ISO/IEC JTC 1/SC 2/WG 2/IRG N1860

parseIDS.html: JavaScript utility for IDS.txt

suzuki toshiya
Hiroshima University

What is “parseIDS.html”?

- A short JavaScript utility to search Hanzis from ids.txt (a collection of IDS for CJK Unified Ideographs)

Why “grep” is insufficient?

- ids.txt is a collection of the most composed expressions
  - something like NFC of Unicode
  - characters including “林” cannot be found by searching “木”
- recursive searching is required
  - The maintainer (Kawabata-san) uses Emacs + LISP
  - They are slightly exotic software on some platforms

NFC-like IDS vs NFD-like IDS

- When one searches something, an IDS of un-coded Hanzi could be wanted
  - ➢NFC-like IDS cannot include un-coded character
    - except of some CDP compatible glyphs
- Searching with NFD-like IDS would be generic and portable

parseIDS.html uses NFC-like IDS

- For some Hanzis, ids.txt provides multiple expressions (e.g. xxx[GT] yyy[J] zzz[K])

- Making NFD-like IDS for all Hanzis will generate huge collection of possible expressions
  - numVariants(component1) x numVariants(component2) x ....

- parseIDS.html is expected to be small utility

Too Many Variation Example

- ids.txt defines 6 variants for 𠂉(U+4E87)

  ⿱⿰𠂉丿[GK]　⿱⿰𠂉𠂉丿[T]　⿰⿰丿丁

  ⿰⿰丿丁　⿱⿰𠂉丿　⿱⿰𠂉𠂉

- 竹(U+7AF9) will have 6x6 = 36 variants
- ⿱𥫱竹(U+25D12) will have...

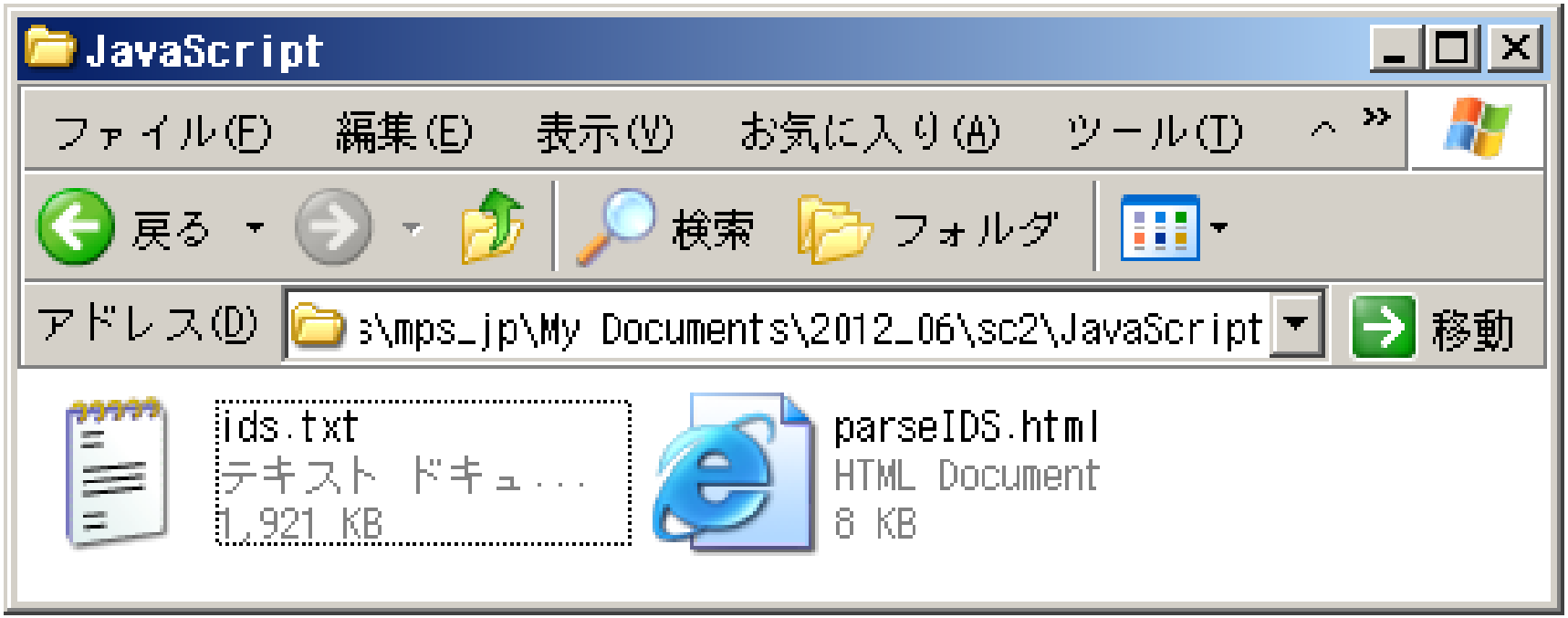
  36^3 = 46,656 variants
- ⿱竹⿱竹竹(U+25DF9) will have...

