The Format of the sentence.dat files for use in Research on Communication through the Language Barrier using encoded Localizable Sentences

William J. G. Overington

Saturday 4 April 2020

This is mostly a thought experiment at present.

Automated localization would be by having a file sentence.dat available. In this 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 four 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.

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

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.

313592|The person is female.

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.

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, which is why the format includes the facility to specify that 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.

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.

This very simple example of the transcript of the text content of a sentence.dat file is not practical for use other than for concept proving, it is just included here so as to provide an indication of the structure of a sentence.dat file. A sentence.dat file for practical use may consist of several hundred lines of text. A version for localization to, say, French, would have the code numbers unchanged, but the equivalent text in French.

```
*sentence.dat
*Test version 2020-04-04 Saturday
*English en-gb-oed
128|The following question has been asked.
129|My answer is as follows.
313125|Is there any information about the following person
please?
313987|The person is safe.
%
```