A preliminary proposal to encode two base characters

William J G Overington

19 October 2015

1. Introduction

This document is a preliminary proposal to encode two characters into Unicode.

The two characters are each intended to be used as a base character for an encoding space produced by using a sequence consisting of a Unicode base character followed by some Unicode tag characters.

I am hoping that each of the two base characters will eventually become encoded into plane 14, though I am not asking the Unicode Technical Committee (hereinafter, UTC) to encode the two characters straight away.

The reason for not asking the UTC to encode the two characters straight away is that in order to be applied each of the two characters needs to be followed by a sequence of tag characters: the encoding of the sequence of tag characters for each of the two base characters to be by applying information from an ISO standard, the ISO standard for each of the two base characters not being the same ISO standard: the two ISO standards do not exist at this time.

The two supposed ISO standards need not refer to electronic communication in any way. One would be a plain paper list of words and code numbers. The other would be a plain paper list of preset sentences and code numbers. The two plain paper lists could in principle be applied in various ways, though my motivation in seeking for them to become produced is so that they could be applied in plain text electronic communication.

A good result from discussion of this document by the UTC, from my viewpoint, would be for the UTC to agree to keep the matter in escrow so that if I can persuade ISO to encode a plain paper list of words and code numbers and a plain paper list of preset sentences and code numbers, then the UTC would at that time encode the two base characters into plane 14 so that the two plain paper lists could each be applied to produce a tagspace accessed by a plane 14 character.

A tagspace is defined in section 2 of this document, on page 2 of this document.

If the UTC were to decide that, or something approaching that, I would then be able to approach ISO with first draft proposals for the two plain paper lists, not comprehensive, more placeholder applications so as to try to get things started, saying to ISO that encoding into Unicode of the two base characters would become possible. The two placeholder plain paper first drafts could then be gradually altered, maybe completely altered, and augmented so as to be able to get things started, just like Unicode started small and has been extended over time.

I opine that encoding uniquely into plain text is important if these two tagspaces are to become used to maximum capability. I have considered markup systems yet, in my opinion, these do not have the maximum flexibility and opportunity for widespread use as do tagspaces encoded directly into Unicode and the International Standard.

2. The concept of a tagspace defined

In this document, a tagspace is defined as follows.

tagspace: An encoding space produced by using a sequence consisting of a Unicode base character followed by one or more Unicode tag characters.

This document suggests the encoding of two Unicode base characters, each of which would be used to produce a tagspace. The two tagspaces thus produced would be separate and would each stand alone, yet could be used together in some circumstances.

3. The suggested PANLEX BASE CHARACTER character

The first base character for a tagspace is

PANLEX BASE CHARACTER

and the tagspace would encode in a language-independent form each word that exists in a language.

The format for encoding in the tagspace is not suggested in this document.

However, in an explanation of the concept a sequence of tag digit characters has been used, sometimes a TAG DIGIT SEVEN character followed by seven tag digit characters, sometimes a TAG DIGIT ONE character followed by one tag digit character, sometimes a TAG DIGIT TWO character followed by two tag digit characters, and so on.

Here is a suggested visible glyph for the character.

The glyph would not typically be displayed in a software-assisted application scenario yet is included for completeness as it could be useful for a graceful fallback display in some circumstances.

In relation to PanLex words, here is an example, the code number chosen is just for this example, it is not intended to go through to encoding.

For example, consider the word apricot.

Suppose that that is encoded as the number 72519430 in the PanLex listing.

Here the 7 is because there are seven digits following the 7, the 0 is because it is the general word, and the 251943 is just chosen as a number to distinguish the word apricot from other words.

The 251943 is just an example number chosen for this explanation.

The use of the leading 7 is because more frequently used words could have lower numbers.

For example, the word and could have the number, say, 12 and the word today could have the number, say, 245 and so on.

Returning to apricot.

There could be variations so as to disambiguate.

For example,

72519430 apricot

72519431 apricot (fruit)

72519432 apricot (tree)

72519433 apricot (colour)

72519434 apricot (flavour)

Suppose that the PANLEX BASE CHARACTER were to become encoded into Unicode.

Then an apricot in the sense of an apricot tree would, just here continuing with the code number of this explanation, be encoded into plain text using the following nine characters.

PANLEX BASE CHARACTER, TAG DIGIT SEVEN, TAG DIGIT TWO, TAG DIGIT FIVE, TAG DIGIT ONE, TAG DIGIT NINE, TAG DIGIT FOUR, TAG DIGIT THREE, TAG DIGIT TWO

The commas are just for clarity in this description, they are not in the message.