  36^4 = 1,679,161 variants

  (more than the number of UCS Hanzi!)

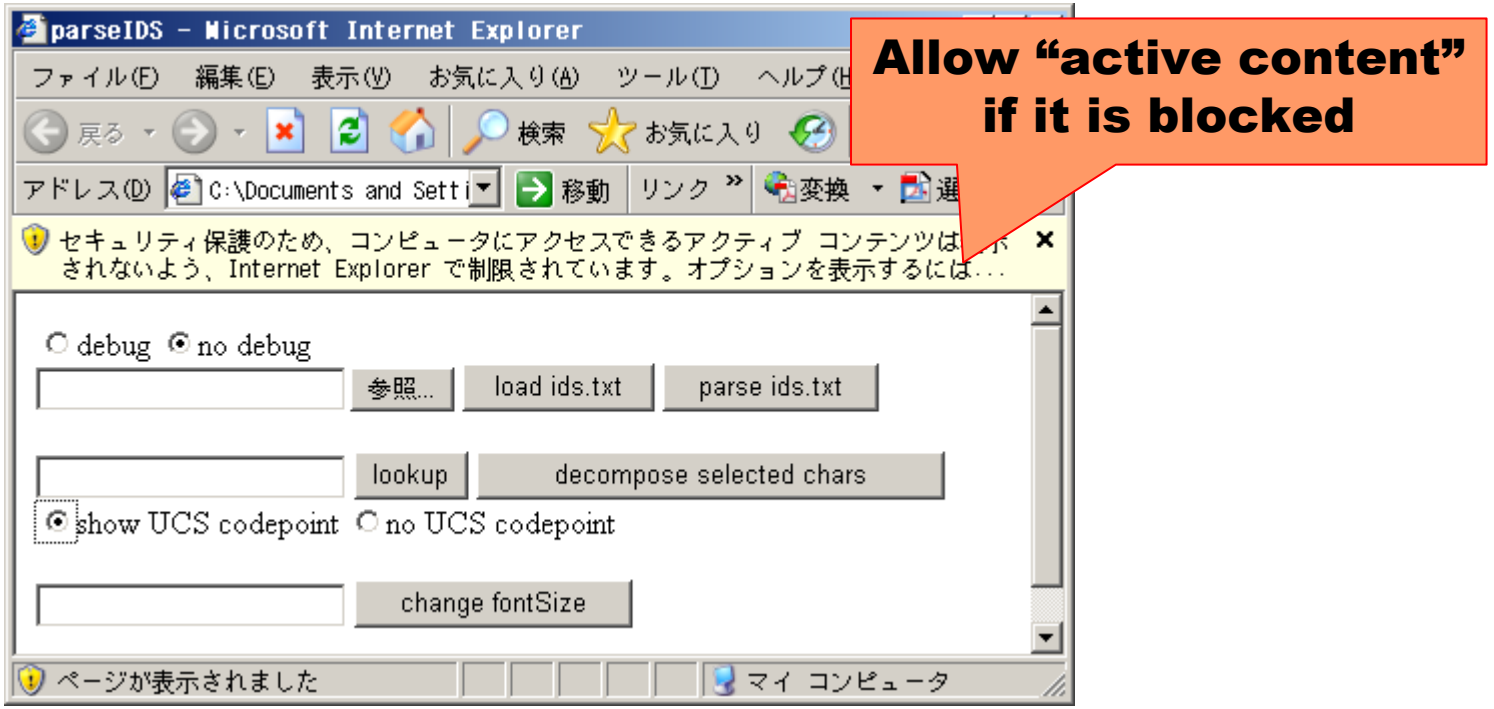
How to use? (installation)

- Place “ids.txt” and “parseIDS.html” to same folder



How to use? (startup)

