The format of the sentence.dat file used for automated localization of encoded localizable sentences

William J G Overington

19 February 2014

In a publication of 17 February 2014, referenced at the end of this publication, the format of a file named sentence.dat is explained.

This research document presents an extension to that format that is considered desirable in practice and also presents some notes about practical production and application of a sentence.dat file.

Section 4 of the publication of 17 February 2014 is as follows.

quote

4. Automated localization

This is a thought experiment at present.

Automated localization would be by having a file sentence.dat available. In the thought experiment the file is a UTF-16 text file, such as can be saved from the WordPad program by selecting saving as a Unicode Text Document.

The sentence.dat file could either be the definitive standardization file, which would use English with en-gb-oed spelling, or could be a copy of that file that has been translated into some other language, with due consideration for localization issues.

The sentence.dat file would consist of a number of lines of text.

A valid line of text would have one of three possible formats.

If the first character of the line is an asterisk, then the line is a comment.

If the first character of the line is an EQUALS SIGN then the line is a heading for a cascading menu for semi-automated message construction; and also the rest of the line is intended to be a localization line as below.

Otherwise the line is intended to be a localization line, yet only is a localization line if it is of the correct structure.

The correct structure for a localization line is as follows.

One or more characters that are not the VERTICAL LINE character.
A VERTICAL LINE character.
Zero or more characters that are not the VERTICAL LINE character.

For example, as follows.

::70000:;|Is there any information about the following person please?

Sentences do not need to be in numerical order, though that would usually help a human reading the file. However, heading sentences for a cascading menu, which could also be used in a message if desired, may well be in a particular part of the code number range.

end quote

While experimenting with practical implementation of a sentence.dat file, the possibility was considered that on some software platforms that there might be complications, while reading characters from the sentence.dat file, regarding detecting the end of the sentence.dat file.

So the following is added to the format of the sentence.dat file.

If the first character of the line is a PERCENT SIGN then the line is the last line of the file.

In a sentence.dat file produced as a Unicode Text Document saved from the WordPad program, lines are separated by two characters, namely CARRIAGE RETURN and LINE FEED, in that order. That is, pressing the return key on the keyboard produces two characters in a Unicode Text Document saved from the WordPad program.

The final five characters of the sentence.dat file are here specified to be as follows.

CARRIAGE RETURN
LINE FEED
PERCENT SIGN
CARRIAGE RETURN
LINE FEED

This is achieved using WordPad by pressing the return key both before and after the PERCENT SIGN has been entered.

It is noted that a Unicode Text Document saved from the WordPad program stores the two bytes of each character with the lower byte before the higher byte.

It is noted that a Unicode Text Document saved from the WordPad program starts with a U+FEFF character, used as a BYTE ORDER MARK. Thus the first two bytes of a sentence.dat file do not represent a character used in the automated localization process.

It is noted that for English and for some other languages that a Unicode Text Document saved from the WordPad program has many bytes that have a value of zero. However, the use of a Unicode Text Document saved from the WordPad program is

deliberately chosen for this system so as to make participation in producing a localized version of a sentence.dat file as straightforward as possible, and with the hope that software developed for automated localization of this system of localizable sentences will work for all languages that can be represented using Unicode characters.

Here is the content of a sentence.dat file produced to use the sentences for which code numbers were stated in the publication of 17 February 2014.

```
*sentence.dat
*Test version 2014-02-19
*English en-gb-oed
=::99100:;|General discussion.
::10000:;|The following version of the localizable sentence system is being used.
::10016:;|The following question has been asked.
::10017:;|My answer is as follows.
=::99700:;|Enquiries after a disaster.
::70000:;|Is there any information about the following person please?
::70224:;|The person is safe.
%
```

The file may be localized into another language, preferably by a native speaker of that language, and a Unicode Text Document saved from the WordPad program, and the file published, keeping the file name as sentence.dat as the idea is that software developed for automated localization of this system of localizable sentences will hopefully work successfully with whatever version, in whatever language, of the sentence.dat file with which it is supplied at any particular time.

Reference:

Overington, William J G
Communication through the language barrier in some particular circumstances by means of encoded localizable sentences
Published by the author on 17 February 2014.
Deposited at the British Library in a pdf file named by copying the title of the publication and replacing each space character with a LOW LINE spacing underscore character, and then adding .pdf at the end to produce the file name.