

Modern Parsing Techniques

All You Need is Generalized GLL

Semyon Grigorev

October 31, 2022

Parsing: Problems and Challenges

Language specification approaches Zoo

- Lexer + Parser vs Scannerless: grammar over tokens vs grammar over characters
- Context-free grammars (CFG): BNF, EBNF, ...
- Data-dependent grammars: utilization of previously parsed data to guide parsing
- Rail diagrams: graphical language
- Parsing Expression Grammars (PEG): CFG with prioritized choice and manual control of lookahead
- ...

Language/grammar properties

- Determinism and ambiguity
- (Hidden) Left recursion
- ...

Technical challenges

- Performance
- Error recovery
- Incremental parsing
- Manual control of behavior

- Adaptive LL(*) Parsing: The Power of Dynamic Analysis
- ALL(*)
- Huge grammar zoo
- Can generate parser in Java, C#, Python, JavaScript, Go, C++, Swift, Dart, PHP
- + Simple error recovery
- + Left recursion
- + Prioritized choice
- + Parsing tree and visitors, listeners for it

Grammar	KB/sec
XML	45,993
Java	24,972
JSON	17,696
DOT	16,152
Lua	5,698
C	4,238
Verilog2001	1,994
Erlang	751

Figure: Throughput of lexing+ parsing; all input preloaded into RAM²

²From "Adaptive LL(*) Parsing: The Power of Dynamic Analysis"

- GLL-based
- Sources
- Practical General Top-Down Parsers
- + Data-dependent grammars
- + Left recursion
- + Prioritized choice
- + Operator associativity
- + Layout rules

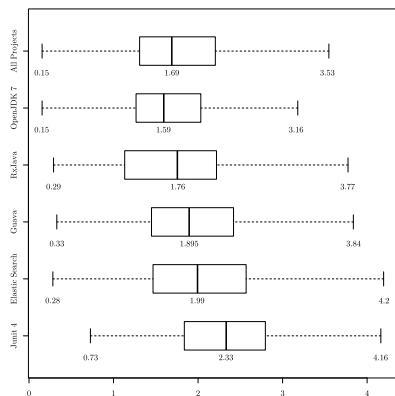


Figure: Iguana performance relative to ANTLR³

³From "Practical General Top-Down Parsers"

- GLR-based
- Sources
- + Huge grammar zoo
- + Incremental
- + Error recovery
- + Operators precedence and associativity
- Custom lexers to handle layout
- Parsers in C with bindings to other languages
- Prioritized choice

SDF3 (Spoofax)

- GLR-based
- + Modular
- + Layout rules
- + Operators associativity
- ? Error recovery
 - Incremental version under development⁴

⁴Incremental scannerless generalized LR parsing

SDF3 (Spoofax)

- GLR-based
- + Modular
- + Layout rules
- + Operators associativity
- ? Error recovery
 - Incremental version under development⁴
 - Part of Spoofax language workbench
 - Sources
 - Documentation
 - Multi-purpose Syntax Definition with SDF3

⁴Incremental scannerless generalized LR parsing

Parser combinators

- 👎 Poor performance
- ✗ Left recursion
- ✗ Ambiguity
- ✗ Error recovery
- ✗ Incremental parsing

Parser combinators

- 👎 Poor performance
- ✗ Left recursion
- ✗ Ambiguity
- ✗ Error recovery
- ✗ Incremental parsing
- 💡 Modern combinator library is not just a set of functions
 - 👍 Good performance
 - ✚ Left recursion
 - ✚ Ambiguity

Parser combinators

- 👎 Poor performance
- ✗ Left recursion
- ✗ Ambiguity
- ✗ Error recovery
- ✗ Incremental parsing
- 💡 Modern combinator library is not just a set of functions
 - 👍 Good performance
 - ✚ Left recursion
 - ✚ Ambiguity
 - ❓ Error recovery
 - ❓ Incremental parsing

