で保持し，類似度の計算に使用する．また，ルートノード以外のノードは，クラスやメソッドを表現しており，識別子 (例:plus(int, int)) を持つ．クラスを表現するノードには名前空間 (例:com.ex.) が付与され，メソッドを表現するノードには引数名リスト (例:[x, y]) が付与される．このような設計により，クラスやメソッドの移動や名称変更といったリファクタリングを効率よく検出できる．

また，ノード間には，継承関係や呼び出し関係などの関係が明示的に付与されており，クラス階層や依存関係に基づく解析を効率的に実施できる．例えば，図 1 における main メソッ

[4]https://docs.python.org/ja/3/library/ast.html

```java
package com.ex;
public class Main {
    public static void main(String[] args) {
        Calculator c = new Calculator();
        int r = c.plus(2, 5);
        System.out.println(r);
    }
}
```

図 1: Java コードの例 (com/ex/Main.java)

```java
package com.ex;
public class Calculator {
    public int add(int x, int y) {
        return x + y;
    }

    /**
     * @deprecated Use add(int, int) instead.
     */
    @Deprecated
    public int plus(int x, int y) {
        return x + y;
    }
}
```

図 2: Java コードの例 (com/ex/Calculator.java)

ドは図 2 における plus メソッドを呼び出しているため，図 3 では main メソッドを表現するノードから plus メソッドを表現するノードへ有向辺が付与されている．

CST のノードは，各プログラミング言語の AST ノードから生成されるが，その表現は言語非依存となるよう設計されている．CST 上で表現される AST ノードの一覧を表 1 に示す．例えば，Java ではクラスやメソッドが CST のノードとして表現される一方，JavaScript や C においても，それぞれの言語における関数やファイルが共通の概念として CST 上に表現される．すなわち，CST は言語間の差異を吸収する中間表現であるといえる．

なお，リビジョン間で追加・削除・変更されたファイルのみを対象として CST が構築される．未変更のファイルはリファクタリング検出において重要な役割を果たさないためである．

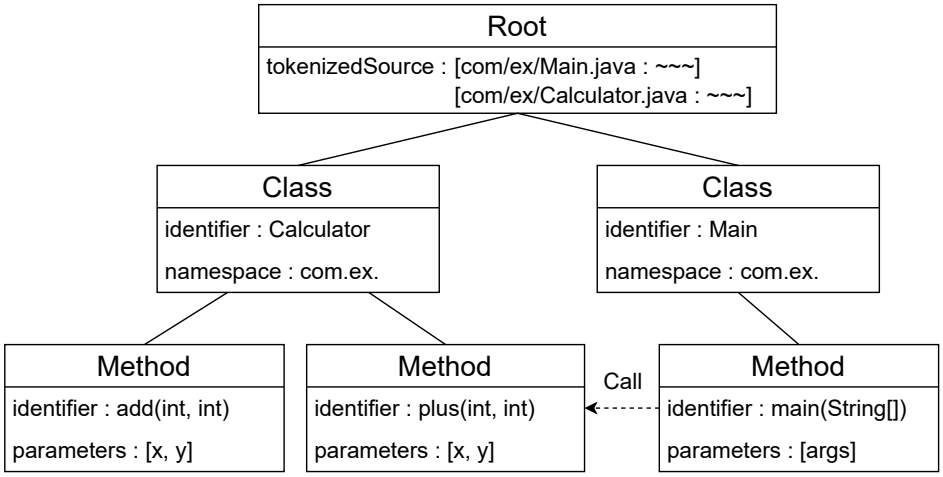

図 3: CST の例

表 1: CST 上で表現される AST ノード [10]

言語	AST ノード
Java	class, enum, interface, and method
C	file and function
JavaScript	file, class, and function

