

Intrinsic Verification of Parsers in Dependent Lambek Calculus

Steven Schaefer¹ Nathan Varner¹ Pedro H. Azevedo de Amorim²
Max S. New¹

November 22, 2024

¹University of Michigan

²University of Oxford



Parsing is Everywhere!

Source Code

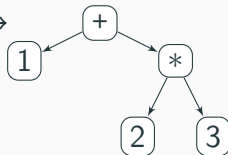
"1+2*3"



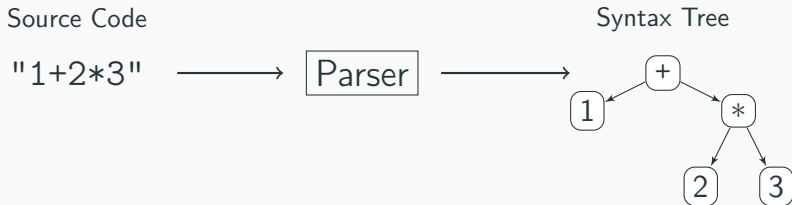
Parser



Syntax Tree

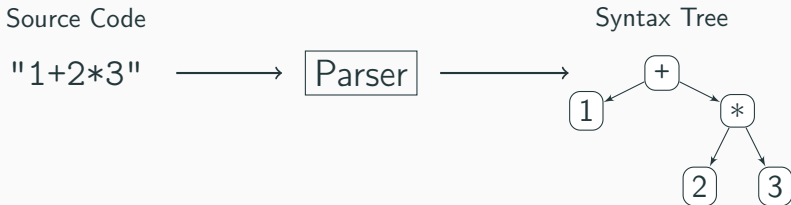


Parsing is Everywhere!



- Incorrect parsers jeopardize verified developments

Parsing is Everywhere!



- Incorrect parsers jeopardize verified developments
- Early versions of CompCert
 - 0 middle or backend bugs, 6 parser bugs (Yang et al., 2011)

A **parser** for grammar A is a partial function $f : \text{String} \rightarrow \text{Parse}(A)$

Parser Correctness

A **parser** for grammar A is a partial function $f : \text{String} \rightarrow \text{Parse}(A)$

f is correct if it is **sound** and **complete**

Parser Correctness

A **parser** for grammar A is a partial function $f : \text{String} \rightarrow \text{Parse}(A)$

f is correct if it is **sound** and **complete**

Soundness of f

$f(w)$, if it exists, is a parse tree of w for the grammar A

A **parser** for grammar A is a partial function $f : \text{String} \rightarrow \text{Parse}(A)$

f is correct if it is **sound** and **complete**

Soundness of f

$f(w)$, if it exists, is a parse tree of w for the grammar A

Completeness of f

If a parse tree of w for A exists, then $f(w)$ is defined

Parser Correctness

A **parser** for grammar A is a partial function $f : \text{String} \rightarrow \text{Parse}(A)$

f is correct if it is **sound** and **complete**

Soundness of f

$f(w)$, if it exists, is a parse tree of w for the grammar A

Completeness of f

If a parse tree of w for A exists, then $f(w)$ is defined

Our Approach

Define a domain-specific language in which all parsers are
sound-by-construction

1. A DSL for Verified Parsing
2. An Example Parser
3. Implementation
4. Future Work

1. A DSL for Verified Parsing

2. An Example Parser

3. Implementation

4. Future Work

Dependent Lambek Calculus

- Non-commutative linear logic

Dependent Lambek Calculus

- Non-commutative linear logic

Grammars	Linear Types
Grammar A	Linear type A

Domain-Specific Language for Verified Parsers

Dependent Lambek Calculus

- Non-commutative linear logic

Grammars	Linear Types
Grammar A	Linear type A
Parse of string w	$w \vdash A$

Dependent Lambek Calculus

- Non-commutative linear logic

Grammars	Linear Types
Grammar A	Linear type A
Parse of string w	$w \vdash A$
Parser	$\text{String} \vdash A \oplus A_{\neg}$

