Forth Recognizer -- Request For Discussion

Author: Matthias Trute
Contact: mtrute@web.de

Version: 1
Date: 3.10.2014
Status: Final

Background

I'm working on a Forth for 8-bit micro-controllers for 8 years now (amforth.sf.net). It is not only a useful tool for serious work but a nice playground to experiment with Forth too.

In 2011 my Forth got a floating point library. Since a micro-controller is (was) a resource constrained system it is not an option to include it permanently. It has to be a loadable module. Therefore I needed a way to keep the core system small but at the same time able to fully handle the new numbers. All but one problem were easy to fix. Only adding the number format to the Forth interpreter turned out to be serious one. I searched the net for ways to extend the Forth interpreter. What I found was having many hooks in the interpreter (U. Hoffman, Euroforth 2008) or a conditional re-compile of the sources with an autotool/configure like build system. Nothing really convinced me or my users. While googling I stumbled across the number parsing prefix discussion in c.l.f in 2007. The ideas sketched there looked promising so I stopped searching and started with them to invent my own solution.

I changed the Forth interpreter into a dumb tool, that delegates all data related work to modules, which can be changed at run-time. That made it possible to load the FP library into a running system that afterwards was able to deal with the new numbers like native ones. Surprisingly the new system had no disadvantages in speed or size compared the old one, something I consider very important on a micro-controller.

Shortly thereafter, Bernd Paysan got interested in what I did (we have regular IRC sessions on Forth topics) and started to implement recognizers in gforth. He suggested changes that further simplified my concept and made it more flexible.

By now we reached a point that justifies the public review. There are two very different Forth's available (both GPL'ed) that implement recognizers. A third implementation is in the proposal (public domain).

A recognizer written for one Forth works without modification for the other ones too. The words used to actually implement a recognizer (mostly string processing) need to be available of course. E.g. I wrote a recognizer for time stamp strings with gforth that converts the hh:mm:ss notation into a double cell number for the seconds since midnight. The code runs on amforth too. Gforth is a 64-bit system on the PC, amforth a 16-bit system on an 8-bit micro-controller (hence the double numbers). With that, something like

```
: test 01:00:01 d. ."  seconds since midnight" ; ok
test 3601 seconds since midnight ok
01:01:00 01:00:01 d+ d. 7261 ok
```

is possible. Similarly strings: everything that starts with a " is a string until the closing " is reached. Further string handling get the addr/len without the enclosing ".

```
: test "A string" type ; ok
test A string ok
" Another string" type ok  Another string
```

Another use case are name-spaces with word lists, without touching ORDER:

```
: test i2c.begin i2c.sendbyte i2c.end ;
```

where `begin/sendbyte/end` are words from the word-list identified with `i2c` (a constant with the wid). The recognizer splits the word at the first dot and uses the left sub-word to get the a word-list. In that word-list it searches with the remaining string and handles the result just like an ordinary dictionary search: interpret, compile (or not found).

Implementations for these examples are available in the respective Forth systems.

Problem

The Forth compiler can be extended easily. The Forth interpreter however has a fixed set of capabilities as outlined in section 3.4 of the standard text: Words from the dictionary and some numbers.

It's not easily possible to use the Forth text interpreter in an application or system extension context. The building blocks (`FIND`, `COMPILE,`, `>NUMBER` etc) are available but there is a gap between them and what the Forth interpreter does. Applications need to use either additional system provided and system specific intermediate words (if available) or have to re-invent the wheel to use e.g. numbers with a sign or hex numbers with the $ prefix.

Some Forth interpreters have ways to add new data types. That makes it possible to use a loadable library to implement new data types to be handled like the built-in ones. An example are the floating point numbers. They have their own parsing and data handling words including a stack of their own.

To actually handle data in the Forth context, the processing actions need to be `STATE` aware. It would be nice if the Forth text interpreter, that maintains `STATE`, is able to do the data processing without exposing `STATE` to the data handling methods. For that the different methods need to be registered somehow.