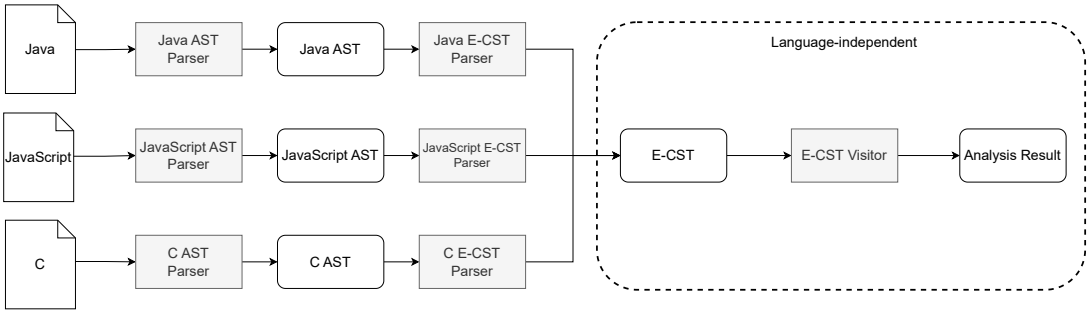

図 4: 提案フレームワーク

3　提案フレームワーク

本フレームワーク (図 4) の目的は，静的解析処理を言語間で共通化することである．そのために，言語非依存なデータ構造として，CST を多様な静的解析でも利用できるように拡張した CST（以降 E-CST）を導入する．

3.1　E-CST

E-CST とは，RefDiff 2.0 において導入された CST を，リファクタリング検出に限らず，より一般的な AST ベースの静的解析にも適用可能とした言語非依存なデータ構造である．CST が持つ言語非依存性を維持しつつ，より多様な静的解析に必要となる情報および操作を追加している．CST は，クラスやメソッド，関数といった要素に着目し，リファクタリング検出に特化した設計となっている．そのため，リファクタリング検出以外の静的解析で重要となる，型情報や API の利用に関する情報などは，十分に表現されていない．E-CST はこの点を補完し，AST ベースの静的解析の基盤として利用可能な表現へ拡張したデータ構造である．

3.1.1　保持する情報

E-CST が保持する情報を表 2 に示す．E-CST では，CST に対して以下の情報を追加した．

- 引数型リスト (parameterTypes)
- 返り値の型 (returnType)
- 非推奨メッセージ (deprecatedMessage)

CST では引数名リスト (parameterNames) のみ保持していたが，E-CST では引数および返り値の型情報を保持することで，メソッド探索や API 間の比較において型情報を用いるこ

表 2: E-CST が保持する情報

情報	例	備考
tokenizedSource	[”private”, ”void”, ...]	ルートノードのみ
identifier	”plus(int, int)”	
namespace	”com.ex.”	クラス，ファイルのみ
parameterNames	[”x”, ”y”]	関数，メソッドのみ
parameterTypes	[”int”, ”int”]	関数，メソッドのみ
returnType	”int”	関数，メソッドのみ
deprecatedMessage	”Use hoge() instead.”	

とができる．また，非推奨メッセージを保持することで，非推奨 API の代替候補探索や API 移行支援などへの応用が期待できる．これらの情報は，従来の CST では直接的に扱われていなかったものであり，本フレームワークを適用可能な静的解析ツールの幅を広げている．

なお，表 2 に示した情報は，文字列型として保持する．identifier，namespace，returnType，deprecatedMessage は文字列型として保持する．parameterNames および parameterTypes は文字列型の順序付きリストとして保持し，引数宣言順に格納する．tokenizedSource はトークンを表す文字列型の順序付きリストとして保持する．

3.1.2 提供する操作

E-CST が提供する操作を表 3 に示す．E-CST では，CST に対して以下の操作を追加した．

- ソースコード全文の復元
- Visitor パターン用の accept() メソッド

まず，ソースコード復元に関する差異について述べる．CST では，単一のノードをソースコードに復元することは可能であったが，ソースコード全文を復元する操作は提供されていなかった．そこで E-CST では，ルートノードを入力とし，各ファイルパスに対応するソースコード文字列の集合を出力する復元操作を定義した．本操作は，ルートノードが保持する tokenizedSource を用いて，各ファイルのソースコードを再構成する．これにより，E-CST 上では保持していない詳細な構文情報を，必要に応じて元のソースコードから参照でき，構造情報とテキスト情報を組み合わせた解析が可能となる．

次に，走査機構の差異について述べる．CST では，ルートノードを起点とする深さ優先探索 (pre-order) を実行できる．しかしこの機構は，単一のコールバック関数を適用する形式の走査であり，ノード型ごとの処理分岐や post-order の深さ優先探索などは定義されていな

