A Tokenizer for Rexx and ooRexx

35th International Rexx Language Symposium
Brisbane, Australia, March 3-6 2024

Josep Maria Blasco
jose.maria.blasco@gmail.com

EPBCN – ESPACIO PSICOANALÍTICO DE BARCELONA
C/ BALMES, 32, 2º 1º — 08007 BARCELONA

March the 4th, 2024
Part I

Introduction: General concepts

- Natural languages and formal languages
- Lexers, tokenizers and parsers
- Clauses, tokens and items
- “Tokenized” programs
- What is a tokenizer good for?
### Natural and formal languages (1/3)

<table>
<thead>
<tr>
<th>Natural languages have...</th>
<th>Programming languages have...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>An alphabet</strong></td>
<td><strong>An alphabet</strong></td>
</tr>
<tr>
<td><strong>Lexical elements:</strong></td>
<td><strong>Lexical elements:</strong></td>
</tr>
<tr>
<td>Words, numerals, acronyms, spaces, punctuation, ...</td>
<td>Identifiers, numbers, strings, comments, whitespace, punctuation, ...</td>
</tr>
<tr>
<td>Natural languages have...</td>
<td>Programming languages have...</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>An <em>alphabet</em></td>
<td>An <em>alphabet</em></td>
</tr>
<tr>
<td><em>Lexical</em> elements</td>
<td><em>Lexical</em> elements</td>
</tr>
<tr>
<td><strong>Syntax rules</strong></td>
<td><strong>Syntax rules</strong></td>
</tr>
<tr>
<td>Non-rigid. Especially</td>
<td>Rigid</td>
</tr>
<tr>
<td>in certain contexts:</td>
<td>Not optional</td>
</tr>
<tr>
<td>Poetry</td>
<td>Can not be bended</td>
</tr>
<tr>
<td>Marketing</td>
<td></td>
</tr>
<tr>
<td>News headlines</td>
<td></td>
</tr>
<tr>
<td>Jokes</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Rigid rules: when we operate with a formal language (logic, physics, chemistry, mathematics, music, programming languages, ...) we want to be completely sure of

- *What* we are saying.
- The *meaning* of what we are saying.

That is, we want to *eliminate the ambiguity* that is inherent to natural languages, by means of clear, unambiguous, definitions of the *syntax* and the *semantics* of the formal language.
An application that reads programs written in a certain programming language and returns the sequence of its lexical elements is called a *lexer* or a *tokenizer*.

⇒ *Beware:* “Token” has two special, different, meanings in the Rexx language.

An application that reads programs written in a certain programming language and returns a representation of its syntax tree is called a *parser*.
A **Rexx clause**: a sequence of whitespace, comments and tokens, ended by a (in many cases implied) semicolon. A *token* may be:

- A **literal string** (including hexadecimal and binary strings).
- A **symbol** (*Chair*, *t.*, *t.i.j*, 25AB, *Soup*, ...).
- A **number** (⇒ a special form of literal string ["-12.3", "4e-2", ...] or symbol [12.34, 5E+12, ...]).
- An **operator character** ("+", "-", "*", ...).
- A **special character** (":", "("", ")", ".", ...).
A desirable property of a lexical analyser is to \textit{return all the components} of a clause, including whitespace and comments, instead of only its \textit{tokens}.

- Our tokenizer will return \textit{all the components} ("items"), not only the tokens.
- This allows to \textit{reconstruct the source program} by collating these items in order.

⇒ Our tokenizer returns \textit{more} than only tokens.
Colloquially, one refers to a program distributed without source as a *tokenized program*. Although this denomination has stuck, it is inexact, since “tokenized” programs are indeed full abstract syntax trees, not a mere sequence of tokens.

⇒ In this presentation, we will use “token” in its proper sense.
What is a tokenizer good for?

- A language processor (i.e., an interpreter or a compiler) has to “understand” a program before running it. To that purpose, it has to first break it into its constituent elements.