This seems a lot, yet this could be from a cascading menu system where the person sending the message would select the word apricot (tree) from a menu and the software would insert the nine characters into the message automatically.

4. The suggested LOCALIZABLE SENTENCE BASE CHARACTER character

The second base character for a tagspace is

LOCALIZABLE SENTENCE BASE CHARACTER

and the tagspace would encode in a language-independent form a finite subset of the grammatically stand-alone complete whole sentences that can be expressed in a language.

The format for encoding in the tagspace is not suggested in this document.

However, in an explanation of the concept a sequence of five tag digit characters has been used.

Here is a suggested visible glyph for the character.

The glyph would not typically be displayed in a software-assisted application scenario yet is included for completeness as it could be useful for a graceful fallback display in some circumstances.

5. Examples of how encoded items in the Localizable Sentence tagspace could be used

A good selection of preset grammatically stand-alone complete whole sentences that can be expressed in a language would be encoded in a language-independent form.

Please note that the sentences are here expressed in capitals as in an encoding standard.

A localized version would use ordinary conventions as to the use of uppercase letters and lowercase letters in those languages where two cases of letters are normally used.

These sentences could include everyday items such as

GOOD DAY.

and

BEST REGARDS,

to sentences such as

IS THERE ANY INFORMATION ABOUT THE FOLLOWING PERSON PLEASE?

in a situation where someone is seeking news of a relative or friend after a disaster in another country.

An example of use.

Is there any information about the following person please?

Margaret Gattenford

Here a localizable sentence and the name of a person are used together. The localizable sentence would become localized into the local language yet the name of the person would go through unaltered.

6. Examples of how encoded items in the PanLex tagspace and the Localizable Sentence tagspace could be used together

This would require some additional items to become encoded into the Localizable Sentence tagspace, yet no additional items would need to become encoded into the PanLex tagspace.

This would open up lots of possibilities for communicating through the language barrier.

There would be some situations where a PanLex word would follow a localizable sentence.

For example, the following localizable sentence.

THE PERSON NEEDS THE FOLLOWING MEDICINE.

That could be followed by a PanLex word.

Also, there could be a general mechanism so as to be able to send many messages through the language barrier. I deliberately did not write that any message could be sent as there could be limits for some very complicated sentences: the scope of what is possible could hopefully become increased as research continues, yet there may perhaps always be some messages that could not be sent in this manner.

The general mechanism would be somewhat more complicated to use, yet maybe software assistance could in the future make this straightforward to apply.

For example, suppose that one wishes to send through the language barrier the following sentence.

The apricot tree is beautiful.

One would need to construct the sentence.

There could be localizable sentences for constructing such a sentence.

For example, the following.

The subject noun of the sentence being constructed is as follows.

That would be encoded as LOCALIZABLE SENTENCE BASE CHARACTER followed by a sequence of tag characters to represent the code for the localizable sentence.

That localizable sentence would be followed by PANLEX BASE CHARACTER followed by a sequence of tag characters.

LOCALIZABLE SENTENCE BASE CHARACTER, TAG DIGIT SIX, TAG DIGIT SEVEN, TAG DIGIT ZERO, TAG DIGIT ZERO, TAG DIGIT ONE

PANLEX BASE CHARACTER, TAG DIGIT SEVEN, TAG DIGIT TWO, TAG DIGIT FIVE, TAG DIGIT ONE, TAG DIGIT NINE, TAG DIGIT FOUR, TAG DIGIT THREE, TAG DIGIT TWO

That is a lot of characters to send the message that the subject noun of the sentence being constructed is an apricot tree, yet whereas localizable sentences are intended for a limited set of sentences, (some just for friendly chat; some for serious applications, such as, for example, seeking news of relatives and friends after a disaster) this combination has the potential to send a message involving any words.

I mentioned a localizable sentence as follows.

The subject noun of the sentence being constructed is as follows.

There could be a number of similar sentences, such as, as examples, the following five sentences.

The direct object noun of the sentence being constructed is as follows.

The verb of the sentence being constructed is as follows.

The verb is in the present tense.

The verb is in the future tense.

The day of the week is as follows.

This technique is interesting because it puts the activity of localizing the sentence about the apricot tree into the target language, as the activity of the recipient of the message, who would typically be a native speaker of the language.

Also the same message could be sent to more than one recipient, where each of the recipients may not necessarily localize the sentence into the same target language.

7. Conclusions

This document suggests possibilities for the future. I ask that consideration please be given to the long-term usefulness of these possibilities in software-assisted communication through the language barrier.