

Toward the Interoperability of Language Resources

July 13-15, in conjunction with the 2007 LSA Summer Institute



Issue Statement

Elizabeth J. Pyatt
Penn State University

The greatest barrier to interoperability for multi-lingual Unicode documents is that input methods vary widely between platform and software configurations. For instance a Web page may require the use of a numeric escape characters, while an XML file for Flash does not recognize numeric codes but only UTF-8 text and a Microsoft Office document may not display characters until you switch to a correct font. In addition, there is still very little reliability in data transfer between systems. Unicode data entered correctly into a database backend may be displayed incorrectly on a Web pages if the HTML codes fails to include Unicode headers; similarly UTF-8 e-mail messages may become garbled if an intervening server fails to support Unicode.

Issues specific to linguistics include the lack of full set of phonetic symbols glyphs in most default fonts, the lack of consistent input tools for phonetic symbols or accented letters and the inability of some programs to display the full set of letter and diacritic combinations. For instance, some browsers cannot display diacritics over open o and many programs do not allow users to easily input macrons (long marks) over vowels.

What can be achieved in this workshop? One approach could be to develop in-house third party tools that the community can guarantee will work across platforms. However, it is important that the output of these tools not be so specialized that it would discourage non-specialists from viewing documents. The proliferation of IPA text and specializes accents within the Wikipedia shows that there is an awareness of proper use of phonetic symbols and scripts in the Internet community. I would argue that it is critical that any specialized tools developed by the community must create documents that are easily viewable by non-specialists (i.e. no special plugins required). An interesting example of a specialized tool with universal output is the accent and symbol palette developed by Wikipedia. The tool is specific to the MediaWiki platform used in the Wikipedia, but the output text is readable in most browsers.

Another approach may be to more vocally join the Unicode and open source communities to promote universal Unicode tools, including bundling of more fonts with phonetic symbols as default system fonts and inclusion more usable input utilities. The emergence of new global markets is a strong incentive to software companies to develop tools, although they tend to focus



on a few “power” languages. However, competition from open source tools that are developed by an international community may be even stronger.

Finally, none of these efforts will be successful unless a set of outreach programs is developed to help linguists learn about Unicode and recommended Unicode practices. Some good examples of outreach are the fonts and tools developed by SIL, the Medieval Unicode Font Initiative and others. A bad is example is to develop a tool, but never market it. A major software company did develop a tool to more easily input macrons, hacheks and breves, but they never marketed it. Few language professionals would discover this tool unless they had time to open and explore it. This is a lost opportunity for everyone.

