

---

Subject: Re: Choosing the best way to go full UNICODE

Posted by [mirek](#) on Thu, 08 Jun 2017 08:26:47 GMT

[View Forum Message](#) <> [Reply to Message](#)

---

cbpporter 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?

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.

---