- Other purposes: a tokenizer is ideally suited to introduce transformations into the sequence of lexical elements that compose a program [Examples: a prettyprinter, a preprocessor (like RXU, see below)],

- and also to compile data about that sequence [Example: a cross-referencer].
A Tokenizer for Rexx and ooRexx

Part II

Tokenizer features

- The specificity of Rexx
- Simple and full tokenizing
- Tokenizing several dialects
- Experimental support for Unicode
The syntax of Rexx is peculiar in several aspects. One of the main ideas behind its design is to make life easy for users, not for language processor implementers.

Example 1: Rexx has *no reserved words*.

```
while = 4
Do while = 1 To (while) While (while < 7)
   Say while
End while
```

⇒ Parsing may be more difficult than with less peculiar languages.
The syntax of Rexx is peculiar in several aspects.

**Example 2:** The concept of token is *counterintuitive*:

- Whitespace is not a token, but, when significant, it may be an operator.
- Some basic constructs like "**", "+=" or ": :" are not a single token but a sequence of several tokens (and may have whitespace and/or comments in between, not that it is a great idea).

```plaintext
a1 = a2 | /* That was a */ - /* (continued) */
| a3 /* concatenation, after all */
```
The syntax of Rexx is peculiar in several aspects.

**Example 3:** The concept of symbol is highly *unusual*:

- It encompasses *variable symbols* (simple, compound or stems),
- *environment symbols*,
- *constant symbols*,
- and *numbers* (⇒ syntax rules are bent to accommodate signs in numbers with an exponent).

More “classical” languages have *identifiers* and *numbers* (as distinct syntactical constructs), but no *constant symbols* or *environment symbols*. 
The syntax of Rexx is peculiar in several aspects.

**Example 1:** Rexx has *no reserved words*.

**Example 2:** The concept of token is *counterintuitive*.

**Example 3:** The concept of symbol is highly *unusual*.

⇒ Our tokenizer will have to take into account all these peculiarities.
Simple and full tokenizing (1/2)

Simple tokenizing: we want the sequence of tokens and separators exactly as they occur in the source file.

For example, if we tokenize “a += 1”, we want to get:

1. "a" (a variable symbol),
2. " " (whitespace, a blank),
3. "+" (an operator character),
4. ":" (another operator character),
5. " " (another blank), and
6. "1" (an integer number symbol).
[Features] Simple and full tokenizing (2/2)

Full tokenizing: we want that some tokens are combined into higher level constructs, and that non-significant separators are discarded.

Tokenizing “a += 1” once more, but now with full tokenizing, we would get:

1. "a" (a variable symbol, with an indication that this is the start of an [extended] assignment),
2. "+=" (an extended assignment operator),
3. "1" (an integer number symbol).
We want to be able to recognize several variants of Rexx:

- Open Object Rexx (ooRexx)
- Regina Rexx
- ANSI Rexx (implemented by Regina)
- ... (in the future?)

Every dialect has its own, slightly different definitions. For example, whitespace in ooRexx includes only HT as other blank characters, but under Regina we also accept VT and FF.
When activated, we accept **five new string suffixes**.

**Low-level Unicode strings**, "String"U, composed of any number of

- Blank-separated hexadecimal code points (with or without a "U+" or "u+" prefix: "61"U == "a", "u+0061"U == "a", "1F680"U == "🚀", "U+1F680"U == "🚀").

- Parenthesized names, alias or labels ("(Rocket)"U == "🚀", 
  "(End-of-line)"U == "0A"X, "(<Control-000A>)"U == "0A"X).

Names, alias and labels are case-insensitive, and blanks, dashes and underscores are ignored.
Low-level **BYTES** strings, "String"\(^{Y}\), composed of bytes.

BYTES strings are explicitly declared to be equivalent to Classic Rexx strings. The "\(^{Y}\)" suffix is useful when unsuffixed strings have been assigned non-classical semantics.