表 3: E-CST が提供する操作

操作	備考
全ノードの走査	深さ優先探索 (pre-order のみ)
関係グラフを辿る	呼び出し関係または継承関係
ノードをソースコードに復元	
ソースコード全文の復元	
Visitor を accept する	Visitor パターン用

い．そこで E-CST では，E-CST Visitor を引数として受け取り，ノード到達時 (pre-order) および子ノード走査後 (post-order) に E-CST Visitor の対応メソッドを呼び出す accept 操作を定義した．本操作は値を返さず，解析結果は E-CST Visitor の内部状態として保持される．これにより，E-CST を変更することなく新たな解析処理を追加可能となり，汎用的な AST ベースの静的解析基盤へと拡張されていることが分かる．

3.1.3 E-CST の例

図 1，図 2 のソースコードから得られる E-CST の例を図 5 に示す．ここで，E-CST の木構造自体は，図 3 の CST と同様である．すなわち，E-CST は CST と同様に，1 リビジョンにつき 1 つ生成される．また，CST と同様に，main メソッドを表現するノードから plus メソッドを表現するノードへ，呼び出し関係を表す有向辺が付与されている．

一方，図 5 においてメソッドを表現するノードに注目すると，本フレームワークにおいて新たに追加された，引数型リスト，返り値の型，非推奨メッセージが保持されている．また，クラスを表現するノードに注目すると，本フレームワークにおいて新たに追加された非推奨メッセージが保持されている．

3.2 処理の流れ

本フレームワークは 3 段階の処理に分けられる (図 4)．

1. **AST の構築**

   解析対象となるソースコードを AST に変換する．ここでは既存の AST パーサを用いる．例えば Java 言語の場合は，JDT の AST パーサを用いる．

2. **E-CST の構築**

   AST をもとに E-CST を構築する．例えば Java の AST をもとに E-CST を構築する場合，Java 専用の E-CST パーサを用いる．すなわち，E-CST パーサは解析対象とな

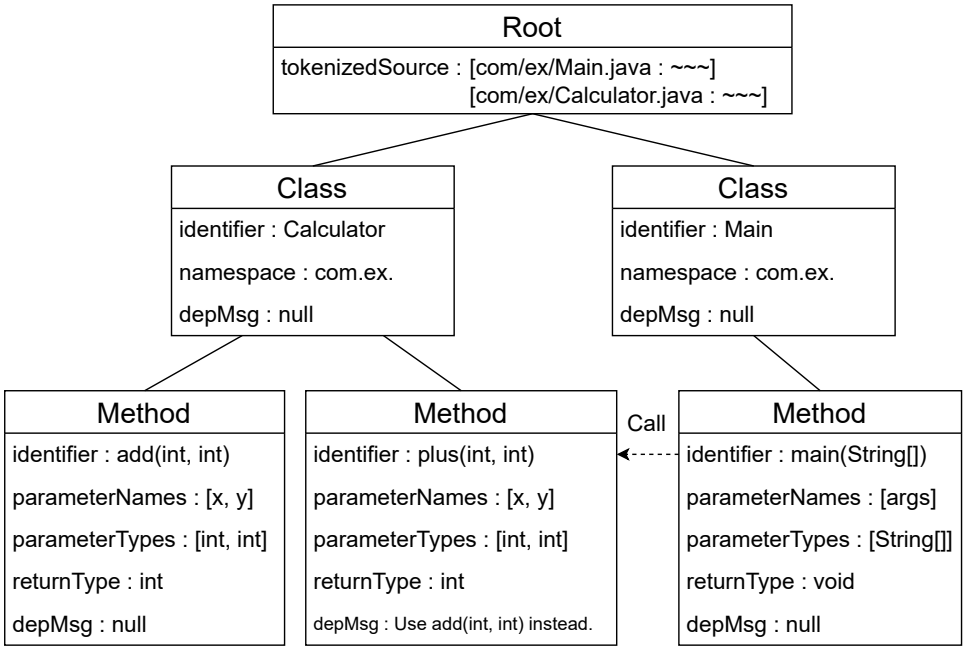

図 5: E-CST の例

る言語ごとに開発する必要がある．

3. 静的解析の実行

E-CST 上で静的解析を実行する．E-CST は言語非依存なデータ構造であるから，解析対象の言語にかかわらず，解析器を再利用できる．解析器としては，Visitor パターンに基づく E-CST Visitor を想定している．これにより，E-CST 側を変更することなく解析器を開発できる．