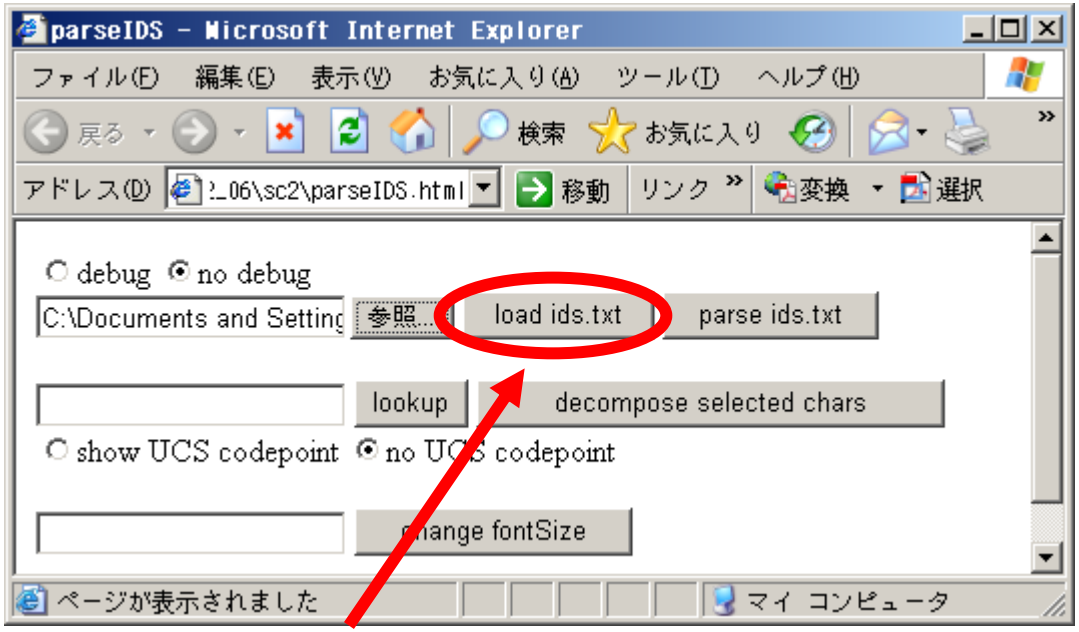
- Open “parseIDS.html” with your web browser



Because ActiveXObject(‘Microsoft.XMLHTTP’) is invoked when executed by MSIE, the security warning will be issued.

How to use? (initialization(1))

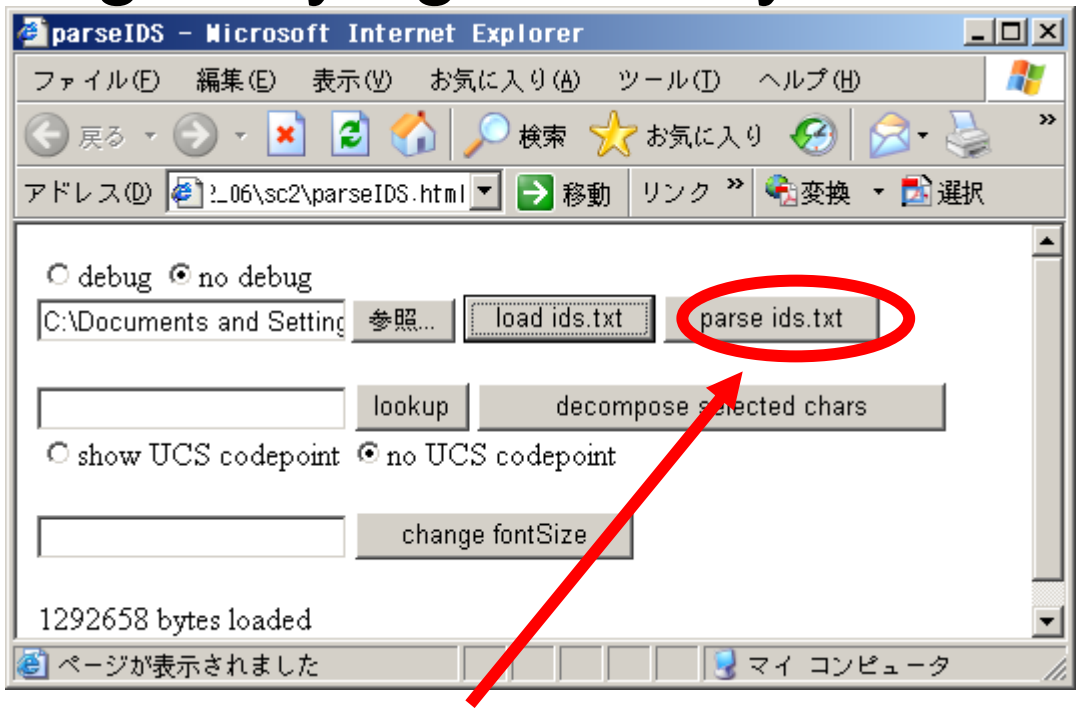
- Select “ids.txt” from the folder where parseIDS.html is placed.
  - The web browsers with HTML5 FileReader API (e.g. Firefox) can handle “ids.txt” in different folders.