Parser combinators

- 👎 Poor performance
- ✗ Left recursion
- ✗ Ambiguity
- ✗ Error recovery
- ✗ Incremental parsing
- 💡 Modern combinator library is not just a set of functions
 - 👍 Good performance
 - ✚ Left recursion
 - ✚ Ambiguity
 - ❓ Error recovery
 - ❓ Incremental parsing
 - 👎 You have no control on nontrivial machinery inside

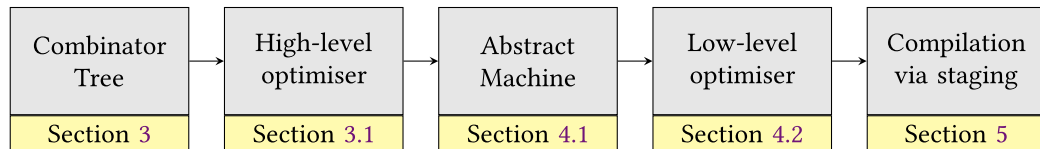


Figure: Parsley internal pipeline⁵

⁵From “Staged Selective Parser Combinators”

⁶From “Staged Selective Parser Combinators”

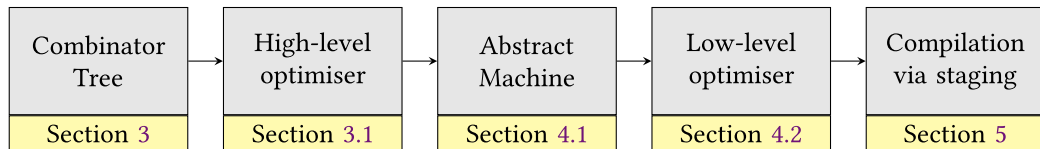


Figure: Parsley internal pipeline⁵

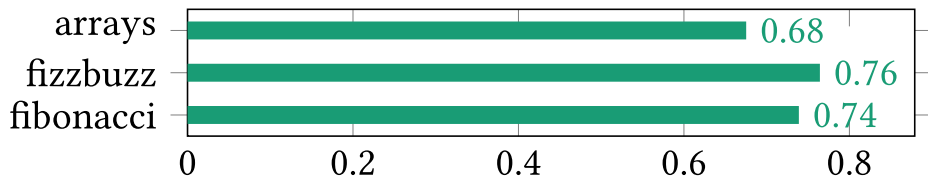


Figure: Parsley performance relative to Bison⁶

⁵From “Staged Selective Parser Combinators”

⁶From “Staged Selective Parser Combinators”

- GLL-like engine for CPS parsers
- Practical, general parser combinators

⁷From “Practical, general parser combinators”

- GLL-like engine for CPS parsers
- Practical, general parser combinators
- + Left recursion
- + Prioritized choice
- + Operator associativity

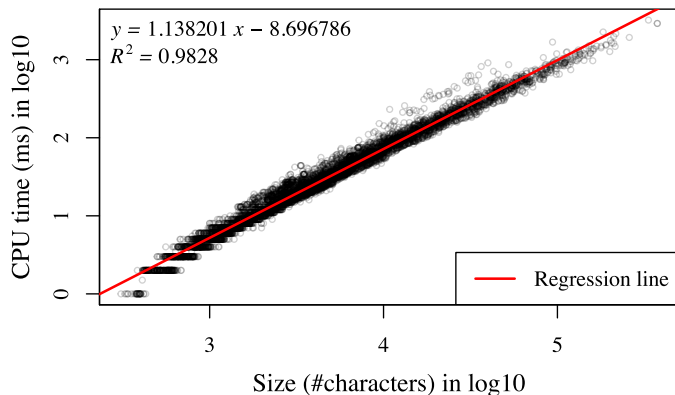


Figure: Meerkat performance on Java files⁷

⁷From “Practical, general parser combinators”

- Parsing Expression GLL
- Sources

- Don't Panic! Better, Fewer, Syntax Errors for LR Parsers
- Problems with lexer

- Natural precise error recovery (LED)
- String-embedded DSL parsing
- ...