4　提案フレームワークの適用

本章では，既存の静的解析ツールである REPFINDER と PYREF に対し，提案フレームワークの適用例を示す．

4.1　適用方法

提案フレームワークの適用は，以下の 3 段階の手順により行う．

1. **E-CST への変換**

   解析対象となるソースコードを E-CST へ変換する．対象言語に対応した E-CST パーサが既に存在する場合はそれを利用し，存在しない場合には，当該言語の AST を入力として E-CST を構築する E-CST パーサを新たに実装する．

2. **静的解析の実行**

   構築した E-CST に対して静的解析を実行する．E-CST は Visitor パターンを採用しているため，解析内容に応じた E-CST Visitor を実装することで，E-CST の構造を変更することなく解析処理を追加できる．この解析器は言語非依存であり，同一の Visitor を複数のプログラミング言語に対して再利用可能である．

3. **解析結果の収集・利用**

   解析結果を収集・利用する．E-CST 上で得られた解析結果は，リファクタリング検出や非推奨 API の検出，コード要素間の関係解析など，さまざまな静的解析に応用できる．

4.2　REPFINDER への適用

4.2.1　REPFINDER の概要

REPFINDER は Java プロジェクトにおいて，ライブラリ更新時に欠落した API の代替となる API を探索する静的解析ツールである．ライブラリ更新時，一部の API は削除されたり非推奨 (deprecated) とされたりするため，欠落することがある．その結果，ライブラリ利用者は欠落 API の代替 API を手動で探す必要があり，これは大きなコストのかかるソフトウェア保守作業となる．REPFINDER では代替 API を自動的に発見するために，以下の 3 段階の探索を行う．

1. JavaDoc 非推奨メッセージをもとに探索

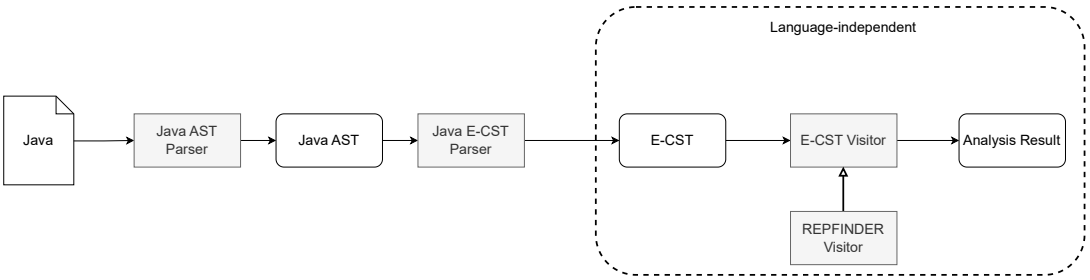

図 6: 提案フレームワークの適用例 (REPFINDER)

2. 自ライブラリを探索

3. 外部ライブラリを探索

3 段階の探索によって，既存手法よりも高い再現率で代替 API を提示できることが報告されている．

REPFINDER は，API 間の構造的関係や継承関係を解析するため，Java の AST に基づいた静的解析を行う．具体的には，クラス階層の探索やメソッドシグネチャの比較を通じて，欠落 API と類似した API を代替 API の候補として抽出する．すなわち，REPFINDER は Java の AST に強く依存している．

その結果，REPFINDER の探索アルゴリズム自体は他のプログラミング言語にも適用可能な性質を持つ一方で，実装上は Java 専用のツールとなっている．

4.2.2 REPFINDER への適用例

REPFINDER へのフレームワークの適用例を図 6 に示す．具体的な適用手順は以下の 3 段階である．

1. **E-CST への変換**

   Java ソースコードを JDT の AST パーサにより AST へ変換し，その AST を入力として E-CST を構築する．クラス，メソッド，継承関係，呼び出し関係，および JavaDoc 非推奨メッセージを E-CST 上のノードおよび関係として表現する．

2. **静的解析の実行**

   REPFINDER の探索処理を，E-CST Visitor を用いて再実装する．E-CST Visitor を継承した REPFINDER Visitor を実装し，継承関係やシグネチャ類似性の解析を，Visitor パターンによる E-CST 走査として実現する．走査では，欠落 API が属していたクラスの親クラスや子クラスを辿り，型やシグネチャが類似する API を候補として収集

