ISO/IEC JTC1/SC2/WG2 N4191-A

Date: 2010-02-10

ADDENDUM TO THE DOCUMENT *ISO/IEC JTC1/SC2/WG2 N4191*

This Addendum should be considered as addendum to chapter 5 *Rendering of sequences* of the document ISO/IEC JTC1/SC2/WG2 N4191.

Apart from rendering there are other problems when using Combining Character Sequences mechanism. Lithuania request to have all Lithuanian letters as encoded including 35 accented letters identified by the named sequences is fully legitimate and should be supported by all means. Lithuania request is supported by the stakeholders from several parties in Lithuania and abroad (ISO/IEC JTC1/SC2/WG2 N4187, 4188, 4189, 4192). The decision from 1996 that encoding of precomposed Latin letters is no longer necessary severely impacts usage of contemporary Lithuanian language in IT applications.

Currently Lithuania is only European country which is forced to use precomposed sequences for it official language. European Union tries to preserve its linguistic diversity promoting all but especially lesser used languages that might be in danger. Lithuanian belongs to the category due to the highest rate of emigration among EU members. Actually it is a Diaspora nation with more than half a million of Lithuanians living abroad. Schoolchildren in the emigrant's communities are provided with a distant computer aided learning possibilities in order to support and develop their native tongue to such an extent that they were able to study at Lithuanian universities. Accentuation in general and accented texts in particular is crucial for their language skills.

Accented text become of paramount importance also for the native speakers living in the country because of the language change. The stress in Lithuanian is not fixed changing its position depending on the accentuation paradigm. Sometime initial, sometimes final parts of the word are stressed. The latter case is in danger as many native speakers tend to stress only initial positions (as it is easier to pronounce a word in this way). If the endings permanently loose their stress, they will be shortened or disappear altogether as it has already happened for another Baltic language, i.e. Latvian. In order to preserve the stress in the final positions more texts have to be accentuated. Attempts are made to do it automatically using special software tools as "Kirciuokle" (http://donelaitis.vdu.lt). However, such tools can be supported only by the national fonts which are not compatible with popular learning environments, e.g. MOODLE.

Recent developments in the field of digitalization of the cultural heritage and other fields raised new challenges coming and important problems for Lithuanian language not having all letters in Unicode as well:

Common research infrastructures based on linguistic resources. Since the appearance of ESFRI, (the European Strategy Forum on Research Infrastructures that is used as a strategic instrument to develop the scientific integration of Europe and to strengthen its international outreach) a number of projects, initiatives and consortia appeared. Some of them, like CLARIN (Common Language Resources and Technologies Infrastructure) are committed to establish an integrated and interoperable research infrastructure of language resources and its technology enabling e-Humanities. Lithuania is a member of the consortium providing its linguistic resources (text and speech) as well as software tools

for natural language processing for an international research community. Since Lithuanian is the oldest living Indo-European language it attracts researchers' attention from all over the world. They use corpora, digital dictionaries and search tools for their research. For their needs accented letters are necessary. Moreover, other languages included in CLARIN infrastructure are fully supported by the ISO/IEC 10646 and the UNICODE standards, leaving Lithuanian language in a worse position.

Search in large text corpora including accented text. As languages tend to shrink with respect to the number of words that are being used in everyday spoken language it is important to compile large text corpora, storing language information as part of national heritage. That is of paramount importance for the lesser used languages, Lithuanian language being one of them. Large corpora serve multiple purposes; therefore they include a great variety of texts, accented texts being part of them. Since Unicode is currently used as the main corpora encoding standard, it is important to have uniform representation for all the letters used (both accented and regular) in the corpus, in order to ensure correct search options. Search of words and their environment (so called concordancing) is one of the basic corpora-related services, implemented either as an on-line service, or as a network service, that can be integrated as a building block in other complex services. Accented letters are important for disambiguation when defining search patterns, as well as in these cases, when corpora are used in machine learning for algorithm training purposes. Moreover, extended search options that include accented letters are vital for research in the field of computational linguistics. For all the above mentioned reasons, search of accented Lithuanian letters should be supported correctly in different operational and design environments, as well as by different application systems.

Search in speech corpora. Taking into account the increasing need of speech analysis and synthesis in different applications (e.g. virtual assistant applications, automatic speech-text recording of medical records, media information, etc.), correct search options should be ensured for speech corpora, including both spoken language and its transcription, the latter including also accented letters. Accents are important for disambiguation, while using speech corpora for algorithm training purposes. Lithuanian speech corpora and speech recognition/synthesis are in a fast development phase right now, and their designers are pointing to the accented letter management problem as one of the most important problems in their design field.

Search in electronic dictionaries. Electronic dictionaries/thesaurus is an important part of the programs for the preservation of the national heritage. Dictionaries explicitly include accented words, and regular dictionary search must support also search patterns with accented letters.

Conventional multi-level search patterns. Accented letters are needed while executing multi-level search, i.e. taking the results of the first search level for narrowing search at the next level. This is normally done by applying the „copy-paste“ procedure, resulting in a failed search for numerous standard applications.

Modern learning systems based on machine learning. Modern learning systems are increasingly using different machine learning (artificial intelligence) approaches. In this case, algorithm training is the main component, requiring large corpora, speech corpora, thesaurus, dictionaries and other linguistic resources with correct search procedures implemented. Here, accented letters are especially important in language learning systems.

Accented Lithuanian letters are very widely used. Their usage spans various media and means of information technology. Presently, the Combining Character Sequences mechanism in most IT products is being supported only episodically. Lithuanian accented letters are mostly used in Lithuania, so it would seem that it would be possible to code the missing letters using the PUA defined by the UNICODE and adopting a local Lithuanian standard for that. This is only temporary and palliative means which does not take into account proliferation of IT usage. Without proper inclusion of all

Lithuanian letters into the UNICODE, the emergence of international software that is non-discriminative (i.e. with Lithuanian full sorting order, search engines, etc.) to Lithuanian seems not likely as well, which raises an issue of the compliance with the license agreements and ROI of the same product in different countries.

The optimal way of solving the problem (as was done with ancient Greek) would be **the appointment of UNICODE positions for the missing Lithuanian letters in the new version of the UNICODE and ISO/IEC 10646 standard**. The accented letters of other languages have their own UNICODE coding for a long time already and neither composition sequences nor PUA are being used for their information processing. The same must be done for the rest 35 Lithuanian accented letters.