ruby -e 'p gets' with japanese charactors gets additional invalid leading chars and caught Encoding::InvalidByteSequenceError

02/16/2022 03:52 PM - YO4 (Yoshinao Muramatsu)

Status: Closed
Priority: Normal
Assignee:
Target version:
ruby -v:

Description
Input a line starting with japanese charactor from console, almost every time ruby gets additional invalid leading characters.

Reproduce process
R:\ruby32\bin>ruby -e 'p gets'

"\xA0" on Windows-31J (Encoding::InvalidByteSequenceError)
from -e:1:in `gets'
from -e:1:in `<main>'

expected result
R:\ruby32\bin>ruby -e 'p gets'

"あ"

your ruby version (ruby -v)
R:\ruby32\bin>ruby -v
ruby 3.2.0dev (2022-02-16T08:57:04Z master 00c7a0d491) [x64-mswin64_140]
R:\ruby32\bin>ver
Microsoft Windows [Version 10.0.19043.1526]

other observations
environment
- On command prompt window with Legacy Console mode, this issue NOT occurs.
- On Windows Terminal, this issue occurs.
- On Windows Sandbox(Japanese Locale), this issue occurs.
- RubyInstaller binaries has same issue

C:\src\git>ruby -v
ruby 3.1.0p0 (2021-12-25 revision fb4df44d16) [x64-mingw-ucrt]
C:\src\git>ruby -Eutf-8 -e 'p gets'

"\xA0" on Windows-31J (Encoding::InvalidByteSequenceError)
from -e:1:in `gets'
from -e:1:in `<main>'

A line starting with single byte charactor(s) got valid value.

R:\ruby32\bin>ruby -e 'p gets'
external encoding affects

- with Windows-31J, second enter key for line input.

R:\ruby32\bin>ruby -EWindows-31J -e 'p gets'
"\xA0\xFF\n" # <= Second enter key required
"\xA0\xFF\n" # <= \xA0\xFF is additional chars

character variations

R:\ruby32\bin>ruby -EWindows-31J -e 'p gets.b'
" # <= \x{82A0}
"\xA0\xFF\x82\xA0\n"

R:\ruby32\bin>ruby -EWindows-31J -e 'p gets.b'
" # <= \x{8140} fullwidth space
"@\x00\x81@\n"

R:\ruby32\bin>ruby -EWindows-31J -e 'p gets.b'
" # <= \x{8141}
"A\x00\x81A\n"

R:\ruby32\bin>ruby -EWindows-31J -e 'p gets.b'
" # <= \x{8142}
"B\x00\x81B\n"

sysread got valid value.

R:\ruby32\bin>ruby -e 'p STDIN.sysread(1024).force_encoding(Encoding::Windows_31J)'
"\xA0\xFF\r\n" # <= valid

STDIN.binmode can not resolve this.

R:\ruby32\bin>ruby -e 'STDIN.binmode; p gets.force_encoding(Encoding::Windows_31J)'
" # <= Second enter key required
"\xA0\xFF\x(82A0)\r\n" # <= invalid

Ruby 3.0 and earlier versions has a different behavior. especially sysread returns invalid.

C:\src\git>ruby -v
ruby 3.0.3p157 (2021-11-24 revision 3fb7d2cadc) [x64-mingw32]

C:\src\git>ruby -Eutf-8 -e 'p gets'
" # <= Second enter key required
"\xA0\xFF\x82\xA0\n" # <= exception not occurs but invalid value
C:\src\git>ruby -EWindows-31J -e 'p gets'
" # <= Second enter key required
"\xA0\xFF\n" # <= also invalid value
C:\src\git>ruby -e 'p STDIN.sysread(1024).force_encoding(Encoding::Windows_31J)'
" \xA0\xFF\x{82A0}\

conclusion
1. ruby 3.1/3.2dev gets return invalid vs sysread return valid
2. ruby 3.1/3.2dev sysread return valid vs 3.0 sysread return invalid
3. The fact that it works fine in legacy console suggests that windows has some issue, but from the previous it looks like ruby can handle it.

**Associated revisions**

**Revision 5d90c601 - 03/16/2022 01:28 AM - YO4 (Yoshinao Muramatsu)**

Avoid console input behavior in windows 10 [Bug #18588]

When ANSI versions of PeekConsoleInput read multibyte character partially, subsequent ReadFile returns wrong data on newer Windows 10 versions (probably since Windows Terminal introduced). To avoid this, use Unicode version of PeekConsoleInput/ReadConsole.

**History**

**#1 - 03/09/2022 02:01 PM - YO4 (Yoshinao Muramatsu)**

It seems to ANSI version of PeekConsoleInput read multibyte character partially, subsequent ReadFile returns wrong data on newer Windows 10 versions.

I reported this to microsoft/terminal ([https://github.com/microsoft/terminal/issues/12626](https://github.com/microsoft/terminal/issues/12626))

To avoid this behavior, we can use Unicode version of PeekConsoleInput/ReadConsole.

PR: [https://github.com/ruby/ruby/pull/5634](https://github.com/ruby/ruby/pull/5634).

**#2 - 03/16/2022 01:29 AM - YO4 (Yoshinao Muramatsu)**

- Status changed from Open to Closed

Applied in changeset [git|5d90c6010999ac11d25822f130b29d37f81755](git|5d90c6010999ac11d25822f130b29d37f81755).

Avoid console input behavior in windows 10 [Bug #18588]

When ANSI versions of PeekConsoleInput read multibyte character partially, subsequent ReadFile returns wrong data on newer Windows 10 versions (probably since Windows Terminal introduced). To avoid this, use Unicode version of PeekConsoleInput/ReadConsole.