する必要がある．E-CST では，継承関係や包含関係が明示的に表現されているため，Visitor パターンを用いた走査によって探索を実現できる．

3. **解析結果の収集・利用**

   欠落 API に対する代替 API 候補を抽出する．

このように，提案フレームワークを適用することで，REPFINDER の探索処理は Java の AST から独立し，E-CST に基づく言語非依存な解析処理として再構成されると考えられる．同様の E-CST パーサを他言語向けに実装することで，REPFINDER の解析器を再利用し，低コストで他のプログラミング言語へ拡張できると考えられる．

4.3 PYREF への適用

4.3.1 PYREF の概要

PYREF は，Python プロジェクトを対象として，リファクタリングを自動的に検出する静的解析ツールである．PYREF は，RefactoringMiner[12] から着想を得た手法を Python 向けに適用し，Python の AST を用いてコード要素をモデル化することで，メソッド単位のリファクタリング検出を実現している．具体的には，RENAME METHOD，ADD PARAMETER，EXTRACT METHOD，MOVE METHOD などを検出する．

PYREF は，リビジョン間で変更されたファイルのみを対象とし，Python の AST からモジュール，クラス，メソッド，文といったコード要素を抽出する．抽出した要素をノードとして表現し，異なるリビジョン間のノード同士を対応付けることで，どのノードがどのノードに変化したか特定し，リファクタリング候補を検出する．この解析処理は，Python の AST に強く依存している．

その結果，PYREF の探索アルゴリズム自体は他のプログラミング言語にも適用可能な性質を持つ一方で，実装上は Python 専用のツールとなっている．

4.3.2 PYREF への適用例

PYREF へのフレームワークの適用例を図 7 に示す．具体的な適用手順は以下の 3 段階である．

1. **E-CST への変換**

   Python ソースコードを，ast モジュールを用いて AST へ変換し，その AST を入力として E-CST を構築する．モジュール，クラス，メソッドといった PYREF の解析に必要なコード要素を，E-CST 上のノードとして表現する．

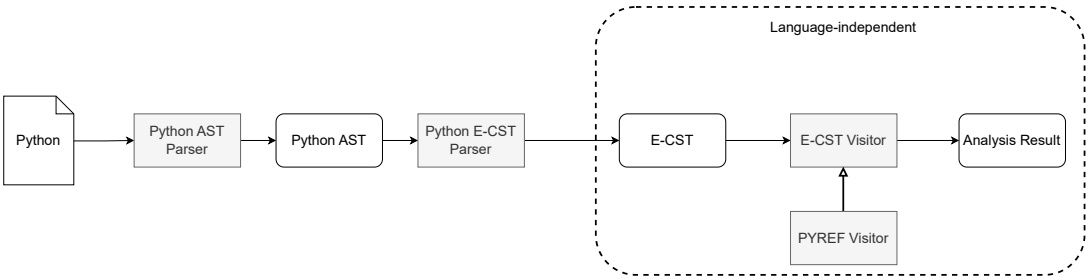

図 7: 提案フレームワークの適用例 (PYREF)

2. **静的解析の実行**

   PYREF が行うリファクタリング検出処理を，E-CST 上の解析として再実装する．E-CST Visitor を継承した PYREF Visitor を実装し，リファクタリング検出処理を，Visitor パターンによる E-CST 走査として実現する．PYREF のリファクタリング検出では，メソッド名，所属クラス，モジュール構造，およびコード要素間の対応関係が重要となる．E-CST では，これらの構造的情報が言語非依存な形式で保持されているため，Python の AST に依存することなく，Visitor パターンによる走査およびノード間関係の解析によって，リファクタリング検出を実現できる．

3. **解析結果の収集・利用**

   メソッド単位のリファクタリング操作を抽出する．

このように，提案フレームワークを適用することで，PYREF のリファクタリング検出処理は Python の AST から切り離され，E-CST に基づく言語非依存な解析処理として再構成されると考えられる．同様の E-CST パーサを他言語向けに実装することで，PYREF の解析器を再利用し，低コストで他のプログラミング言語へ拡張できると考えられる．

