Unicode in Stak Scheme

@raviqqe

November 15, 2025

Contents

Stak Scheme

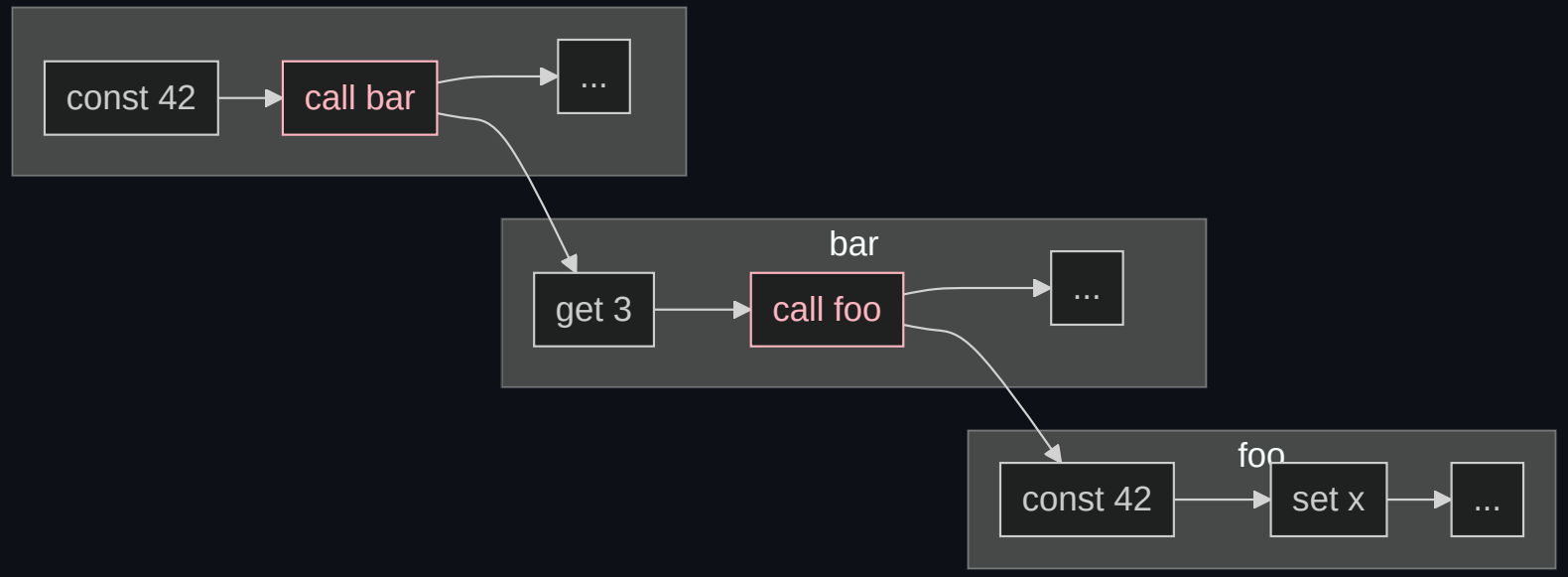
- A bytecode compiler and virtual machine (VM) for Scheme
  - The compiler is written in Scheme.
  - The VM is written in Rust.
- It implements the R7RS-small standard.

Progress

- Backtrace on errors
- Unicode in `(scheme char)`

Backtrace on errors

- On the VM, instructions are in a linked list.
- On a call instruction, the VM calls its procedure storing a return address of the **current** call instruction.
  - We cannot go back when it points to the next instruction.



Example

Source code:

```
(import (scheme base))

(define (foo)
  (error "Oh, no!" 42)
  #f)

(let ()
  (foo)
  #f))
```

Output:

```
Oh, no! 42 [error foo eval #f]
```

Unicode in `(scheme char)`

- Stak Scheme already supports Unicode in I/O.
  - i.e. UTF-8 encoding
- Now, its `(scheme char)` library also supports Unicode.
- Unicode defines multiple tables for character properties.
  - e.g. categories, and case mappings
- The tables can be fairly large...
  - One of Stak Scheme's goals is small memory footprints.

Encoding Unicode tables

Example: Upper to lower case mapping

1. Parse a table.

```
; A -> a, B -> b, C -> c, ... Z -> z
((65 97) (66 98) (67 99) #| ... |# (90 122))
```

2. Calculate differences between rows.
   - Small integers are encoded into small bytes in bytecode encoding.

```
((65 97) (1 1) (1 1) #| ... |# (1 1))
```

3. Apply run-length encoding.

```
((65 97) (24 . 1))
```

Asymmetric mapping

```
(import (scheme base) (scheme char) (scheme write))

(write
  (map
    (lambda (char) (cons char (char->integer char)))
    (list #\ß (char-upcase #\ß) (char-downcase #\ß))))
```

```
> stak ~/foo.scm # same for guile and gosh
((#\ß . 223) (#\ß . 223) (#\ß . 223))
> chibi-scheme ~/foo.scm
((#\ß . 223) (#\ß . 7838) (#\ß . 223))
```

Asymmetric mapping

- UnicodeData file format
  > Uppercase mapping
  >
  > These mappings are always one-to-one, not one-to-many or many-to-one.
- ß - Wikipedia
  > Because ⟨ß⟩ had been treated as a ligature, ... it had no capital form in early modern typesetting. ... A capital was first seriously proposed in 1879, but did not enter official or widespread use.

🤯🤯🤯

Future work

- LZSS bytecode compression
- Soft float in Scheme
- DSW garbage collection