www.LICT4D.asia/fonts/Urdu_Nasta'leeq

Sarmad Hussain
Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
sarmad.hussain@nu.edu.pk

Introduction

Information has now become such an integral part of our global society, that its access is considered as a basic human right [1, 2]. Moreover, progress of rural and urban developing populations is also getting increasingly dependent upon access to information [3]. This is specifically applicable to Asia which houses the largest developing population. 32 out of 50 Asian countries are considered to be under-developed, having annual per capita income of less than $500 [4]. Access to information is thus critical for these countries.

ICT have emerged as the dominant carrier of information across the globe and therefore access to information can be equated with access to ICT, specifically the Internet. Asians form about 61% of the world population [5] and since 2001 they have also become the largest group of Internet users. Currently, 173 million Asians use the internet, which is 27% of the total internet users across the globe [6]. However, these users form only about 4.5 % of the total Asian population [5], which shows that there is enormous potential for Internet usage in Asia, and therefore an immense opportunity for growth through access to information through the Internet.

Investments have been put into developing ICT infrastructures in Asia. Nevertheless, the persisting digital divide attests that the current path towards providing connectivity and technology infrastructure alone would not enable the majority of Asian populations to benefit from the present information availability [7]. In addition to being most populous, Asia is also the most culturally and linguistically diverse region of the world. There are 2197 languages spoken in Asia, which is the largest number of languages spoken in any one region. Only about 20% of these people can communicate in English [8, 9], majority of whom reside in the developed countries in Asia. Currently, English is the lingua franca for ICT, with majority of content being in English [10]. This makes English language information available on ICT inaccessible to a large majority of Asians. This particularly affects those living in the rural areas of developing countries in Asia. Therefore these populations cannot effectively utilize the information, even if infrastructure is available to access it.

Unless these large non-English speaking populations have the ability to do computing in local languages, they will not be able to use ICT for development effectively. This is reaffirmed by Digital Review of Asia Pacific: "The ready availability of relevant local language content is critical for the development of productive capacity … one of the challenges … of Internet diffusion [lies] in the dominance of the English language … content … with little relevance to many Asia-Pacific Internet users" [11]. Thus, the solution is providing relevant content in local languages to these developing populations. Localization of ICT is as critical as the infrastructure. This is being termed as LICT4D (Localized Information and Communication Technologies for Development).

Access can be broadly categorized as being able to (i) *generate* and (ii) *retrieve* information in local languages. Each of these categories has its own challenges. One of the fundamental challenges to both generation and retrieval of localized information is the availability of fonts, which would enable the Asians to display information in their languages.

Font Development Challenges of Asian Writing Systems

Enabling Asian writing systems pose a unique challenge to English and western-centric ICT. On ICT, English has a character-based left-to-right context-free non-cursive writing system.

Asian writing systems introduce much more variety and complexity. These may be categorized as below:

i) **Variety in units of writing system.** Though some writing systems are character-based, there are others which are syllable-based or based on pictograms/ideograms, e.g. Figure 1 below shows some Chinese characters and their possible evolution from pictures.

