

Testing Gran Paradiso (Firefox 3) Alpha 5 on Windows XP for Sinhala-Unicode compatibility

Test Environment:

Windows XP with Service Pack 2 & Sinhala Enabling Pack

In addition, Microsoft Internet Explorer was used to try alternatives.

Tested Pages:

1. Wikipedia – Sinhala – Article: Sri Lanka
http://si.wikipedia.org/wiki/%E0%B7%81%E0%B7%8A%E2%80%8D%E0%B6%BB%E0%B7%93_%E0%B6%BD%E0%B6%82%E0%B6%9A%E0%B7%8F%E0%B7%80
2. Sinhala Website of Language technology Research Laboratory – University of Colombo School of Computing
<http://www.ucsc.cmb.ac.lk/ltr/?lang=si&page=home&style=default>
3. Gmail: <http://mail.google.com/mail/> for Gmail had issues with ZWJ previously.

Identified Issues:

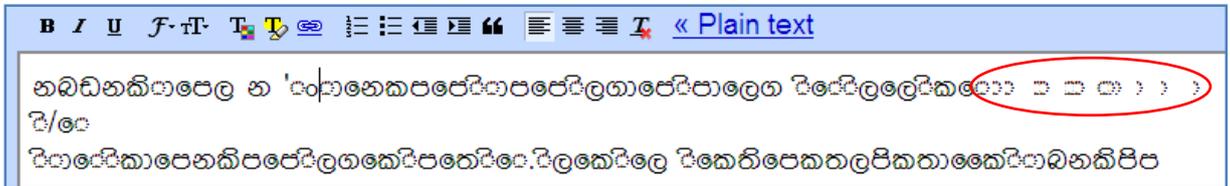
1. Sinhala rendering works fine (ZWJ is not filtered-out any more) but some issues are there with spacing between words. Firefox 2 had an issue where it filters-out the zero-width-joiner (ZWJ) character from text leading to improper rendering of words.

ඉන්දියන් සාගරයේ මුතු ඇටය නමින් විරුදාවලිය ලත් ශ්‍රී ලංකාව, ආසියා මහද්වීපයට අයත් කුඩා දූපතකි. එය ඉන්දියාවට පහලින් දකුණු පසට වන්නට පිහිටා ඇත. එහි සම්පූර්ණ ප්‍රමාණය වර්ග කිලෝමීටර 65610 කි. ශ්‍රී ලංකාවේ අග නගරය ශ්‍රී ජයවර්ධනපුර කෝට්ටේ වේ. "ශ්‍රී ලංකා ප්‍රජාතාන්ත්‍රික සමාජවාදී ජනරජය" එහි නිලනාමය වෙයි. අවුරුදු 2500 කටත් වඩා වැඩි ප්‍රේමාඩ ලිඛිත ඉතිහාසයකට උරුමකම් කියන ශ්‍රී ලංකා ඉතිහාසය, ක්‍රි.පූ. 6 වන සියවසේදී කුමරුගේ ආගමනයත් සමඟ ආරම්භ වූ බැව් කියත ශ්‍රී ලංගොඩ මානවයා සොයා ගැනීමෙන් පසු එහි ඉතිහාසය ගනන් බැලිය නොහැකි දුරකට දිවයයි. ප්‍රථම රාජධානිය ලෝකපුරාධිපුරය සැලකෙන අතර, රාජ්‍ය පරපුරේ අවසාන යමනනුවර නගරය විය.

Source:

http://si.wikipedia.org/wiki/%E0%B7%81%E0%B7%8A%E2%80%8D%E0%B6%BB%E0%B7%93_%E0%B6%BD%E0%B6%82%E0%B6%9A%E0%B7%8F%E0%B7%80

- Text-area for composing emails in Gmail does not refresh properly when the existing text is edited. The same test showed no issue with other web-sites containing text-areas.



This is after deleting some text from the middle of the first line.

- Saved web-pages with Sinhala filenames are unable to load into Firefox 3 when the file is double-clicked or drag-dropped. This was there on Firefox 2 as well.