- Push “load” button.

How to use? (initialization(2))

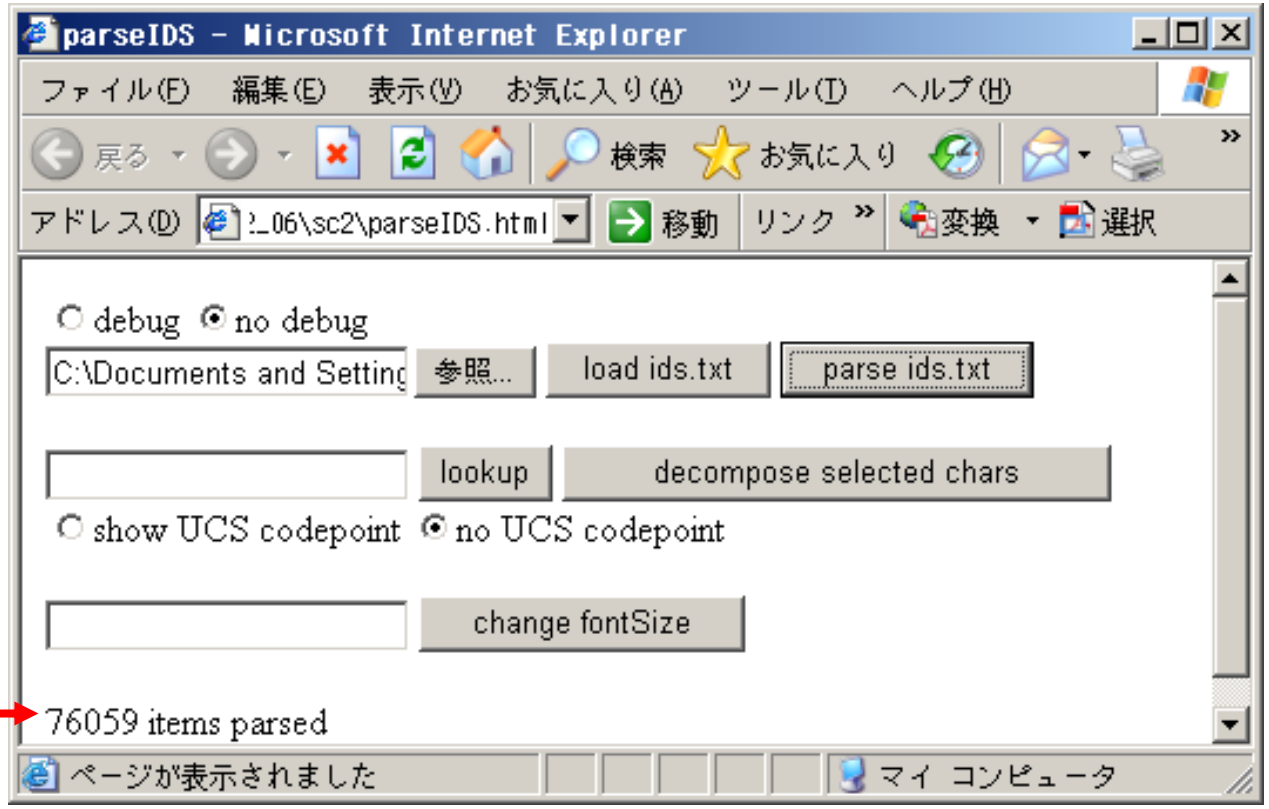
- If loading finishes successfully, you will get a message saying “xxxx bytes loaded”.



- Push “parse” button.