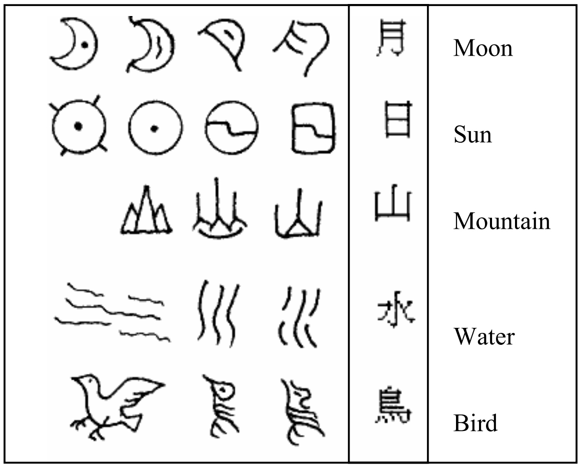
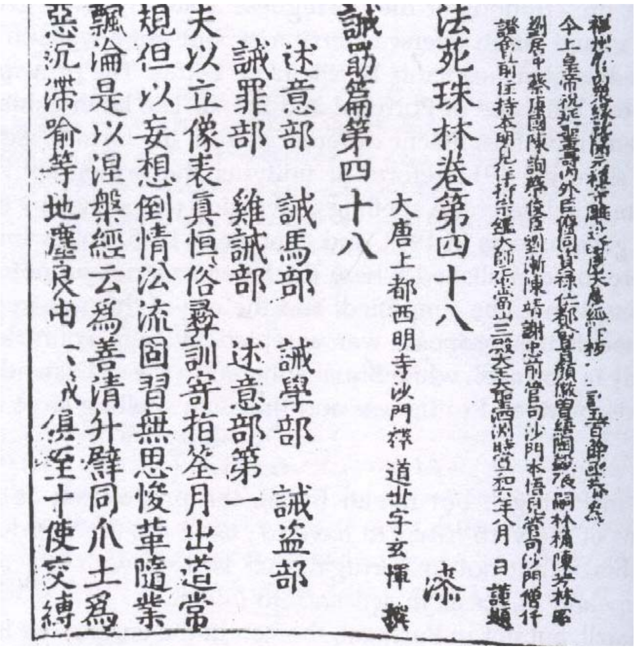


Figure 1: Chinese Pictographs [12]

ii) **Variety in writing direction.** Asians scripts are written left to right (L to R), right to left (L to R), top down. There are also writing systems which are written L to R and R to L and are referred to as bidirectional writing systems.



[illegible]大藏經[illegible]五百餘卷[illegible]
今上皇帝祝延聖壽內外臣僚同資祿位都勸首顏微曾緒閣[illegible]張嗣林楠陳若林郎
劉居中蔡康國陳詢蔡俊臣劉漸陳靖謝忠前管句沙門本悟見管句沙門僧什
證會副住持本明見住持傳法[illegible]大師[illegible]當山三殿[illegible]宣和六年八月 日謹題
法苑珠林卷第四十八 添
大唐上都西明寺沙門釋 道世字玄惲 撰
誡勗篇第四十八
述意部 誡馬部 誡學部 誡盜部
誡罪部 雜誡部 述意部第一
夫以立像表真恆俗彝訓寄指詮月出道常
規但以妄想倒情沿流罔習無思悛革隨業
飄淪是以涅槃經云爲善清升辟同爪土爲
惡沉滯喻等地塵良由[illegible]俱至十使交縛

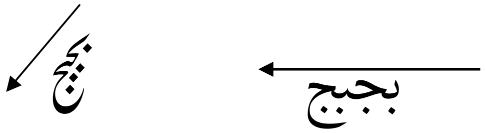
(a)

زبان کے معنی کا تعلق بولنے والے کی بہ نسبت سننے والے کے ذہن
سے زیادہ ہے۔

(b)

Figure 2: (a) Chinese Written Top Down (Vertically) [13] (b) Urdu Written in Arabic Script (Naskh Style) from Right to Left

iii) **Variety in baseline usage.** Unlike English which is written horizontally along a single baseline, there are writing systems which employ complex use of baseline, e.g. diagonal writing (top-right to bottom-left) in Nasta'leeq as shown below and explained in the next section.