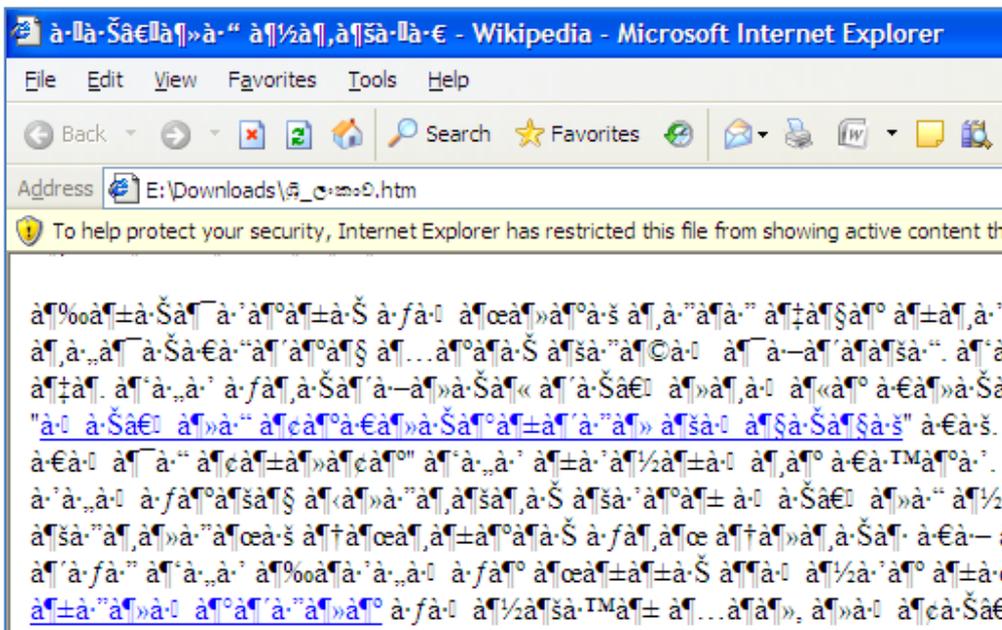
The above mentioned Wikipedia article was saved as “ශ්‍රී ලංකාව.htm” from Firefox 3, and when tried to load (by double-clicking / drag-dropping), it gave the error:

File not found

Firefox can't find the file at /E:/Downloads/?????_?????.htm.

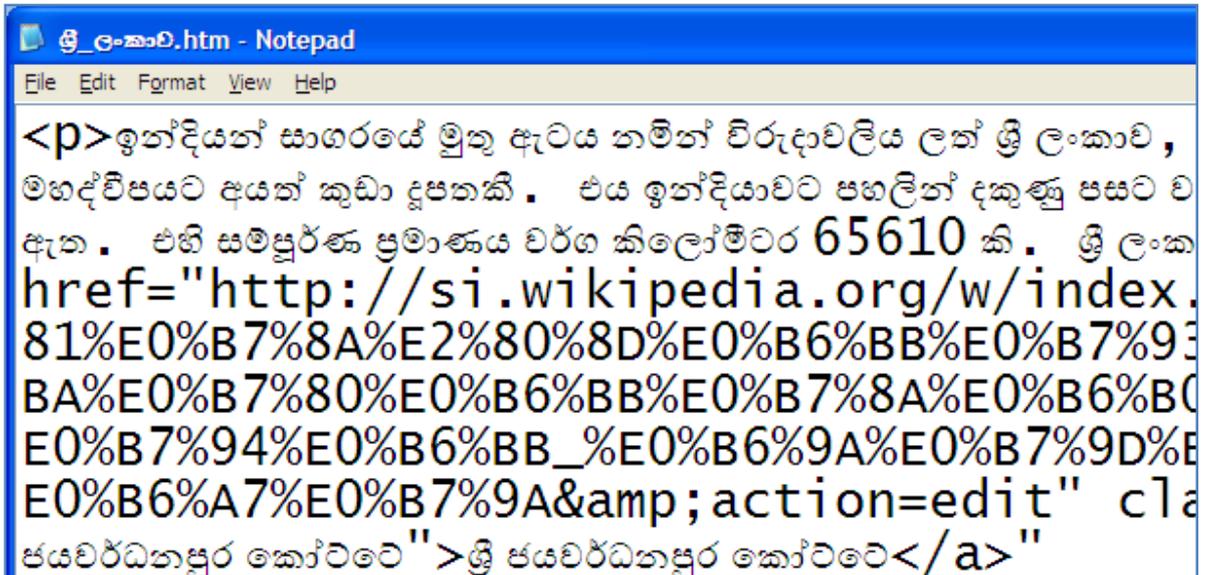
- * Check the file name for capitalization or other typing errors.
- * Check to see if the file was moved, renamed or deleted.