5　関連研究

本章では，静的解析を対象とした多言語拡張可能なフレームワークに関する関連研究について述べる．

5.1　LiSA

LiSA(Library for Static Analysis)[4] は，CFG ベースの静的解析ツールを対象とした多言語拡張可能なフレームワークである．現時点では Java，Python，Go の 3 言語を対象としている．

LiSA の基本的な設計思想は，解析対象のプログラミング言語に依存する処理と，言語非依存な解析アルゴリズムを明確に分離する点にある．具体的には，各言語のソースコードはフロントエンドによって LiSA 内部の中間表現である LiSA CFG へ変換され，その後の解析はすべてこの CFG 上で実行される．この設計思想は，LLVM における中間表現である LLVM IR を用いた多言語対応と同様であり，一度解析基盤を実装すれば，複数言語に対して共通の解析処理を再利用できる．

なお，LiSA では，データフロー解析および抽象解釈に基づく数値解析を中心に，幅広い静的解析を実装可能である．

このように，LiSA は CFG に基づく言語非依存な中間表現を導入することで，多言語対応かつ拡張性の高い静的解析フレームワークを実現している．

5.2　GAST

GAST(Generic Abstract Syntax Tree)[8] は，さまざまなプログラミング言語で記述されたソースコードを，言語非依存な AST へ変換することを目的とした静的解析基盤である．共通の構造を持つ GAST へ変換することで，単一の解析アルゴリズムによる静的解析を可能にする．現時点では Java, Python, C#などの言語に対応している．

GAST の基本的な設計思想は，各プログラミング言語固有の構文的差異を GAST 上で吸収し，構文構造に基づく解析を言語非依存に実施できるようにする点にある．各言語のソースコードは，既存のパーサを用いて AST へ変換された後，あらかじめ定義された対応規則に基づいて GAST へ変換される．この変換の正当性は，AST と GAST の構造的同値性を検証するバリデーション機構によって保証される．具体的には，ノード数や階層構造，親子関係が変化しないことが確認される．

GAST は主に，プログラムの構造やコード要素間の関係に着目した静的解析を想定している．具体的には，コードクローン検出，ソフトウェアメトリクス (サイクロマティック複雑度など) の計測などへの利用が期待されている．また，GAST は異なるプログラミング言

語で実装されたプログラムを，共通の表現上で比較できる点に特徴がある．これにより，言語を跨いだコードクローン検出や構造的類似度の解析など，従来は困難であった言語横断的な解析が可能となる．

GAST と E-CST の差異について述べる．GAST は各言語の AST の構造を可能な限り保持しつつ，それらを言語非依存な構造へ変換することを目的としている．そのため，文や式，制御構造といった細かい粒度の構文要素まで詳細に表現する汎用的な AST 基盤であり，コードクローン検出やメトリクス計測など，構文構造全体を対象とする解析に適している．これに対し，本研究で提案した E-CST は，AST 全体の統一を目的とするものではなく，クラス，メソッド，フィールドといった粗い粒度のコード要素に着目する点に特徴がある．すなわち，GAST は AST の一般化によって汎用的な多言語解析基盤を目指しているのに対し，E-CST は解析目的に基づく AST の一部の抽象化によって簡潔かつ軽量なデータ構造を目指している．また，GAST では AST との構造的同値性を維持することが設計上重要であるのに対し，E-CST では元の AST 構造を完全には保持せず，解析に不要な構文情報を意図的に捨象する．この点において，両者は設計方針および想定する解析粒度が本質的に異なる．

このように，GAST は汎用的な AST 統合基盤であるのに対し，E-CST は特定の静的解析を効率的に実現するという目的に特化したデータ構造であるという点で差異がある．

5.3 UNICOEN

UNICOEN[13] は，複数プログラミング言語に対応するソースコード処理フレームワークであり，言語とツールの多対多の関係を多対一に簡略化することを目的としている．そのために，シンタックスに基づく言語非依存な統合コードモデルを提供し，各言語のソースコードを統合コードモデルへマッピングすることで，共通 API 上で解析および変形処理を記述可能としている．統合コードモデルは木構造を有し，クラス・メソッド・ブロックなどを再帰的に保持する AST に近い構造を持つ．

