

Encoding of Lithuanian Accented Letters

Vladas TUMASONIS

1. Main alphabet

Lithuanian is, by its grammatical structure, one of the most archaic languages among the living Indo-European languages. It is spoken by approximately 5 million people and is subject to linguistic studies at many universities all over the world.

The main Lithuanian alphabet consists of the Latin alphabet (excluding *Q, q, W, w, X, x*) with 18 extra letters with diacritics (9 capital and 9 small):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Ą | Č | Ę | Ė | Į | Š | Ų | Ū | Ž |
| ą | č | ę | ė | į | š | ų | ū | ž |

These letters belong to the Lithuanian main alphabet and are included in several 8-bit single-byte coded character sets (ISO/IEC 8859-13 [1998], MS CP 1257, IBM CP 775, etc.). Thus there is no problem to use them.

2. Extended alphabet (with accented letters)

Lithuanian words do not have a fixed stress position (AMBRAZAS 1997). The stress may fall on every syllable of the word. Its constitutive function manifests itself in distinguishing a word from a combination of words. The second function of word stress is its distinctive function, which distinguishes otherwise identical words by the place where the stress falls, e.g.:

| | |
|----------------|---------------|
| <i>añtis</i> | bosom |
| <i>ántis</i> | duck |
| <i>áukštas</i> | high; tall |
| <i>aũkštas</i> | floor; storey |

For word stressing (or accenting), three accent marks (or diacritical marks in ISO terms) are used: grave accent, acute accent and tilde (cir-

cumflex). The position of the stress depends on the stress pattern (or accentual paradigm) of the word and its morphological structure (cf. the examples above).

Word stress is thus expressed by means of accented letters. There are 68 accented letters in the Lithuanian language (cf. figure 1).

| | | | | | | |
|---|---|----|---|---|---|---|
| À | Á | Ã | Ą | Ā | | |
| à | á | ã | ą | ā | | |
| È | É | Ē | Ė | Ĕ | É | Ě |
| è | é | ē | ė | ĕ | é | ě |
| Ì | Í | Ī | Į | Ĭ | Ý | Ÿ |
| ì | í | ī | į | ĭ | ý | ÿ |
| Ĵ | | Ľ | | Ṁ | | Ñ |
| ĵ | | l̄ | | ṁ | | ñ |
| Ò | Ó | Õ | | Ř | | |
| ò | ó | õ | | ř | | |
| Ù | Ú | Ū | Ū | Ů | Ú | Ů |
| ù | ú | ū | ų | ů | ú | ů |

Figure 1: Lithuanian accented letters

Together with the main letters, the accented letters constitute the extended alphabet.

The use of accented letters goes back to the first Lithuanian writings. Even some of the first printed Lithuanian books were accented, for example Daukša's "Kathechismas" (DAUKŠA 1595) and "Postilla Catholica" (DAUKŠA 1599). In present-day publishing practice, all dictionaries, special vocabularies and encyclopaedias are accented. Accented letters are also used in textbooks for schools, reference books, linguistic texts, and in the publication of laws.

In common press publications (newspapers, fiction, etc.), only the letters of the main Lithuanian alphabet are used. Here, accented letters are used only in those words where they have a distinctive function.

3. 8-bit single-byte coding (National standard code tables)

For the encoding of accented letters, there are three national 8-bit single-byte standard code tables used in Lithuania. The basic Lithuanian accented letter code table is valid for a UNIX environment (cf. Pr LST 1564-2 1998; the second half of this table is shown in figure 2). It defines the basic character repertoire including accented letters. This code table is conformant with ISO/IEC 8859-13, i.e. the codes of all Lithuanian main letters in both tables are the same.

Commonly used and very important graphic characters are retained. The repertoire of this table is optimal for linguistic text processing.

The code table used for a Windows environment contains the basic character repertoire and extra phonetic symbols in rows 8 and 9. This code table is conformant with ISO/IEC 8859-13.