Options DefaultString Codepoints
/* --> Now a string is a CODEPOINTS string by default */

a = "ئة " /* A CODEPOINTS string, 1 code point */
b = "نة ﺰ "\(^{Y}\) /* A BYTES string, 8 bytes */
CODEPOINTS strings, "String" P, composed of Unicode code points.

"String" has to be valid UTF-8, or a syntax error will be raised.

```plaintext
zoo = "チンパンジー シマウマ " P
Say Length(zoo)    /* 3 (3 code points) */
Say zoo[2]        /*  */
**Features** Experimental support for Unicode (4/5)

**GRAPHEMES** strings, "String"G, composed of Unicode extended grapheme clusters.

"String" has to be valid UTF-8, or a syntax error will be raised.

```plaintext
Options Coercions Promote
glue = "(Zero Width Joiner)"U
family = ".face""glue"face""glue"face""glue"G /* <-- Note the "G" */
Say Length(family) /* 1 (1 grapheme cluster) */
Say family /* face */
```
TEXT strings, "String"T, composed of Unicode extended grapheme clusters automatically normalised to NFC.

jose = "Jose"T
joseacute = jose"301"U /* "301"U is the acute accent */
Say C2X(joseacute[4]) /* 39A9 (not 65CC81) */
Say Reverse(joseacute) /* ésoJ */
A Tokenizer for Rexx and ooRexx

Part III

Using the tokenizer

Using the tokenizer
  Installation
  Choosing the right tokenizer
  Creating a tokenizer instance
  Load the tokenizer constants
  Choosing simple or full tokenizing
  Choosing detailed or undetailed tokenizing

Structure of the returned items
  Returned items are Rexx stems
  Class and subclass
  Location
  Value
  Other attributes
  Error handling
[Usage] Installation

- To use the tokenizer in conjunction with all the Tutor-defined Unicode Rexx features, follow the Tutor installation instructions, and load the Unicode libraries using:

  ::Requires "Unicode.cls"

- If you do not need Unicode features, you can load a standalone version of the tokenizer:

  ::Requires "Rexx.Tokenizer.cls"
Choosing the right tokenizer

Choose the class that represents the tokenizer variant you want to run:

- ooRexx.Tokenizer, for programs written in ooRexx.
- Regina.Tokenizer, for programs written in Regina.
- ANSI.Rexx.Tokenizer, for programs written in ansi Rexx.

If you need Unicode features, choose one of

- ooRexx.Unicode.Tokenizer,
- Regina.Unicode.Tokenizer or
- ANSI.Rexx.Unicode.Tokenizer.
Creating a tokenizer instance

To create a tokenizer instance, you will first need to construct a `Rexx` array containing the source program to tokenize.

```rexx
/* Assume the source program resides in a file */
/* Read the whole file into an array */
source = CharIn(inFile,,Chars(inFile))~makeArray
```

This array will then be passed as an argument to the `new` method of the corresponding tokenizer class, to get an instance of the tokenizer for this particular program source.

```rexx
/* Now create a tokenizer instance */
tokenizer = .ooRexx.Tokenizer~new(source)
/* Or .Regina.Tokenizer, etc. */
```
[Usage] Load the tokenizer constants

You should load the tokenizer symbolic constants contained in the tokenizer `tokenClasses` constant by using the following code fragment:

```
Do constant over tokenizer~tokenClasses
    Call Value constant[1], constant[2]
End
```

This will allow you to identify the token classes and subclasses returned by the tokenizer, like `END_OF_SOURCE`, `SYNTAX_ERROR`, `VAR_SYMBOL` or `ASSIGNMENT_INSTRUCTION`.

All constants have one byte values.
Choosing simple or full tokenizing

Depending on the characteristics of your program, you may want to choose simple tokenizing (using the getSimpleToken method), or full tokenizing (using getFullToken):

```c
/* Two possible reasons to exit the loop */
extit_conditions = END_OF_SOURCE || SYNTAX_ERROR