Domain-Specific Language for Verified Parsers

Dependent Lambek Calculus

- Non-commutative linear logic

Grammars	Linear Types
Grammar A	Linear type A
Parse of string w	$w \vdash A$
Parser	$\text{String} \vdash A \oplus A_{\neg}$
Parse transformer	$\Delta \vdash A$

Domain-Specific Language for Verified Parsers

Dependent Lambek Calculus

- Non-commutative linear logic

Grammars	Linear Types
Grammar A	Linear type A
Parse of string w	$w \vdash A$
Parser	$\text{String} \vdash A \oplus A_{\neg}$
Parse transformer	$\Delta \vdash A$

Implemented in Agda and instantiated with verified parsers for regular expressions and selected context-free grammars

Why Linearity?

An ordered and linear type system

Why Linearity?

An ordered and linear type system

"cat" ⊢ A

Why Linearity?

An ordered and linear type system

"cat" \vdash A

$x : 'c', y : 'a', z : 't' \vdash (x, (y, z)) : 'c' \otimes 'a' \otimes 't'$

Why Linearity?

An ordered and linear type system

"cat" $\vdash A$

$x : 'c', y : 'a', z : 't' \vdash (x, (y, z)) : 'c' \otimes 'a' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (y, (x, z)) : 'a' \otimes 'c' \otimes 't'$

Why Linearity?

An ordered and linear type system

"cat" \vdash A

$x : 'c', y : 'a', z : 't' \vdash (x, (y, z)) : 'c' \otimes 'a' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (y, (x, z)) : 'a' \otimes 'c' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (x, (y, (z, z))) : 'c' \otimes 'a' \otimes 't' \otimes 't'$

Why Linearity?

An ordered and linear type system

"cat" $\vdash A$

$x : 'c', y : 'a', z : 't' \vdash (x, (y, z)) : 'c' \otimes 'a' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (y, (x, z)) : 'a' \otimes 'c' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (x, (y, (z, z))) : 'c' \otimes 'a' \otimes 't' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (y, z) : 'a' \otimes 't'$

Why Linearity?

An ordered and linear type system

"cat" $\vdash A$

$x : 'c', y : 'a', z : 't' \vdash (x, (y, z)) : 'c' \otimes 'a' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (y, (x, z)) : 'a' \otimes 'c' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (x, (y, (z, z))) : 'c' \otimes 'a' \otimes 't' \otimes 't'$

$x : 'c', y : 'a', z : 't' \not\vdash (y, z) : 'a' \otimes 't'$

These restrictions ensure parsers are **sound-by-construction**

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string \mathbf{I}

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap , \multimap

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$
- Indexed disjunction $\bigoplus_{x:X} A\ x$, indexed conjunction $\big\&_{x:X} A\ x$ over a non-linear type X

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$
- Indexed disjunction $\bigoplus_{x:X} A\ x$, indexed conjunction $\big\&_{x:X} A\ x$ over a non-linear type X
- Indexed inductive linear types

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$
- Indexed disjunction $\bigoplus_{x:X} A\ x$, indexed conjunction $\big\&_{x:X} A\ x$ over a non-linear type X
- Indexed inductive linear types

Non-linear Types

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$
- Indexed disjunction $\bigoplus_{x:X} A\ x$, indexed conjunction $\big\&_{x:X} A\ x$ over a non-linear type X
- Indexed inductive linear types

Non-linear Types

- Ordinary dependent types

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$
- Indexed disjunction $\bigoplus_{x:X} A\ x$, indexed conjunction $\big\&_{x:X} A\ x$ over a non-linear type X
- Indexed inductive linear types

Non-linear Types

- Ordinary dependent types
- $\uparrow: \text{LinTy} \rightarrow \text{NonLinTy}$

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$
- Indexed disjunction $\bigoplus_{x:X} A\ x$, indexed conjunction $\big\&_{x:X} A\ x$ over a non-linear type X
- Indexed inductive linear types

Non-linear Types

- Ordinary dependent types
- $\uparrow: \text{LinTy} \rightarrow \text{NonLinTy}$
 - $\uparrow A$ is the type of A -parses for the empty string