これに対し，本研究で提案した E-CST は，言語とツールの関係を一般化するフレームワークではなく，AST ベースの特定の静的解析を効率的に実現するための目的に特化した中間表現である．UNICOEN が構文構造を網羅的に保持する言語非依存 AST 基盤であるのに対し，E-CST はクラス，メソッド，フィールドといった粗い粒度のコード要素に限定して情報を抽出し，解析に不要な構文情報は保持しない．

また，UNICOEN は解析のみならずコード変形も含めた汎用的ツール開発を想定し，ツール開発者向け API および言語拡張者向け API を提供するフレームワークである．一方，E-CST は独立した API 基盤を提供するものではなく，言語間で構造差異を吸収しつつリファクタリングの対応関係を同一基準で比較するためのデータ構造である．

このように，UNICOEN が言語非依存な AST フレームワークによって多様なツール開発

を支援する基盤であるのに対し，E-CST は特定の静的解析を効率的に実現するという目的に特化したデータ構造であるという点で，設計思想と適用範囲が異なる．

6　考察

本章では，E-CST に対する考察を述べる．

REPFINDER および PYREF への適用を通じて，本フレームワークは，異なるプログラミング言語を対象とし，かつ異なる解析目的を持つ AST ベースの静的解析ツールに対して，共通の解析基盤を提供可能であることを示した．この結果は，本フレームワークが高い汎用性を有することを示唆している．

従来，これらの静的解析ツールはそれぞれ対象言語の AST に強く依存して実装されており，多言語への拡張は困難であった．しかし，本フレームワークを用いることで，解析処理を言語非依存な部分として切り出すことができ，また前処理となる言語依存部分を E-CST パーサに分離できる可能性が示された．

一方で，本フレームワークは AST ベースの静的解析を対象としており，すべての解析を E-CST 上で完全に表現できるわけではない．例えば，式レベルの詳細な構文解析や制御フローを考慮する解析では，E-CST のみでは情報が不足する場合がある．このような制約を踏まえると，本研究のアプローチは，CFG ベースの解析基盤である LiSA と相補的な関係であると考えられる．

7　今後の課題

本章では，今後の課題について述べる．

7.1　E-CST パーサの整備

現時点では E-CST パーサは未整備であり，特に Python 用パーサは未実装である．Java, JavaScript，C についても，RefDiff 2.0 で用いられている CST パーサは存在するものの，E-CST パーサは未実装である．今後は，E-CST パーサを実装し，実際に REPFINDER や PYREF を多言語に拡張することで提案フレームワークの有効性を示す必要がある．

7.2　他の言語非依存な中間表現との連携

例えば，LiSA CFG が表現する制御フロー情報と，E-CST が保持する構造的情報を組み合わせることで，単一の中間表現では困難であった，より高度な静的解析の実現が期待される．

8　あとがき

本研究では，AST ベースの静的解析ツールを対象とした，多言語拡張可能なフレームワークとして E-CST を提案した．E-CST は，CST が持つ言語非依存性を維持しつつ，型情報や非推奨メッセージといった静的解析に有用な情報を拡張したデータ構造である．

また，既存の静的解析ツールである REPFINDER および PYREF に適用することで，解析処理を対象言語の AST から切り離し，言語非依存に再構成できる可能性を示した．これにより，単一言語向けに開発された AST ベースの静的解析ツールを，低コストで多言語に拡張できることが期待される．E-CST は，API 移行支援，リファクタリング検出，コード構造解析など，さまざまな AST ベースの静的解析への応用が可能であり，今後の静的解析基盤となり得る．

謝辞

大阪大学大学院情報科学研究科コンピュータサイエンス専攻 肥後 芳樹 教授には，中間報告会にて途中経過を発表し，議論する機会をいただき，研究および発表に関して多くの貴重なご助言を賜りました．肥後 芳樹 教授のご指導により本研究を進めることができました．心より深く感謝申し上げます．

大阪大学大学院情報科学研究科コンピュータサイエンス専攻 松下 誠 准教授には，毎週のミーティングをはじめ，多くの時間を割いて直接ご指導いただきました．研究の方針から細部に至るまで貴重な助言を賜るとともに，本論文の執筆および報告に関しても添削や相談にご協力いただきました．松下 誠 准教授の多大なご支援により，本研究を進めることができました．心より深く感謝申し上げます．