Whenever the Forth text interpreter is mentioned, the standard words `EVALUATE` (CORE), `INCLUDE-FILE` (FILE), `INCLUDED` (FILE), `LOAD` (BLOCK) and `THRU` (BLOCK) are expected to act likewise.

Solution

The monolithic design of the Forth interpreter is factored into three major blocks: First the interpreter. It maintains `STATE` and organizes the work. Second the actual data parsing. It is called from the interpreter and analyses strings (usually sub-strings of `SOURCE`) if they match the criteria for a certain data type. These parsing words are grouped as a stack to achieve an order of invocation. The result of the parsing words is handed over by the interpreter to data specific handling methods. There are three different methods for each data type depending on `STATE` and to `POSTPONE` the data.

The combination of a parsing word and the set of data handling words to deal with the data is called a recognizer. There is no strict 1:1 relation between the parsing words and the data handling sets. A data handling set for e.g. single cell numbers can be used by different parsing words.

Proposal

XY. The optional Recognizer word set

XY.1 Introduction

The algorithm of the Forth text interpreter as described in section 3.4 is modified as follows. All subsections of 3.4 apply unchanged.

a. Skip leading spaces and parse a name. Leave if the parsing area is empty.

b. For each element of the recognizer stack, starting with the top element, call its parsing method with the sub-string "name" from step a).

Every parsing method returns an information token and the parsed data from the analyzed sub-string if successful. Otherwise it returns the system provided failure token `R:FAIL` and no further data.

Continue with the next element in the recognizer stack until either all are used or the information token returned from the parsing word is not the system provided failure token `R:FAIL`.

c. Use the information token and do one of the following

   1. if interpreting execute the interpret method associated with the information token.
   2. if compiling execute the compile method associated with the information token.

d. Continue with a)

A recognizer consists of two elements: a parsing word (`REC:FOO`) and one or more information tokens returned by the parsing words that identify the parsed data and provide methods to perform the various semantics of the data (interpret, compile and postpone). A parsing word can return different information tokens. A particular information token can be used by different parsing words.

There is a system provided information token called `R:FAIL`. It is used if no other token is applicable. This failure token is associated with the system error actions if used in step c).

The parsing word of a recognizer has the stack effect

```
REC:FOO ( addr len -- i*x R:FOO | R:FAIL )
```

"addr/len" is a sub-string provided by the Forth text interpreter inside `SOURCE`. The parsing word must not change the string content. It can change `>IN` however.

"i*x" is the result of the text parsing of the string found at "addr/len". `R:FOO` is the information token that the interpreter uses to execute the interpret, compile or postpone actions for the data "i*x".

All three methods are called with the "i*x" data as left by the parsing word.

```
R:FOO:METHOD ( ... i*x -- j*y )
```

They can have additional stack effects, depending on what `R:FOO:METHOS` actually does.

The data items "i*x" don't have to be on the data stack, they can be at different places, when applicable. E.g. floating point numbers have a stack of their own. In this case, the data stack contains the `R:FOO` information only.

The names `R:FOO`, `REC:FOO` and `R:FOO:METHOD` don't have to actually exist, except the `R:FAIL` name.

A Forth system shall provide recognizers for integer numbers (both single and double precision) and the word look-up in the dictionary. They shall be ordered in a way that the word look-up is called first followed by the one(s) for numbers.

There shall be at least 4 recognizer slots available for application use.

XY.2 Additional terms and notations

Information token: A single cell number. It identifies the data type and a method table to perform the data processing of the interpreter. A naming convention suggests that the names start with `R:`.

Recognizer: A combination of a text parsing word that returns information tokens together with parsed data if successful. The text parsing word is assumed to run in cooperation with `SOURCE` and `>IN`. A naming convention suggests that the names start with `REC:`.

XY.3 Additional usage requirements

XY.3.1 Environment Queries

Obsolete.

XY.4 Additional documentation requirements

XY.4.1 System documentation

XY.4.1.1 Implementation-defined options

No additional options.