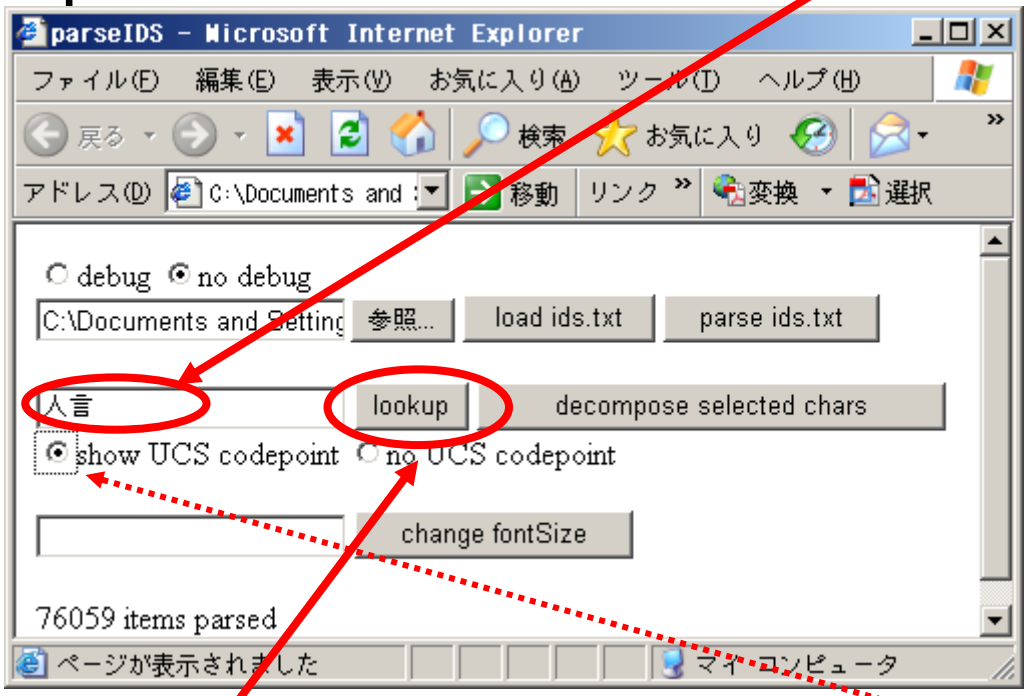
How to use? (initialization(3))

- If parsing finishes successfully, you will get a message saying “xxxx items parsed”.



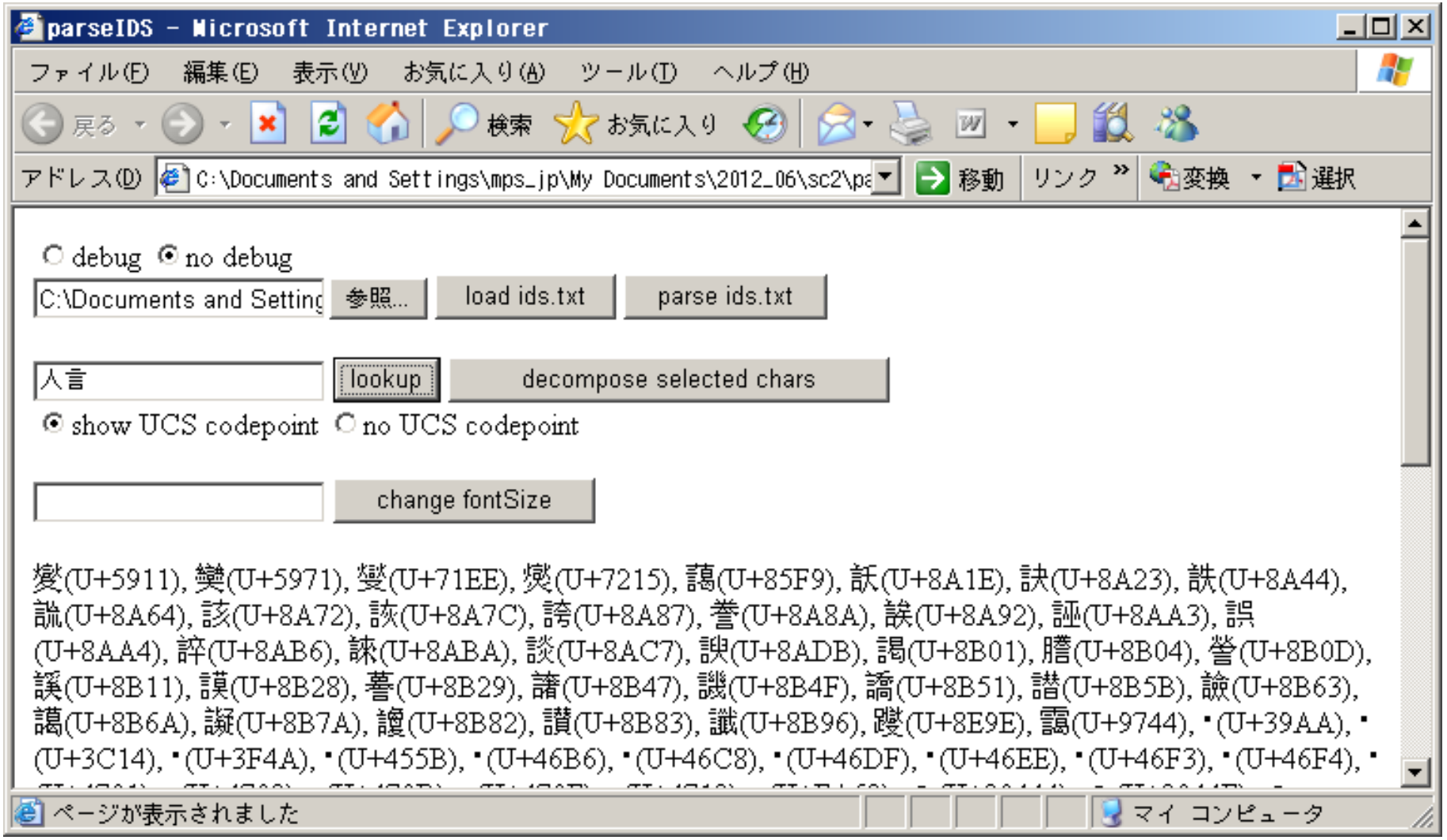
How to use? (search)