Types in Dependent Lambek Calculus

Linear Types

- Characters in the alphabet — i.e. 'a'
- Empty string I
- Concatenation \otimes
- Linear function types \multimap, \multimap
- Disjunction $(\oplus, 0)$, conjunction $(\&, \top)$
- Indexed disjunction $\bigoplus_{x:X} A\ x$, indexed conjunction $\big\&_{x:X} A\ x$ over a non-linear type X
- Indexed inductive linear types

Non-linear Types

- Ordinary dependent types
- $\uparrow: \text{LinTy} \rightarrow \text{NonLinTy}$
 - $\uparrow A$ is the type of A -parses for the empty string
 - $!$ in linear logic, \Box in separation logic

1. A DSL for Verified Parsing

2. An Example Parser

3. Implementation

4. Future Work

- Context-free grammar of balanced parentheses

- Context-free grammar of balanced parentheses
- Define as an inductive linear type

Dyck Grammar

- Context-free grammar of balanced parentheses
- Define as an inductive linear type

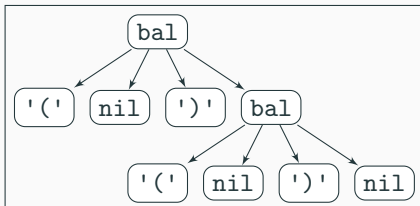
```
data Dyck : LinTy where  
  nil : ↑ Dyck  
  bal : ↑('(' —o Dyck —o ')' —o Dyck —o Dyck)
```


Dyck Grammar

- Context-free grammar of balanced parentheses
- Define as an inductive linear type

```
data Dyck : LinTy where  
  nil : ↑ Dyck  
  bal : ↑('(' —o Dyck —o ')' —o Dyck —o Dyck)
```

"()()" ⊢ bal l1 nil r1 (bal l2 nil r2) : Dyck



A Parser for the Dyck Grammar

Dyck Parser

A parser for Dyck is a function

$$\uparrow (\text{String} \multimap \text{Dyck} \oplus \text{Dyck}_{\neg})$$

where $\text{Dyck} \& \text{Dyck}_{\neg} \cong 0$

A Parser for the Dyck Grammar

Dyck Parser

A parser for Dyck is a function

$$\uparrow (\text{String} \multimap \text{Dyck} \oplus \text{Dyck}_{\neg})$$

where $\text{Dyck} \& \text{Dyck}_{\neg} \cong 0$

Soundness: guaranteed by the linear type system

A Parser for the Dyck Grammar

Dyck Parser

A parser for Dyck is a function

$$\uparrow (\text{String} \multimap \text{Dyck} \oplus \text{Dyck}_{\neg})$$

where $\text{Dyck} \& \text{Dyck}_{\neg} \cong 0$

Soundness: guaranteed by the linear type system

Completeness: guaranteed by the disjointness of Dyck and Dyck_{\neg}

A Parser for the Dyck Grammar

Dyck Parser

A parser for Dyck is a function

$$\uparrow (\text{String} \multimap \text{Dyck} \oplus \text{Dyck}_{\neg})$$

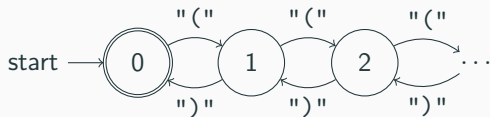
where $\text{Dyck} \& \text{Dyck}_{\neg} \cong 0$

Soundness: guaranteed by the linear type system

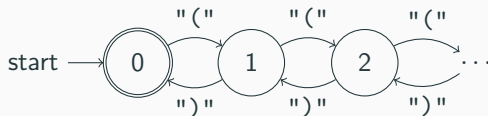
Completeness: guaranteed by the disjointness of Dyck and Dyck_{\neg}

Need an automaton to define the parser!