XY.4.1.2 Ambiguous conditions

- An empty recognizer stack.
- Changing the content of the parsed string during parsing.

XY.4.2 Program documentation

- No additional dependencies.

XY.5 Compliance and labelling

The phrase "Providing the Recognizer word set" shall be appended to the label of any Standard System that provides all of the Recognizer word set.

XY.6 Glossary

XY.6.1 Recognizer words

DO-RECOGNIZER (addr len -- i*x R:FOO | R:FAIL) RECOGNIZER

Apply the string at "addr/len" to the elements of the recognizer stack. Terminate the iteration if either a recognizer returns a information token that is different from `R:FAIL` or the stack is exhausted. In this case, return `R:FAIL`, otherwise `R:FOO`.

"i*x" is the result of the parsing word. It may be on other locations than the data stack. In this case the stack diagram should be read accordingly.

There is an ambiguous condition if the recognizer stack is empty.

GET-RECOGNIZERS (-- rec-n .. rec-1 n) RECOGNIZER

Return the execution tokens `rec-1 .. rec-n` of the parsing words in the recognizer stack. `rec-1` identifies the recognizer that is called first and `rec-n` the word that is called last.

The recognizer stack is unaffected.

MARKER ("<spaces>name" --) RECOGNIZER

Extend `MARKER` to include the current recognizer stack in the state preservation.

R:FAIL (-- R:FAIL) RECOGNIZER

A constant cell sized information token with two uses: first it is used to deliver the information that a specific recognizer could not deal with the string passed to it. Second it is a predefined information token whose elements are used when no recognizer from the recognizer stack could handle the passed string. These methods provide the system error actions.

The actual numeric value is system dependent and has no predictable value.

RECOGNIZER: (XT-INTERPRET XT-COMPILE XT-POSTPONE "<spaces>name" --) RECOGNIZER

Skip leading space delimiters. Parse name delimited by a space. Create a recognizer information token "name" with the three execution tokens. The implementation is system dependent.

The words for XT-INTERPRET, XT-COMPILE and XT-POSTPONE are called with the parsed data that the associated parsing word of the recognizer returned. The information token itself is consumed by the interpreter.

SET-RECOGNIZERS (rec-n .. rec-1 n --) RECOGNIZER

Set the recognizer stack to the recognizers identified by the execution tokens of their parsing words `rec-n .. rec-1`. `rec-1` will be the parsing word of the recognizer that is called first, `rec-n` will be the last one. If n is not a positive number, an ambiguous condition is met. A minimum recognizer stack shall include the words for dealing with the dictionary and integer numbers.

XY.7 Reference Implementation

The code has as little as possible dependencies (basically only CORE). The implementations in gforth and amforth differ and use highly system specific strategies. The code has been tested on gforth 0.7.0.

```
\ create a simple 3 element structure
: RECOGNIZER: ( XT-INTERPRET XT-COMPILE XT-POSTPONE "<spaces>name" -- )
   CREATE SWAP ROT , , ,
;

\ system failure recognizer
:NONAME -13 THROW ; DUP DUP RECOGNIZER: R:FAIL

\ helper words to decode the data structure created by
\ RECOGNIZER: The knowledge they represent is used inside
\ POSTPONE and the text interpreter only.
: _R>POST ( R:FOO -- XT-POSTPONE ) CELL+ CELL+ @ ;
: _R>COMP ( R:FOO -- XT-COMPILE  )       CELL+ @ ;
: _R>INT  ( R:FOO -- XT-INTERPRET)             @ ;

\ contains the recognizer stack data
\ first cell is the current depth.
10 CELLS BUFFER: rec-data
0 rec-data ! \ empty stack

: SET-RECOGNIZERS ( rec-n .. rec-1 n -- )
   DUP rec-data !
   BEGIN
     DUP
   WHILE
     DUP CELLS rec-data +
     ROT SWAP ! 1-
   REPEAT DROP
;

: GET-RECOGNIZERS ( -- rec-n .. rec-1 n )
   rec-data @ rec-data
   BEGIN
     CELL+ OVER
   WHILE
     DUP @ ROT 1- ROT
   REPEAT 2DROP
   rec-data @
;
```

```
: DO-RECOGNIZER   ( addr len -- i*x R:FOO | R:FAIL )
    rec-data @
    BEGIN
      DUP
    WHILE
      DUP CELLS rec-data + @
      2OVER 2>R SWAP 1- >R
      EXECUTE DUP R:FAIL <> IF R> DROP 2R> 2DROP EXIT THEN DROP
      R> 2R> ROT
    REPEAT
    DROP 2DROP R:FAIL
;
```

