Subject: Strings with national specific characters are wrongly sorted - Sort
Posted by Klugier on Wed, 25 Aug 2021 11:01:50 GMT
View Forum Message <> Reply to Message

Hello,

Today I found that sort returns wrong results in term of special characters:

```
#include <Core/Core.h>

using namespace Upp;

CONSOLE_APP_MAIN
{

 Sort(vec);

 for (const auto s : vec)
 {
  Cout() << s << "\n";
 }
}
```

The results are:

Ala
Edward
Zbig

and should be:

Ala

Edward

Zbig

This is probably corner case, because this world doesn't exist in Polish, but anyway the error is there. I believe it is more serve when these character are in the middle of the string and we have a lot of such words.

Here is the article about Polish alphabet and the order of letters.

Klugier

Subject: Re: Strings with national specific characters are wrongly sorted - Sort
Posted by busiek on Thu, 26 Aug 2021 11:09:55 GMT
View Forum Message <> Reply to Message

For Polish alphabet I use such function to map wchar into something comparable according to our needs:

```
auto CharRepr = [](wchar c)
{
   wchar d = ToLower(c);

   return MakeTuple(IsLetter(c), ToAscii(d), c);
};
```

Then, to compare, I map each WString into the appropriate vector of tuples.

Cheers,
Kuba

Subject: Re: Strings with national specific characters are wrongly sorted - Sort
Posted by mirek on Fri, 27 Aug 2021 07:51:58 GMT
View Forum Message <> Reply to Message

Klugier wrote on Wed, 25 August 2021 13:01Hello,

Today I found that sort returns wrong results in term of special characters:

```
#include <Core/Core.h>

using namespace Upp;

CONSOLE_APP_MAIN
{

 Sort(vec);

 for (const auto s : vec)
 {
  Cout() << s << "\n";
 }
}
```

The results are:

Ala
Edward
Zbig

and should be:

Ala

Edward

Zbig

This is probably corner case, because this world doesn't exist in Polish, but anyway the error is there. I believe it is more serve when these character are in the middle of the string and we have a lot of such words.

Here is the article about Polish alphabet and the order of letters.

Klugier

This is not error, base [W]String comparison simply compares character values.

You need to use NLS specific sorting in this situation - LanguageInfo::Compare. That said, it really is specifically defined just for CZ and even there it would need improvement, OTOH the generic routine should at least work better that the result you get.

BTW, language specific sorting is extremely difficult topic if it should be done right in many languages...

Mirek

Subject: Re: Strings with national specific characters are wrongly sorted - Sort
Posted by mirek on Fri, 27 Aug 2021 07:57:47 GMT
View Forum Message <> Reply to Message

BTW, as I am revisiting this issue after a long long time, it looks like we should connect here to host platform APIs in the future.
