Subject: Re: Choosing the best way to go full UNICODE
Posted by cbpporter on Thu, 08 Jun 2017 08:43:00 GMT
View Forum Message <> Reply to Message

mirek wrote on Thu, 08 June 2017 11:26cbpporter wrote on Thu, 08 June 2017 10:00My priority is CodeEditor right now, but right after I do want to take care on my side of canonical decomposition too.

What do you plan?

I guess you missed my saga: http://www.ultimatepp.org/forums/index.php?t=msg&th=9945&start=0&

I'm guessing I have about 3-4 hours more of work and I'll have a preview version done. Then I'll upload it there and include it into my daily builds and test it for a couple of weeks.

mirek wrote on Thu, 08 June 2017 11:26
Quote:
I'll gather stats like what % of characters can be decomposed and into how many characters on average to come up with an optimal scheme. For Latin languages I expect a few hundred with at most 3 characters in decomposition, with most having 2.

That is my estimate too. I even thing that 3 codepoints is so sparse, that the basic table should only store 2 (which means single base char + single combining mark) - that will allow for more dense table, and 3 codepoint characters should be handled as exception.

That's why I'm gathering data to make informed decisions. I'll get back to you with the stats.

How do you want to handle compatibility decomposition? Like:
http://www.fileformat.info/info/unicode/char/0149/index.htm

My plan is to have a flag for compatibility decomposition vs normal ones, with it being off by default. I'm not sure, but I think you can exclude them all if you don't want to bother with. Unicode can be complicated.