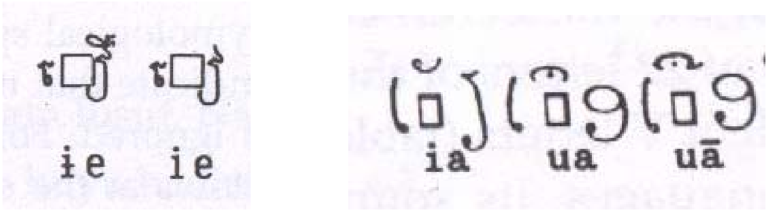


(b) (a)

Figure 3: (a) Urdu Written in Horizontal Naskh Style (b) Urdu Written in Diagonal Nasta'leeq Style

(Character Sequence ب ج ب ج Written Cursively in Both Cases)

iv) **Non-Monotonic writing.** In English characters start from left and are written towards the right. One does not come back left of an existing character to write a subsequent character. However, many writing systems in Asia show non-monotonic behavior. This is shown by Khmer and Lao vowels in Figure 4. The box represents a consonantal character and as the figure indicates, the vowels are written on both sides of consonants.



(a) (b)

Figure 4: Non-Monotonic Writing of Vowels in (a) Khmer, and (b) Lao [13]

v) **Marks in writing systems.** English has very limited sets of marks in its alphabet, e.g. the dot on letter 'i' and 'j'. However, many Asian scripts are much more rich the use of marks. Figure 5 shows how different vowel-marks modify the base letter for ‘k’.

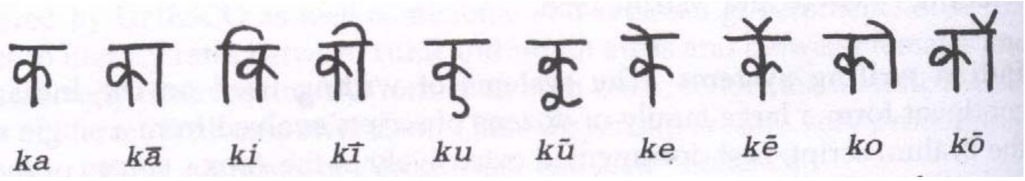


Figure 5: Use of Marks in Devanagari script [13]

vi) **Cursive behavior.** Many Asian writing systems are written by joining individual characters within a word or a phrase. The latter property is exhibited when there is no use of space between words and poses a significant challenge to determine end of line breaks. Loa is an example of such language.

vii) **Context sensitivity.** In many writing systems of Asia the shape of character is not fixed but depends on the shapes of characters around it, thus the shape of the letter is sensitive to the context in which it is occurring. The following example in Figure 6 shows various shapes of the same character 'bay' of Arabic in different contexts.

ب بب ببب

Figure 6: Context Dependent Shapes of Letter 'bay' in Arabic

These complexities pose significant challenges to technology. This study presents a case study of Urdu Nasta'leeq font highlighting the motivation, challenges, successes and lessons learnt. This font has been developed by Center for Research in Urdu Language Processing at National University of Computer an Emerging Sciences in Lahore, Pakistan, and supported through the Small grants program by IDRC, APDIP UNDP and APNIC.

Nasta'leeq Case Study

Urdu, the national language of Pakistan, is spoken by more than 60 million speakers in over 20 countries [14]. Urdu is written in Arabic script. Arabic script has many traditional writing styles, including Naskh (mostly used for Arabic language), Kufi, Divani, Sulus, Riqa, etc. Some of these styles are shown in Figure 7 below.

Naskh and Taleeq styles of writing were combined into the very spatially concise Nasta'leeq writing style.

Nasta'leeq writing system for Urdu is character based, bidirectional (mainly R to L), diagonal, non-monotonic, cursive, context sensitive writing system with a significant number of marks (dots and other diacritics). This makes Nasta'leeq one of the most complex writing styles and challenging to develop.

These complexities of the Nasta'leeq writing style had up till now precluded the development of a character-based font. Thus, it had not been possible to develop and publish text-based HTML web-pages for Urdu and images of scanned Urdu text had to be put on the internet for publishing. This made websites very memory intensive and were not accessible through slow internet connections, impeding the widespread usage of Urdu in electronic communication and publication. However, recent advances in font-technologies (e.g. Open Type Font) have made it possible to model and implement, to an acceptable level, Nasta'leeq type complex writing systems.

