Communication through the language barrier in some particular circumstances by means of encoded localizable sentences

William J G Overington

17 February 2014

This research document presents a system which it is hoped will assist communication through the language barrier in some particular circumstances.

This is by means of encoded localizable sentences.

1. An introductory example

As an introductory example, suppose that a gentleman who speaks English but not French wishes to communicate with a lady who speaks French but not English.

Suppose that the gentleman wishes to send a message in the situation of seeking information about a relative or friend after a disaster.

In a manual system he can look through a list and find that the following sentence is within the list of available sentences.

Is there any information about the following person please?

The code for that sentence is as follows, namely COLON COLON SEVEN ZERO ZERO ZERO ZERO COLON SEMICOLON.

::70000:;

The gentleman wishes to send the following message.

Is there any information about the following person please?
Margaret Gattenford

So he sends an email including the following text.

::70000:;
Margaret Gattenford

The lady receiving the email could look up the code ::70000:; in a list, a list in the French language, yet she could have an automated system and automated localization switched on, and so the lady would not then view on the screen of her computer the sequence ::70000:; with the name Margaret Gattenford below it: she would instead view a sentence in French with the name Margaret Gattenford below it.

The lady then decides to reply.

The message that she chooses to send, constructed from the list of sentences in the system, would localize into English as follows.

The following question has been asked.
Is there any information about the following person please?
Margaret Gattenford
My answer is as follows.
The person is safe.

If the lady has an automated system she could construct the reply using a cascading menu system, so she need not view the code numbers at all. The message would appear in French upon the screen of her computer.

The gentleman receives the following email.

::10016:;
::70000:;
Margaret Gattenford
::10017:;
::70224:;

The gentleman does not have an automated decoding system yet is able to decode the message manually using a list, a list in English.

2. Plain text and markup and markup bubbles

Traditionally, some text files are plain text and some text files are markup text, with a typical situation that when a text file is markup text that the whole file is structured to a markup format. For example, HTML files and XML files.

This system is different in that the markup is contained as markup bubbles within a plain text file, each markup bubble containing the code for one localizable sentence. Thus a text file for message communication with this system is a plain text file which includes within it one or more markup bubbles.

A localizable sentence markup bubble has the following structure.

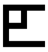
A localizable sentence markup bubble opening bracket, then a number, typically five, of digit characters, then a localizable sentence markup bubble closing bracket.

During the research, special symbols have been designed for a localizable sentence markup bubble opening bracket and for a localizable sentence markup bubble closing bracket; however, as the special symbols are not encoded as characters in an official standard, for practical purposes of interoperability when sending messages from one computer to another computer, the localizable sentence markup bubble opening bracket is represented by the two character sequence COLON COLON and the localizable sentence markup bubble closing bracket is represented by the two character sequence COLON SEMICOLON as it is thought unlikely that those two sequences would occur within plain text situations. Please note that the two character sequence SEMICOLON SEMICOLON was not used, partly because it could possibly occur in some source code written in a computer programming language, such as, for example, Pascal, though SEMICOLON SEMICOLON was not used because, in a Pascal-style way of design thinking, COLON SEMICOLON seemed to express the intended meaning in a better way.

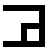
The digits used are the ordinary digits.

Digits are used so that the characters inside the localizable sentence markup bubble are not from the Latin alphabet, so as to make the system as language neutral and script neutral as possible within the limits of being able to achieve a workable system.

For completeness, here are illustrations of the special symbols for a localizable sentence markup bubble opening bracket and for a localizable sentence markup bubble closing bracket. In order to display these symbols in this document a research font with the symbols encoded at Private Use Area code points has been used.



A localizable sentence markup bubble opening bracket.



A localizable sentence markup bubble closing bracket.

3. Limitations of the system

It is perhaps important, for the avoidance of doubt, to state what this system does not do.

This system is not machine translation from one language to another.

This system does not seek to encode all possible sentences.

This system is limited in its uses, though for some uses it may well be of great practical usefulness and for some other uses it may be of some cultural usefulness.

The communication does not flow in the same way as natural conversation. This is so as to minimize the number of sentences that need encoding, to avoid complicated grammatical structures and so as to avoid grammatical carry-forward: for example, if it is desired to mention a flower, describing the colour of the flower, one uses two localizable sentences.

There is a flower.
The colour is yellow.

This is because using one sentence with both the word yellow and the word flower in the same sentence would need either more sentences in order to cover other colours of flower, or a more complicated system to use an adjective as a parameter.

Also using two sentences with the second sentence starting with the word It in the English standardization document could lead to problems because in some other languages, for some nouns, the noun in the previous sentence could be masculine, and for some nouns, the noun in the previous sentence could be feminine: and in some other languages, for some nouns, the noun in the previous sentence could be masculine, and for some nouns, the noun in the previous sentence could be feminine, and for some nouns, the noun in the previous sentence could be neuter.

However, making the adjective describe the colour means that the gender of the adjective agrees with the gender of the word colour, thus resolving the problem.

The purpose of the system is to assist communication through the language barrier, not to be able to translate any text from one language to another language.

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.

5. Symbols

The research upon which this document is based has gradually developed over many years, mostly, yet not entirely, as thought experiments.

At an earlier stage of the research the idea was to encode each localizable sentence as a character, each localizable sentence having its own code point in the character map. There are also symbols for display for some sentences, that is, a distinct symbol for each localizable sentence. Thus if automated localization were not available, the symbol could be displayed using a special font. Various problems arose and later the present system was devised so as to avoid those problems yet still produce a useful system.

However, it is still possible to use the symbol glyphs as they can, if so desired, be included in an OpenType font as unmapped glyphs and accessed by OpenType glyph substitution using an OpenType dlig (discretionary ligatures) table by substituting a nine character localizable sentence code with a glyph of the symbol. Font-based OpenType substitution would only occur after non-font-based automated localization of localizable sentences, so if automated localization took place and all of the localizable sentence codes used in the email message were in the sentence.dat file, then there would be no nine character localizable sentence code or codes still available in the incoming message.

6. Various versions of the sentence.dat file

Most standards are such that a later version, a higher numbered version, replaces an earlier version.

A notable exception is in relation to standardization of PDF/A documents, where lower numbered versions continue in coexistence with higher numbered versions.

In relation to localizable sentences, it is a current research topic as to how various versions would coexist so as to avoid confusion.

The intention is that for sentence.dat files that lower numbered versions continue in coexistence with higher numbered versions and also that a higher numbered version need not necessarily contain all of the localizable sentences in a lower numbered version. This need not necessarily lead to confusion as some versions could be comprehensive for general communication through the language barrier, while others could be for specialist purposes and there is no need for a sentence.dat file for one specialist purpose to include the sentences specific to another specialist purpose. Thus a version number can here be regarded as an index number to identify a particular version within an indexed list of versions.

This can be facilitated by including in all versions the following localizable sentence that states which version is being used.

The following version of the localizable sentence system is being used.

Here is an example use to indicate that version 2 is being used.

```
::10000:;
2
```

Please note that the above example use is only to show how the localizable sentence would be used. At the time of writing this document, no versions have been defined.

7. Conclusions

The present system using markup bubbles represented using colon, semicolon and digit characters is a system that could potentially be useful in some particular circumstances.

A sentence.dat file in English could be produced and a sentence.dat file in another language could be produced by translation of that file.

Various permanent versions of the sentence.dat file could be produced so as to assist communication through the language barrier in various scenarios.