The code table used for DOS contains basic characters and box drawing symbols and is conformant with IBM CP 775 for the Baltic States. The usage of a DOS environment is still popular in publishing houses.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|------------|----------|----------|
| 8 | 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 |
| 9 | 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 |
| A | NBSP 160 | Ą 161 | Ę 162 | Ė 163 | į 164 | Ĺ 165 | Ų 166 | Ń 167 | Ė 168 | Ń 169 | Ė 170 | Ų 171 | Ų 172 | SHY 173 | Ų 174 | Ų 175 |
| B | Ī 176 | ą 177 | ę 178 | ė 179 | ´ 180 | Ī 181 | ¶ 182 | · 183 | é 184 | ñ 185 | ě 186 | ř 187 | ų 188 | Ų 189 | ũ 190 | ú 191 |
| C | Ą 192 | Į 193 | À 194 | Á 195 | Ä 196 | Å 197 | Ę 198 | Ą 199 | Č 200 | É 201 | È 202 | Ê 203 | Ë 204 | Ì 205 | Í 206 | Ī 207 |
| D | Š 208 | Į 209 | Ò 210 | Ó 211 | Ý 212 | Õ 213 | Ö 214 | Ų 215 | Ų 216 | Ù 217 | Ú 218 | Ū 219 | Û 220 | Ÿ 221 | Ž 222 | ß 223 |
| E | ą 224 | į 225 | à 226 | á 227 | ä 228 | å 229 | ę 230 | ą 231 | č 232 | é 233 | è 234 | ê 235 | ë 236 | ì 237 | í 238 | ĭ 239 |
| F | š 240 | į 241 | ò 242 | ó 243 | ý 244 | õ 245 | ö 246 | ų 247 | ų 248 | ù 249 | ú 250 | ū 251 | û 252 | ÿ 253 | ž 254 | ÿ 255 |

Figure 2: UNIX code table for Lithuanian accented letters (second half)

4. Multiple-Octet coding in ISO/IEC 10646-1 (UCS codes)

All letters of the main Lithuanian alphabet have UCS codes (codes in ISO/IEC 10646-1 1993) or UNICODE codes. With Lithuanian accented letters, however, the situation is more complicated.

| | | | | | | |
|---|---|---|---|---|---|---|
| À | Á | Ã | Ą | Ą | | |
| à | á | ã | ą | ą | | |
| È | É | Ë | Ė | Ė | Ė | Ë |
| è | é | ë | ė | ė | ė | ë |
| Ì | Í | Ĩ | Į | į | Ý | Ÿ |
| ì | í | ĩ | į | į | ý | ÿ |
| Ĵ | | Ł | | Ń | | Ñ |
| ĵ | | ł | | ń | | ñ |
| Ò | Ó | Õ | | Ŕ | | |
| ò | ó | õ | | ŕ | | |
| Ù | Ú | Û | Ū | ū | Ū | Û |
| ù | ú | û | ū | ū | ū | ũ |

Figure 3: Lithuanian accented letters with respect to UCS codes

As was mentioned above, Lithuanian accented letters are Latin script letters with grave accent, acute accent or tilde. Thus some Lithuanian accented letters are also common letters of other languages. For example, LATIN LETTER A WITH ACUTE is also used in Irish, Icelandic, Portuguese, Slovak, etc., while LATIN LETTER N WITH TILDE is also met with in Basque, Breton, and Spanish. This is why they have been assigned UCS codes.

All in all, there are 33 Lithuanian accented letters that have proper UCS codes, and 35 accented letters that have no proper UCS codes. The non-shadowed letters shown in figure 3 have UCS codes, the shadowed letters have none.

At present, a proposal to add these Lithuanian accented letters to ISO/IEC 10646-1 is being prepared. This proposal submits 35 characters

to be included in the existing UNICODE block LATIN EXTENDED-B. The category of these characters is A. According to their structure, all Lithuanian accented letters may be expressed by composite sequences using a Latin script letter and one or two combining characters (diacritics). The proposed letter names reflect this structure and are in accordance with the “Character naming guides” as given in Annex K of ISO/IEC 10646-1. Examples:

a) With one combining character:

Ṁ LATIN CAPITAL LETTER M WITH TILDE

ṁ LATIN SMALL LETTER M WITH TILDE

b) With two combining characters:

Ą LATIN CAPITAL LETTER A WITH OGONEK AND ACUTE

ą LATIN SMALL LETTER A WITH OGONEK AND ACUTE

A peculiar problem is raised by the small letters “i” and “i with ogonek”. The Lithuanian letter “i” is always written with a dot above. All accented forms of “i” should also be dotted (cf. the examples given in 5 below). In ISO/IEC 10646-1 all corresponding forms are dotless. For example, LATIN SMALL LETTER I WITH ACUTE in fact specifies “Latin small letter dotless i with acute”. For Lithuanian, we ought to retain a dot above so that we should define these letters as LATIN SMALL LETTER I WITH DOT ABOVE AND ACUTE instead (or maybe even as LATIN SMALL LETTER DOTLESS I WITH DOT ABOVE AND ACUTE).

5. Samples

From KEINYS (1993), 350:

laik̄klis (2) *tech.* prietaisas ar įtaisas kam laikyti:
Spyruoklės, šepetio, ritės l.
laĩkin|as, ~à (3^b) kurį laiką esantis ar trunkantis, ne-
nuolatinis, neamžinas: *L. reiškinyš.* ~à *tarnyba.* ~aĩ
prv.: Derybos ~aĩ nutrauktos. ~ũmas (2)
laĩkinink|as, ~é *dk.* (1) 1. *sport.* teisėjas, fiksuojantis
laiką. 2. palaikiui apmokamas darbininkas

From Mišiolas (1982), 75: Garbē táu, Diēve, visātos Kūrējau!
 Iš tāvo dosnūmo tūrime vỹno,
 kurī aukójame táu.
 Tàs vỹnmedžio iř žmogaūs dārbo vaišius
 tāps mūms dvāsiniu gērimu.

| | | | |
|---------------------------------------|-----|---------|---------|
| From LAIGONAITĖ & ZINKEVIČIUS (1997), | V. | mažì | māžos |
| 38 (note the accented i): | K. | mažų | mažų |
| | N. | mažíems | mažóms |
| | G. | mažùs | mažàs |
| | Įn. | mažaís | mažomìs |
| | Vt. | mažuosè | mažosè |

References

- AMBRAZAS, V. (ed., 1997): Lithuanian Grammar. Vilnius: Baltos lankos.
- DAUKŠA, M. (1595): *Kathechismas arba mokslas kiekvienam krikščionii priwalvs*. Facsimile edition in: Mikalojaus Daukšos 1595 Metų katekizmas. / Katechismus von Mikalojus Daukša vom Jahre 1595. Vilnius: Mokslo ir enciklopedijų leidykla 1995.
- DAUKŠA, M. (1599): *Postilla Catholicka. Tái est: Išguldimas Ewangeliu kiekvienos Nedelos ir szwetes per wissús metús*. Facsimile edition of title page and some other pages in: K. KORSAKAS ir J. LEBEDYS (ed.), Lietuvių literatūros istorijos chrestomatija. Feodalizmo epocha. Vilnius: Valstybinė Grožinės Literatūros Leidykla 1957, 60-68.
- ISO/IEC 8859-13 (1998): Information technology – 8-bit single-byte coded graphic character sets – Part 13: Latin alphabet No. 7.
- ISO/IEC 10646-1 (1993): Information technology – Universal Multiple-Octet Coded Character Set (UCS) – Part 1: Architecture and Basic Multilingual Plane.
- KEINYS, St. (ed., 1993): Dabartinės lietuvių kalbos žodynas [Dictionary of The Modern Lithuanian Language]. Vilnius: Mokslo ir enciklopedijų leidykla.
- Mišiolas (1982): *Romos Mišiolas. Gedulinis Mišiolas* [*Missale Romanum. Missale Parvum*]. Kaunas / Vilnius.
- LAIGONAITĖ, A. ir ZINKEVIČIUS, Z. (1997): Lietuvių kalba. Mokomoji knyga X klasei [The Lithuanian Language. Textbook for the 10th class]. Kaunas: Šviesa.
- Pr LST 1564-2 (1998): Information technology – 8-bit single-byte character coding – Part 2: Lithuanian accented letter and phonetic character set.