- Enter some Hanzis to the form at the left of “lookup” button.



- Push “lookup” button.
  - If you want UCS codepoint too, check “show UCS”

Search Result

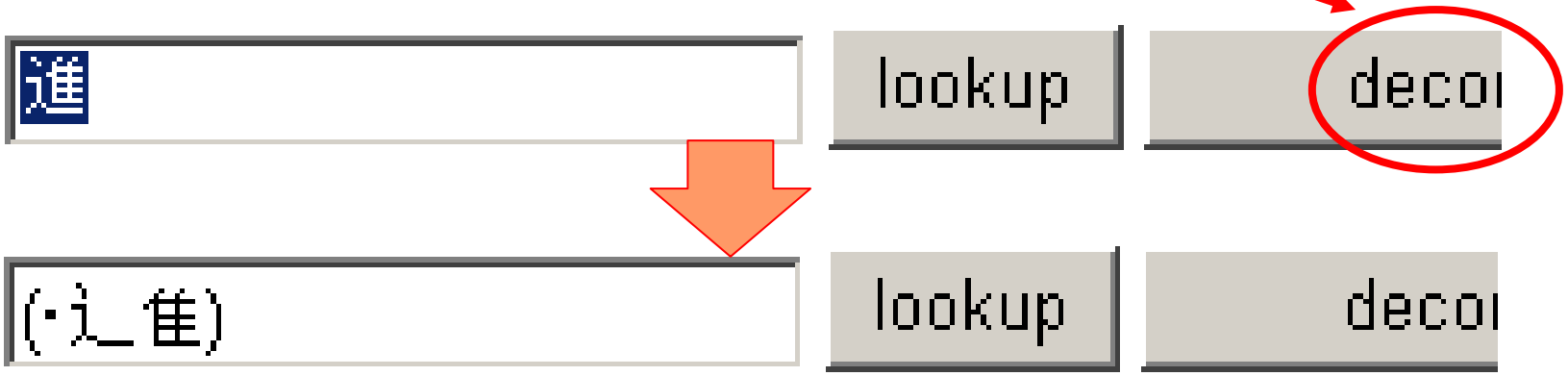


The result is hard to understand?

- You will get the list of Hanzis that include cover all given Hanzis as components.
- The result for “田丁” will include “町”, “甿”, “畸”, etc.
  - Different from the result of “grep 田丁 ids.txt”

How to use “decomposition”

- Some Hanzi input methods are not easy to input the radicals.
  - e.g. MS-IME for Japan: “shinnyou” → “之繞” (not “辶”)
- “Decompose” button replaces the selected Hanzis in “lookup” form
  - After the decomposition, you can remove unrequired components.



Any comments?

feature requests, bugs, etc

→ mpsuzuki@hiroshima-u.ac.jp

appreciations

→ kanji-database-contact@lists.sourceforge.net