The procedure was repeated with Internet Explorer (file double-clicking and drag-dropping) and it did not report an error but the text was displayed as split multi-bytes.





When the file was opened with Notepad, however, everything was properly displayed. And the file encoding was found to be UTF-8.



- There appears to be another issue displaying Sinhala in some web-pages in Firefox 3 while the same page displays fine with Firefox 2.

Firefox 3 showin the LTRL webpage with problems of Sinhala-Unicode character lookups.

භාෂා තාක්ෂණ පර්යේෂණාගාරය - Gran Paradiso

File Edit View History Bookmarks Tools Help

http://www.ucsc.cmb.ac.lk/ltr/?lang=si&page=home&style=default

භාෂා තාක්ෂණ පර්යේෂණාගාරය

භාෂා තාක්ෂණ පර්යේෂණාගාරය

කොළඹ විශ්වවිද්‍යාලයේ පරිගණක අධ්‍යයනායතනය

පරිශීලක උදවු භාෂාව තෝරන්න:

නිවෙස් පිටුව

- ව්‍යාපෘති
- PAN දේශීයකරණය
 - පළමු අදියර
 - දෙවන අදියර
- පර්යේෂණ
 - වෙනත් ව්‍යාපෘති
- අප ගැන
 - දූෂේ කණ්ඩායම්
 - හවුල්කරුවන්
 - ආයතන ප්‍රවේශ
- මූලාශ්‍ර
 - බාහත හැකි
 - ප්‍රකාශන
 - පෝච්චා
 - සබැඳි
- සටහන්
- සාකච්ඡා
- විකි ය

පරිගණක දේශීයකරණ කටයුතු සහ භාෂා පැහැදිලිකරණ කටයුතු හා පර්යේෂණ අරමුණු කරගෙන, 2004 දී භාෂා තාක්ෂණ පර්යේෂණාගාරය පිහිටුවනු ලැබීය.

ජාත්‍යන්තර පර්යේෂණ හා සංවර්ධන මධ්‍යස්ථානයක් කැපවීමට අරමුණු ලැබීමත් සමඟම, භාෂා තාක්ෂණ පර්යේෂණාගාරය පුද්ගලික දේශීයකරණ ව්‍යාපෘතිය ආරම්භ කළේ භාෂා පැහැදිලිකරණයේ මූලික අවශ්‍යතා කීපයක් වන සිංහල වාක් සංඛ්‍යාවක් (Text Corpus), ශබ්දකෝෂයක් (Lexical Resource), ලේඛන කථනයට හැරවීමේ මෘදුකාංගයක් (Text-To-Speech engine) සහ මුද්‍රිත අකුරු හඳුනා ගැනීමේ මෘදුකාංගයක් (Optical Character Recognition application) එළි දැක්වීමේ අරමුණ සහිතවය. මේ අරමුණු ආර්ථකථි ඉටු කරගත් අපි මිලහ පියවර ලෙස දේශීයකරණයට අදාළ වන තාක්ෂණයන්, පසුගිය කාල විකවානුව තුළ අප ලබාගත් අත්දැකීම් හා දැනුමින් බෙහෙවින්ම අපේක්ෂා කරමු.

කවද, භාෂා තාක්ෂණ පර්යේෂණාගාරය මූලික දේශීයකරණ කටයුතු සඳහා තොරතුරු හා සන්නිවේදන තාක්ෂණ නියෝජිතායතනය (ICTA) සමඟ සහයෝගීත්වයෙන් කටයුතු කරන අතර, අවයංක්‍රීය ව්‍යාකරණ ප්‍රවර්ග හඳුනාගැනීම (Part-of-Speech Tagging), අවයංක්‍රීය වචන වෙනස් කිරීම (Automatic Word Clustering), කථනය හඳුනාගැනීම (Speech Recognition) සහ පරිගණකාශ්‍රිත භාෂා පරිවර්තනය (Machine Translation) සාදි වූ භාෂා පැහැදිලිකරණ කේන්ද්‍රයේ විවිධ පර්යේෂණ කටයුතුවල ද නියැලෙමින් සිටියි.