Dyck Traces

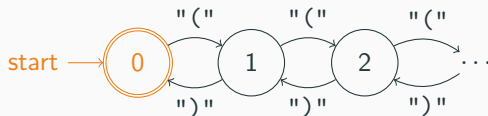


Dyck Traces



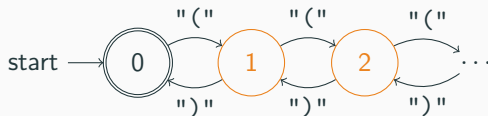
```
data Trace : Bool → ℕ → LinTy where
  eof : ↑ (Trace true 0)
  leftovers : ∀ n → ↑ (Trace false (suc n))
  push : ∀ b n → ↑('(' -> Trace b (suc n) -> Trace b n)
  pop : ∀ b n → ↑(')' -> Trace b n -> Trace b (suc n))
  unexpected : ↑(')' -> ⊥ -> Trace false 0)
```

Dyck Traces



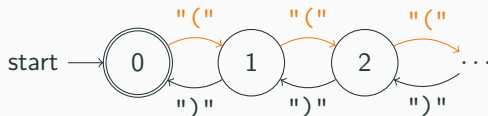
```
data Trace : Bool → ℕ → LinTy where
  eof : ↑ (Trace true 0)
  leftovers : ∀ n → ↑ (Trace false (suc n))
  push : ∀ b n → ↑('(' -> Trace b (suc n) -> Trace b n)
  pop : ∀ b n → ↑(')' -> Trace b n -> Trace b (suc n))
  unexpected : ↑(')' -> ⊥ -> Trace false 0)
```


Dyck Traces



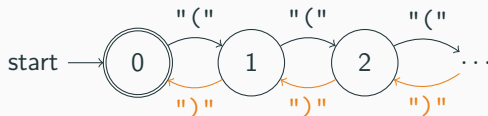
```
data Trace : Bool → ℕ → LinTy where
  eof : ↑ (Trace true 0)
  leftovers : ∀ n → ↑ (Trace false (suc n))
  push : ∀ b n → ↑('(' -> Trace b (suc n) -> Trace b n)
  pop : ∀ b n → ↑(')' -> Trace b n -> Trace b (suc n))
  unexpected : ↑(')' -> ⊥ -> Trace false 0)
```

Dyck Traces



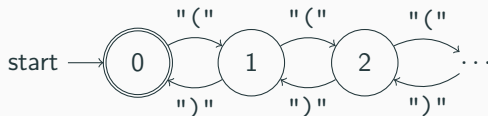
```
data Trace : Bool → ℕ → LinTy where
  eof : ↑ (Trace true 0)
  leftovers : ∀ n → ↑ (Trace false (suc n))
  push : ∀ b n → ↑('(' -> Trace b (suc n) -> Trace b n)
  pop : ∀ b n → ↑(')' -> Trace b n -> Trace b (suc n))
  unexpected : ↑(')' -> ⊥ -> Trace false 0)
```

Dyck Traces



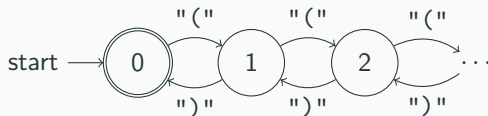
```
data Trace : Bool → ℕ → LinTy where
  eof : ↑ (Trace true 0)
  leftovers : ∀ n → ↑ (Trace false (suc n))
  push : ∀ b n → ↑('(' -> Trace b (suc n) -> Trace b n)
  pop : ∀ b n → ↑(')' -> Trace b n -> Trace b (suc n))
  unexpected : ↑(')' -> ⊥ -> Trace false 0)
```

Dyck Traces



```
data Trace : Bool → ℕ → LinTy where
  eof : ↑ (Trace true 0)
  leftovers : ∀ n → ↑ (Trace false (suc n))
  push : ∀ b n → ↑('(' -> Trace b (suc n) -> Trace b n)
  pop : ∀ b n → ↑(')' -> Trace b n -> Trace b (suc n))
  unexpected : ↑(')' -> ⊤ -> Trace false 0)
```

Dyck Traces



```
data Trace : Bool → ℕ → LinTy where
  eof : ↑ (Trace true 0)
  leftovers : ∀ n → ↑ (Trace false (suc n))
  push : ∀ b n → ↑('(' -> Trace b (suc n) -> Trace b n)
  pop : ∀ b n → ↑(')' -> Trace b n -> Trace b (suc n))
  unexpected : ↑(')' -> ⊤ -> Trace false 0)
```