`POSTPONE` is outside the Forth interpreter:

```
: POSTPONE (  "<spaces>name" -- )
  PARSE-NAME DO-RECOGNIZER
  _R>POST    \ get the XT-POSTPONE from R:FOO
  EXECUTE
; IMMEDIATE
```

A.XY Informal Appendix

A.XY.1 Forth Text Interpreter

The Forth text interpreter turns into a generic tool that is capable to deal with any data type. It maintains `STATE` and calls the data processing methods according to it.

```
: INTERPRET
  BEGIN
    PARSE-NAME ?DUP IF DROP EXIT THEN \ no more words?
    DO-RECOGNIZER ( addr len -- i*x R:FOO | R:FAIL)
    STATE @ IF _R>COMP ELSE _R>INT THEN \ get the right XT from R:*
    EXECUTE \ do the action.
    ?STACK  \ simple housekeeping
  AGAIN
;
```

A.XY.2 Example Recognizer

The first example looks up the dictionary for the word and returns the execution token and the header flags if found. The data processing is the usual interpret/compile action. The Compile actions checks for immediacy and act accordingly. A portable postpone action is not possible. Amforth and gforth do it in a system specific way.

```
\ find-name is close to FIND. amforth specific.
256 BUFFER: find-name-buf
: place ( c-addr1 u c-addr2 ) 2DUP C! CHAR+ SWAP MOVE ;
: find-name ( addr len -- xt +/-1 | 0 )
    find-name-buf place find-name-buf
    FIND DUP 0= IF NIP THEN ;

: immediate? ( flags -- true|false ) 0> ;
```

```
:NONAME ( i*x XT flags -- j*y )  \ INTERPRET
  DROP EXECUTE ;
:NONAME ( XT flags -- )          \ COMPILE
  immediate? IF COMPILE, ELSE EXECUTE THEN ;
:NONAME ( XT flags -- )          \ POSTPONE
  immediate? IF COMPILE, ELSE POSTPONE LITERAL POSTPONE COMPILE, THEN ;
RECOGNIZER: R:WORD

: REC:WORD ( addr len -- XT flags R:WORD | R:FAIL )
   find-name ( addr len -- XT flags | 0 )
   ?DUP IF R:WORD ELSE R:FAIL THEN ;

\ prepend the find recognizer to the recognizer stack
GET-RECOGNIZERS ' REC:FIND SWAP 1+ SET-RECOGNIZERS
```

The second example deals with floating point numbers. The interpret action is a do-nothing since there is nothing that has to be done in addition to what the parsing word already did. The compile action takes the floating point number from the FP stack and compiles it to the dictionary. Postponing numbers is not defined, thus the postpone action here is printing the number and throwing an exception.

```
:NONAME ;                    ( -- ) (F: f -- f) \ INTERPRET
:NONAME POSTPONE FLITERAL ; ( -- ) (F: f -- )  \ COMPILE
:NONAME FS. -48 THROW ;      ( -- ) (F: f -- )  \ POSTPONE
RECOGNIZER: R:FLOAT

: REC:FLOAT ( addr len -- (F: -- f ) R:FLOAT | R:FAIL )
  >FLOAT IF R:FLOAT ELSE R:FAIL THEN ;

\ append the float recognizer to the recognizer stack
' REC:FLOAT GET-RECOGNIZERS 1+ SET-RECOGNIZERS
```

Experience

First ideas to dynamically extend the Forth text interpreter were published in 2005 at comp.lang.forth by Josh Fuller and J Thomas: Additional Recognizers?