Do Forever
    item = tokenizer~getSimpleToken /* Or getFullToken */
    /* Exit on error or end of source */
    If Pos(item[class], exit_conditions) > 0 Then Leave
    /* ==> Do things with the itemn */
End
```
[Usage] Choosing detailed or undetailed tokenizing

If you have chosen to use the full tokenizer, you will also have to decide if you want to get *detailed* or *undetailed* results from your `getFullToken` method calls. You can do that when creating your tokenizer instance, by using a second, optional, argument of the `new` class method:

```plaintext
/* A second, boolean and optional, argument of */
/* the 'new' method determines if tokenizing */
/* will be detailed or not. */

tokenizer = .ooRexx.Tokenizer~new(source, .true)
```
[Usage] Returned items are **REXX** stems

The result of a call to `getSimpleToken` (or `getFullToken`) is a **REXX** stem:

```plaintext
token. = tokenizer~getSimpleToken
```

Each stem has a number of predefined indexes (we sometimes call them “properties” or “attributes”), like `token.class`, `token.subclass`, `token.location` and `token.value`. Results of full tokenizing and special tokens like `SYNTAX_ERROR` may have additional properties.
Token.class and token.subclass describe the nature of the returned token. Examples:

- `token.class == VAR_SYMBOL & token.subclass == SIMPLE_VAR`: a variable symbol which is not a stem or a compound symbol.
- `token.class == KEYWORD_INSTRUCTION & token.subclass == CALL_INSTRUCTION`: a Call instruction (full tokenizing only).
- `token.class == BLANK`: whitespace.
- `token.class == STRING & token.subclass == TEXT_STRING`: a TEXT string, specified with the "T" suffix.
Token.location is a string containing four integers separated by blanks which describe the location and extent of the returned token:

"startLine startCol endLine endCol"

The token starts at line startLine, column startCol, and extends until line endLine, column endCol - 1. StartLine and endLine always have the same value, except for multi-line comments and ooRexx resources.
In most cases, `token.value` is the value of the token as it appears in the source program.

Comments and **ooRexx** resources return a placeholder (but you can reconstruct the original token value by resorting to `token.location` and inspecting the source code).

Some few token classes return values which are interpreted. For example, hexadecimal and binary strings are converted to character strings, and Unicode strings are replaced by their UTF-8 representations.
Some few item classes return stems with additional attributes.

As we have seen, SYNTAX_ERROR returns a number of additional attributes to fully describe the error.

Additionally, detailed full tokenizing may return “ignored” (or “absorbed”) tokens in the token.absorbed array (more about that below).
When an error is encountered, tokenizing stops, and a special item is returned. Its class and subclass will be SYNTAX_ERROR, and a number of special attributes will be included, so that the error information is as complete as possible.

```plaintext
item.class = SYNTAX_ERROR
item.subclass = SYNTAX_ERROR
item.location = location of the error in the source file
item.value = main error message

/* Additional attributes, specific to SYNTAX_ERROR */
item.number = the error number, in the format major.minor
item.message = the main error message (same as item.value)
item.secondaryMessage = secondary error message
item.line = line number where the error occurred
```
If you want to print error messages that are identical to the ones printed by `ooRexx`, you can use the following code snippet:

```plaintext
If item.class == SYNTAX_ERROR Then Do
  line = item.line
  Parse Value item.number With major"."minor
  Say
  /* inFile is the input file name, and array contains the source */
  Say Right(line,6) "*-*" array[line]
  Say "Error" major "running" inFile "line" line":" item.message
  Say "Error" major"."minor": " item.secondaryMessage
  /* -major should be returned when a syntax error is encountered */
  Return -major
End
```
Part IV

Testing the tokenizer

The *InspectTokens* program
- The *InspectTokens* program
- Simple tokenizing: an example
- Undetailed full tokenizing: an example
- Detailed full tokenizing: an example
The InspectTokens program

InspectTokens.rex resides in the parser subdirectory.

C:\Unicode>InspectTokens
InspectTokens.rex -- Tokenize and inspect a .rex source file
------------------------------------------------------------

Format:

[rexx] InspectTokens[.rex] [options] [filename]

Options (starred descriptions are the default):

- h,  -help                  Print this information
- d,  -detail,  -detailed    Perform a detailed tokenization (*)
- nd, -nodetail, -nodetailed Perform an undetailed tokenization
- f,  -full                  Use the full tokenizer (*)
- s,  -simple                Use the simple tokenizer
- u,  -unicode               Allow Unicode extensions (*)
- nu, -nounicode             Do not allow Unicode extensions
- o,  -oorexx                Use the Open Object Rexx tokenizer (*)
- r,  -regina                Use the Regina Rexx tokenizer
- a,  -ansi                  Use the ANSI Rexx tokenizer

C:\Unicode>
Simple tokenizing: an example

Assume that test.rex contains a single line, \( i = i + 1 \).
[Testing] Undetailed full tokenizing: an example

C:\Unicode>InspectTokens -full -nodetailed test.rex
  1 [1 1 1 1] END_OF_CLAUSE (BEGIN_OF_SOURCE): ''
  2 [1 1 1 2] ASSIGNMENT_INSTRUCTION (SIMPLE_VAR): 'i' ✋
  3 [1 2 1 5] OPERATOR (ASSIGNMENT_OPERATOR): '=' ✋
  4 [1 5 1 6] VAR_SYMBOL (SIMPLE_VAR): 'i'
  5 [1 6 1 9] OPERATOR (ADDITIVE_OPERATOR): '+' ✋
  6 [1 9 1 10] NUMBER (INTEGER): '1'
  7 [1 10 1 10] END_OF_CLAUSE (END_OF_LINE): ''
Took 0.002000 seconds.

C:\Unicode>

Lines that have changed are marked with a ✋ emoji.
Detailed full tokenizing: an example

```plaintext
C:\Unicode>InspectTokens -full -detailed test.rex
  1 [1 1 1 1] END_OF_CLAUSE (BEGIN_OF_SOURCE): ''
  2 [1 1 1 2] ASSIGNMENT_INSTRUCTION (SIMPLE_VAR): 'i'
  3 [1 2 1 5] OPERATOR (ASSIGNMENT_OPERATOR): '='
     ---> Absorbed:
  1 [1 2 1 3] BLANK: ' '  
  2 [1 3 1 4] OPERATOR: '=='  
  3 [1 4 1 5] BLANK: ' '  
  4 [1 5 1 6] VAR_SYMBOL (SIMPLE_VAR): 'i'
  5 [1 6 1 9] OPERATOR (ADDITIVE_OPERATOR): '+'
  6 [1 9 1 10] NUMBER (INTEGER): '1'
  7 [1 10 1 10] END_OF_CLAUSE (END_OF_LINE): ''
Took 0.002000 seconds.
```

Lines that are new are marked with a 👍 emoji.
A Tokenizer for Rexx and ooRexx

Part V

RXU, the Rexx Preprocessor for Unicode

RXU, the Rexx Preprocessor for Unicode
An example run of RXU
How does the preprocessor work?
An example run of RXU (1/3)

Let us create a test2.rxu file with the following content:

```
  Options DefaultString Text
  var = "_sorted" || "(Lobster)"U
  Say "'var'" is a 'StringType(var) "string of length" Length(var)
```

If we now run the preprocessor against this file, we will get the following output:

```
C:\Unicode>rxu test2
"_sorted" is a TEXT string of length 2

C:\Unicode>
```

This worked as expected!

But *how, and why?*
An example run of RXU (2/3)

Let us now run the preprocessor with the -keep option: this keeps a copy of the generated .rex file (instead of deleting it):

1. Do; !Options = DefaultString Text; Call !Options !Options; Options !Options; End
2. var = (!DS("こんにちは")) || (Bytes("こんにちは"))
3. Say (!DS('') var || (!DS('' is a')) StringType(var) (!DS("string of length"))) !Length(var)
4. ::Requires 'Unicode.cls'

- A line-by-line translation
- A blank line and ::Requires 'Unicode.cls' are added at the end of the translated program.
- The Options instruction gets a complex translation. [..../..]
Do; !Options = DefaultString Text; Call !Options !Options; Options !Options; End

var = (!DS("וס")) || (Bytes("וס"))
Say (!DS('"'))var || (!DS('" is a' )) StringType(var) (!DS("string of length")) !Length(var)

::Requires 'Unicode.cls'

- Unsuffixied "string" ⇒ !(DS("string")). !DS implements Options DefaultString.
- "(Lobster)"U ⇒ (Bytes("וס"))
- New built-in functions, like StringType(), appear as-is.
- Existing built-in functions, like Length(), have a "!" character prepended to their name.
Example 1: Translating `Length()`.

- We should translate function and procedure calls *only*, including `Call Length`

  instructions, but not variable names, method calls or internal routines.

- We can do (most of) that with only a few symbols of context.

- *(But we can not handle internal routines called `Length`).*
Example 2: Translating strings [1/2]. An unsuffixed string "string" gets translated to !(DS("string)). When an Options DefaultString instruction is found, the setting is stored in .local~Unicode.DefaultString (default is "TEXT").

::Routine !DS Public
    Use Strict Arg string
    Select Case Upper(.Unicode.DefaultString)
    When "BYTES" Then Return Bytes(string)
    When "CODEPOINTS" Then Return Codepoints(string)
    When "GRAPHEMES" Then Return Graphemes(string)
    When "TEXT" Then Return Text(string)
    Otherwise Return String
    End
Example 2: Translating strings [2/2]. P, G, and T strings have to be checked for UTF-8 well-formedness, and T strings have to be additionally normalised to NFC, if needed.

The translation of a Unicode \( \text{U} \) string has to be enclosed in a call to \texttt{Bytes()}\texttt{()}, but \textit{only in certain contexts}:

\[
\text{"(Duck)"U: Say "(Duck)"U}
\]

\[
/\ast \text{If we translate to} /\ast
\]
\[
\text{BYTES(\"\&\text{uml}\")}: \text{Say BYTES(\"\&\text{uml}\")} /\ast \text{--&gt; Syntax error} /\ast
\]
\[
/\ast \text{We should instead translate to} /\ast
\]
\[
\text{\"\&\text{uml}\": Say BYTES(\"\&\text{uml}\")} /\ast \text{OK} /\ast
\]
A Tokenizer for Rexx and ooRexx

Part VI

Conclusions

Conclusions

Further work
Acknowledgements
Resources
Questions?
[Conclusions] Further work

▶ Evolve the tokenizer into a full abstract syntax tree parser.
▶ Improve rxu, the Rexx preprocessor for Unicode, to take advantage of the tokenizer enhancements (for example, calls to internal functions with the same name as built-in functions will not be translated).
▶ Explore the development of new tools, like a cross-referencer for Rexx and ooRexx.
▶ Possibility of new, most probably more powerful, language extensions.
▶ ...

...
Tutor, and the Rexx tokenizer, could not have been developed without the intense debates, general creativity and overwhelming feedback of the RexxLA Architecture Review Board (ARB), for which I am deeply indebted.

I also want to extend my gratitude to Laura Blanco, Mireia Monforte, David Palau and Amalia Prats, students of my Psychoanalysis and Logic course at EPBCN, where I also teach some Rexx programming, for their persistence, unwavering interest, and candid feedback.

Finally, I have to thank my colleagues at EPBCN, for being loving, caring and supportive, and for bearing with me during the long periods where I immersed myself in Rexx matters, disappearing from the common world. Special thanks should go to Silvina Fernández and Olga Palomino, who have attended several essay sessions. Silvina Fernández has also taken care to operate our ElGato Stream Deck during my talks.
[Conclusions] Resources


[Conclusions] Questions?

Thank you!

Questions?