Title: Lithuanian Ad-hoc Report

Date: 2012-2-14
Source: Lithuanian Ad-hoc Committee

An ad-hoc committee on Lithuanian met in Mountain View on February 14, 2012. The following were in attendance:

- Tero Aalto, Debbie Anderson, Peter Constable, Michael Everson, Algirdas Krupovnickas, Mike Ksar, Evaldas Kulbokas, Alain LaBonté, Michel Suignard, Uma Umamaheswaran, Ken Whistler, Satoshi Yamamoto

The ad-hoc meeting was chaired by Peter Constable.

The ad-hoc committee considered concerns related to accented Lithuanian text that have been raised in various documents submitted to WG2, and that were presented by Lithuanian experts during the ad-hoc committee meeting.

Some of the contributions submitted to the WG2 document register assumed problems related to text input and display. N4193, for instance, points to problems in text display and recommends that software and font vendors implement adequate support for display of Latin text with combining characters.

While such issues exist in some implementations and are real user concerns, Lithuanian experts clarified that users are even more significantly impacted by other implementation issues, such as searching. The range of concerns for Lithuanian users are discussed in Lithuanian contributions N4191 and N4191-A.

The remedy proposed in N4191 is to encode new, pre-composed, Latin characters for Lithuanian accented letters. Considerable time was spent evaluating the efficacy of this solution. Various experts provided input explaining that this approach would not provide the assumed benefits in the long term. Chief among the problems is the effect of normalization, which would lead to data using pre-composed characters getting converted into the current decomposed representation. In other words, even if new characters are encoded, the need for implementations to support Lithuanian text using the current decomposed representation will still remain.

Rather than being viewed as an encoding problem, experts suggested that it would be useful to consider this a problem with inadequate software implementations. In particular, a serious commitment is needed from industry for commercial software products and other widely-used software (e.g., operating systems, core libraries, online search and other Web services) to provide support for Latin accented characters. A particular implementation concern that was identified was *asymmetric search*.

> Note: An example of an asymmetric search that was discussed was having a query string with an un-accented letter "e" match strings with "e", but also with strings with accented letters "ė" or "ę". It was noted that asymmetric search was a topic that would be discussed later in the week during the OWG-Sort meeting.

In view of these considerations, the Lithuanian experts were recommended to pursue alternate approaches that would more likely improve software implementations in private or commercially-available software products to meet the requirements of Lithuanian users. Some specific suggestions given include:

- Noting that issues involving accented letters impact hundreds of languages (e.g. Navajo, various African languages), leverage that broader customer impact in presenting a business case for software vendors to improve their implementations.
- Consider the potential role that the ISO/IEC 14651 standard might have either in stating requirements or in designing adequate implementations involving searching or sorting.
- Consider the potential role that that the Unicode CLDR project (http://cldr.unicode.org/) might have in capturing requirements for Lithuanian text processing, in driving awareness of those requirements, and in improving implementations in products that make use of CLDR data.
  > Note: In relation to this, Lithuanian experts are encouraged to actively participate in the Unicode CLDR project.
- Continue to engage software vendors directly in order to clarify requirements and to provide feedback on existing issues with their products. Also, consider potential business incentives that may motivate vendors to make needed improvements.

Recommendations to WG2

In view of the above discussion within the ad-hoc committee, it is recommended that WG2 *not* take action to encode additional characters as was proposed in N4191. There may be other actions that WG2 may want to consider in response to Lithuanian concerns—actions that would increase the likelihood that software implementers would be aware of requirements for Lithuanian or other languages facing similar challenges. For example, there may be ways to draw more attention to named USIs and to their significance in relation to providing adequate support for the particular language(s) for which given named USIs were defined. The ad-hoc does not have specific proposals of this nature to present, but suggests that this general suggestion be given consideration.