A specific solution to deal with number prefixes was roughly sketched by Anton Ertl at comp.lang.forth in 2007 with https://groups.google.com/forum/#!msg/comp.lang.forth/r7Vp3w1xNus/Wre1BaKeCvcJ

There are a number of specific solutions that can at least partly be seen as recognizers in various Forth's:

- prefix-detection in ciforth
- W32Forth uses its "chain" concept to achieve similar effects.
- various commercial Forth's seem to have ways to extent the interpreter.

A first generic recognizer concept was implemented in amforth version 4.3 (May 2011). The design presented in this RFD is implemented with version 5.3 (May 2014). gforth has recognizers since 2012, the ones described here since June 2014.

Existing recognizers cover a wide range of data formats like floating point numbers and strings. Others mimic the back-tick syntax used in many Unix shells to execute OS sub-process. A recognizer is used to implement OO notations.

Most of the small words that constitute a recognizer don't need a name actually since only their execution tokens are used. For the major words a naming convention is suggested: `REC:<name>` for the parsing word of the recognizer "name", and `R:<name>` for the information token word created with `RECOGNIZER:` for the data type "name".

There is no `REC:FAIL` that would be the companion of the system provided `R:FAIL`. It's simply

```
: REC:FAIL ( addr len -- R:FAIL )
  2DROP R:FAIL ;
```

That way, `REC:FAIL` can be seen as the parsing word of the recognizer that is always present as the last one in the recognizer stack and that cannot be deleted.

A Forth system that uses recognizers in the core has words for numbers and dictionary look-ups. These recognizers are useful for other data formats and use cases as well. They shall be named identically as shown in the table:

Name	Stack items	Comment
`R:NUM`	`( -- n R:NUM)`	single cell numbers, based on `>NUMBER`
`R:DNUM`	`( -- d R:DNUM)`	double cell numbers, based on `>NUMBER`
`R:FLOAT`	`( -- f R:FLOAT)`	floating point numbers, based on `>FLOAT`
`R:WORD`	`( -- XT flags R:WORD)`	words from the dictionary, `FIND`

The matching parsing words should be available as

Name	Stack effect
`REC:NUM`	`( addr len -- n R:NUM | d R:DNUM | R:FAIL )`
`REC:FLOAT`	`( addr len -- f R:FLOAT | R:FAIL )`
`REC:WORD`	`( addr len -- XT flags R:WORD | R:FAIL )`

Test cases

The hardest and ultimate test case is to use the interpreter with recognizers enabled. Some parts can be tested separately, however.

```
T{ GET-RECOGNIZERS SET-RECOGNIZERS -> }T
T{ GET-RECOGNIZERS SET-RECOGNIZERS GET-RECOGNIZERS -> GET-RECOGNIZERS }T
T{ s" unknown word" DO-RECOGNIZER  -> R:FAIL }T
```

The system provided recognizers, if available, work as follows:

```
T{ s" 1234"  DO-RECOGNIZER -> 1234  R:NUM   }T
T{ s" 1234." DO-RECOGNIZER -> 1234. R:DNUM  }T
T{ s" 1e3"   DO-RECOGNIZER ->  1e3  R:FLOAT }T
T{ s" DO-RECOGNIZER" DO-RECOGNIZER -> ' DO-RECOGNIZER -1 R:WORD }T
```

Points To Discuss

Provide words to create the system defaults in the tradition of `FORTH` and `ONLY` for word lists?

What is the semantics of postponing data apart from words from the dictionary?

```
POSTPONE 42
POSTPONE "a string"
POSTPONE `a system command`
```

Change history

- 2014-10-03 Version 1