大阪大学大学院情報科学研究科コンピュータサイエンス専攻 Kula Raula Gaikovina 教授には，中間報告会にて研究に関するご助言・ご指摘をいただきました．加えて，報告会後も時間を割いて疑問へのご回答，および研究に関する最新情報などをご紹介いただきました．心より深く感謝申し上げます．

大阪大学大学院情報科学研究科コンピュータサイエンス専攻 Olivier Nourry 助教には，中間報告会などにおいて研究に関する多くのご助言・ご指摘をいただきました．加えて，報告会後も時間を割いて研究に関連する知識について丁寧にご説明いただきました．心より深く感謝申し上げます．

最後に，本研究に関してあらゆるご助言・ご助力をくださった大阪大学大学院情報科学研究科コンピュータサイエンス専攻肥後研究室の皆様に心より深く感謝申し上げます．

参考文献

[1] Alfred V. Aho, Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley, 2nd edition, 2007.

[2] Frances E. Allen. Control flow analysis. *ACM SIGPLAN Notices*, Vol. 5, No. 7, pp. 1–19, 1970.

[3] Hassan Atwi, Bin Lin, Nikolaos Tsantalis, Yutaro Kashiwa, Yasutaka Kamei, Naoyasu Ubayashi, Gabriele Bavota, and Michele Lanza. Pyref: Refactoring detection in python projects. In *IEEE 21st International Working Conference on Source Code Analysis and Manipulation (SCAM)*, pp. 136–141, 2021.

[4] Pietro Ferrara, Luca Negrini, Vincenzo Arceri, and Agostino Cortesi. Static analysis for dummies: Experiencing lisa. In *Proceedings of the 10th ACM SIGPLAN International Workshop on the State Of the Art in Program Analysis (SOAP)*, pp. 1–6, 2021.

[5] David Hovemeyer and William Pugh. Finding bugs is easy. *ACM SIGPLAN Notices*, Vol. 39, No. 12, pp. 92–106, 2004.

[6] Kaifeng Huang, Bihuan Chen, Linghao Pan, Shuai Wu, and Xin Peng. Repfinder: Finding replacements for missing apis in library update. In *Proceedings of the 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 266–278, 2021.

[7] Chris Lattner and Vikram Adve. Llvm: A compilation framework for lifelong program analysis & transformation. In *Proceedings of the 2004 International Symposium on Code Generation and Optimization (CGO)*, pp. 75–85, 2004.

[8] Jason Leiton-Jimenez, Luis Barboza-Artavia, Antonio Gonzalez-Torres, Pablo Brenes-Jimenez, Steven Pacheco-Portuguez, Jose Navas-Su, Marco Hernandez-Vasquez, Jennier Solano-Cordero, Franklin Hernandez-Castro, Ignacio Trejos-Zelaya, and Armando Arce-Orozco. Gast: A generic ast representation for language-independent source code analysis. *ENFOQUE UTE*, Vol. 14, No. 4, pp. 9–18, 2023.

[9] Caitlin Sadowski, Edward Aftandilian, Alex Eagle, Liam Miller-Cushon, and C. Jaspan. Lessons from building static analysis tools at google. *Communications of the ACM*, Vol. 61, No. 4, pp. 58–66, 2018.

[10] Danilo Silva, João Paulo da Silva, Gustavo Santos, Ricardo Terra, and Marco Tulio Valente. Refdiff 2.0: A multi-language refactoring detection tool. *IEEE Transactions on Software Engineering*, Vol. 47, No. 12, pp. 2786–2802, 2020.

[11] Danilo Silva and Marco Tulio Valente. Refdiff: Detecting refactorings in version histories. In *IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, pp. 269–278, 2017.

[12] Nikolaos Tsantalis, Ameya Ketkar, and Danny Dig. Refactoringminer 2.0. *IEEE Transactions on Software Engineering*, Vol. 48, No. 3, pp. 930–950, 2022.

[13] 坂本一憲, 大橋昭, 太田大地, 鷲崎弘宜, 深澤良彰. Unicoen: 複数プログラミング言語対応のソースコード処理フレームワーク. 情報処理学会論文誌, Vol. 54, No. 2, pp. 945–960, 2013.