මීට අමතරව, භාෂා තාක්ෂණ පර්යේෂණාගාරය විවිධ රාජ්‍ය හා පුද්ගලික ආයතන වලට භාෂා පැහැදිලිකරණය සහ දේශීයකරණය පිළිබඳව තාක්ෂණික උපදෙස් හා උපකාර සපයයි. එමෙන්ම අප පර්යේෂණාගාරය භාෂා පැහැදිලිකරණ කේන්ද්‍රයේ පුහුණුකරුවෙකු (instructor) ලෙසද උපකාර සපයන්නෙකු (help desk) ලෙසද කටයුතු කරමු.

කොළඹ විශ්වවිද්‍යාලයේ පරිගණක අධ්‍යයනායතනය

35, 8 වන මාවත

කොළඹ 07

ශ්‍රී ලංකාව

☎ [+94 11] 2581245(-7) ext. 532

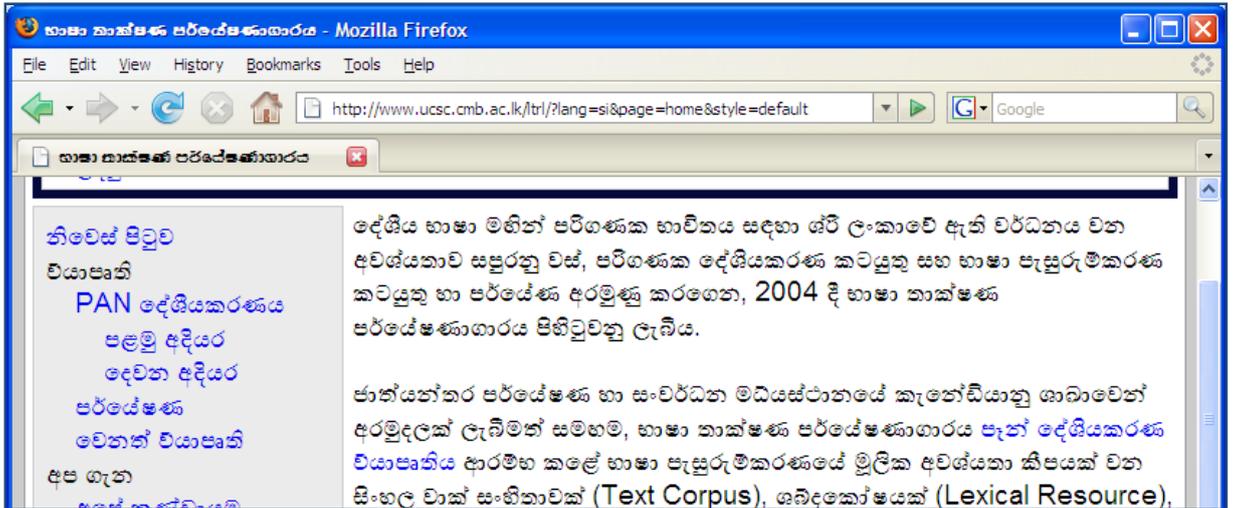
☎ [+94 11] 2587239 ATTN: LTRL

✉ ltr@ucsc.cmb.ac.lk

කොළඹ විශ්ව විද්‍යාලය

කොළඹ විශ්ව විද්‍යාලය, 80 දශකයේ අග භාගයේ සිට, සිංහල කොන්ට්‍රි නිර්මාණය, ඩොස් (DOS) හා වින්ඩෝස් (Windows) මෙහෙයුම් පද්ධති සඳහා සිංහල යතුරු පුවරු ධාවක (keyboard drivers) නිර්මාණය සාදි කටයුතු වල නියැලෙමින් පරිගණක දේශීයකරණයේ පුරෝගාමියෙකු බවට පත්වී සිටියි. 1990 දශකයේ දී උපාධි අපේක්ෂක ව්‍යාපෘති ලෙස දේශීය භාෂා ශබ්දකෝෂ, අක්ෂර විභාගය පරික්ෂක සාදි මෘදුකාංග පර්යේෂණ කිරීමටත්, සිංහල විකවානුවටත් නියැලී ඇත.

Firefox 2 showing the same page without any character lookup issues.



Source: <http://www.ucsc.cmb.ac.lk/ltr/?lang=si&page=home&style=default>

Tested by:

Eranga Jayalatharachchi (dumith.eranga@gmail.com)

Language Technology Research Laboratory

University of Colombo School of Computing

<http://www.ucsc.cmb.ac.lk/ltr/>