Deterministic Automaton

$$\text{String} \cong \text{Trace true } 0 \oplus \text{Trace false } 0$$

A Parser for the Dyck Grammar

Deterministic Automaton

$$\text{String} \cong \text{Trace true } 0 \oplus \text{Trace false } 0$$

The Dyck Grammar is Recognized by the Automaton

$$\text{Dyck} \cong \text{Trace true } 0$$

A Parser for the Dyck Grammar

Deterministic Automaton

$$\text{String} \cong \text{Trace true } 0 \oplus \text{Trace false } 0$$

The Dyck Grammar is Recognized by the Automaton

$$\text{Dyck} \cong \text{Trace true } 0$$

$$\begin{aligned} \text{String} &\multimap \text{Trace true } 0 \oplus \text{Trace false } 0 \\ &\multimap \text{Dyck} \oplus \text{Trace false } 0 \end{aligned}$$

A Parser for the Dyck Grammar

Deterministic Automaton

$$\text{String} \cong \text{Trace true } 0 \oplus \text{Trace false } 0$$

The Dyck Grammar is Recognized by the Automaton

$$\text{Dyck} \cong \text{Trace true } 0$$

$$\text{String} \multimap \text{Trace true } 0 \oplus \text{Trace false } 0$$

$$\multimap \text{Dyck} \oplus \text{Trace false } 0 \quad \text{A parser!}$$

A Parser for Dyck Traces

Looks like an ordinary functional program

```
parse :  
  ↑(String → &[ n ∈ ℕ ] ⊕[ b ∈ Bool ] Trace b n)  
parse nil zero = σ true eof  
parse nil (suc n) = σ false leftovers  
parse (cons (σ '(' a) w) n =  
  let σ b tr = parse w (suc n) in  
  σ b (push a tr)  
parse (cons (σ ')' a) w) zero =  
  σ false (unexpected a _)  
parse (cons (σ ')' a) w) (suc n) =  
  let σ b tr = parse w n in  
  σ b (pop a tr)
```

1. A DSL for Verified Parsing

2. An Example Parser

3. **Implementation**


4. Future Work

Denotational semantics

$$\llbracket A \rrbracket : \text{String} \rightarrow \text{Set} \quad \llbracket A \rrbracket w = \{\text{parse trees of } w \text{ for } A\}$$

Denotational semantics

$$\llbracket A \rrbracket : \text{String} \rightarrow \text{Set} \quad \llbracket A \rrbracket w = \{\text{parse trees of } w \text{ for } A\}$$

Implementation of Dependent Lambek Calculus shallowly embedded in Cubical Agda 

- $\text{LinTy} := \text{String} \rightarrow \text{Set}$
- **Pros:** reuses the Agda typechecker
- **Cons:** poor performance and enforced combinatory style

Dyck grammar ($LL(0)$)

Dyck grammar (LL(0))

Arithmetic expressions with a binary operation (LL(1))

Dyck grammar (LL(0))

Arithmetic expressions with a binary operation (LL(1))

Regular expressions

- Equivalences between NFAs, DFAs, and regexes
- Thompson's construction, powerset construction

1. A DSL for Verified Parsing

2. An Example Parser

3. Implementation

4. Future Work

Parsing

- LL/LR parser generator

Parsing

- LL/LR parser generator
- Semantic actions

Parsing

- LL/LR parser generator
- Semantic actions

Implementation

Parsing

- LL/LR parser generator
- Semantic actions

Implementation

- Performance, ergonomics

Parsing

- LL/LR parser generator
- Semantic actions

Implementation

- Performance, ergonomics

Typechecking?

Dependent Lambek Calculus

Grammars	Linear Types
Grammar A	Linear type A
Parse of string w	$w \vdash A$
Parser	$\text{String} \vdash A \oplus A_{\neg}$
Parse transformer	$\Delta \vdash A$

Grammars are types! Automata are types!

Language and examples implemented in Cubical Agda 🙌

- github.com/maxsnew/grammars-and-semantic-actions

Preprint of this work

- maxsnew.com/docs/lambek.pdf