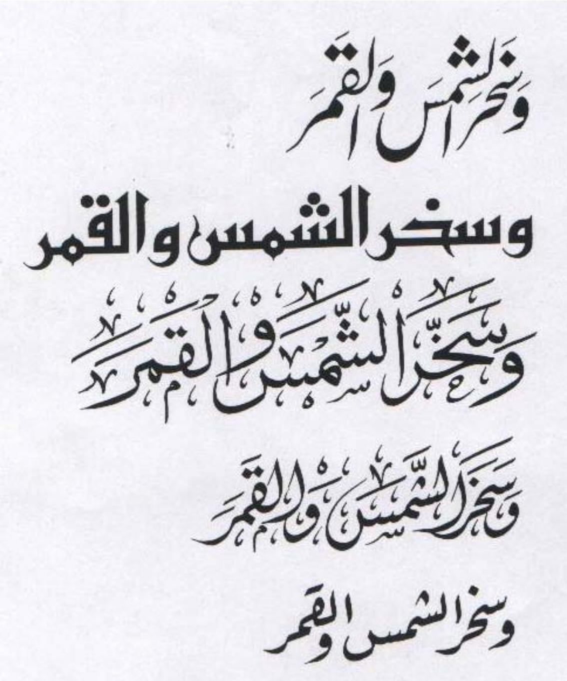


Figure 7: Various Writing Styles for Arabic Script (Unpublished Work Written by Jameel ur Rahman, Calligrapher of Nafees Nasta'leeq and Nafees Naskh Fonts)

To develop Nafees Nasta'leeq font an orthographic analysis of Nasta'leeq writing style was done. The following section briefly explains the findings of the work.

Challenges of Nasta'leeq Writing Style

Nasta'leeq is a complex cursive style of writing Arabic script based languages e.g. Urdu and Persian. Each letter has precise writing rules, relative to the width of the flat nib of the pen, called *qat*. The measurement of some letters in terms of *qat* is given in Figure 8.

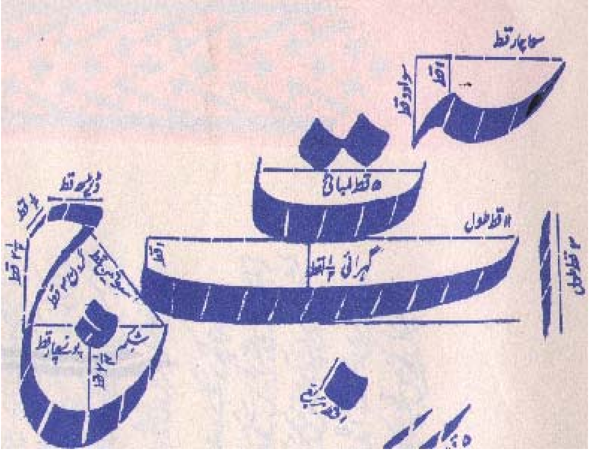


Figure 8: Measurements of Individual Letters in *Qat*

As Nasta'leeq is a writing style for Arabic script, it inherits its bidirectional nature, where the characters are written from R to L but numbers are written from L to R. In addition, some symbols take arguments above them (e.g. the digits indicating the year for Sanah sign). The bidirectional nature is shown in Figure 9 below. The arrows indicate the direction in which individual words are written while generally writing from right to left.

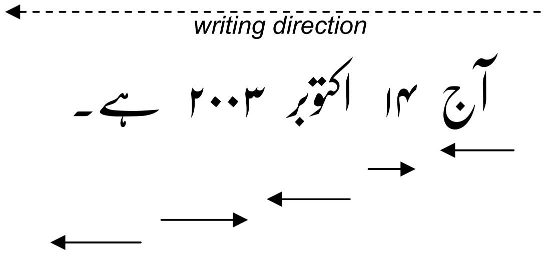
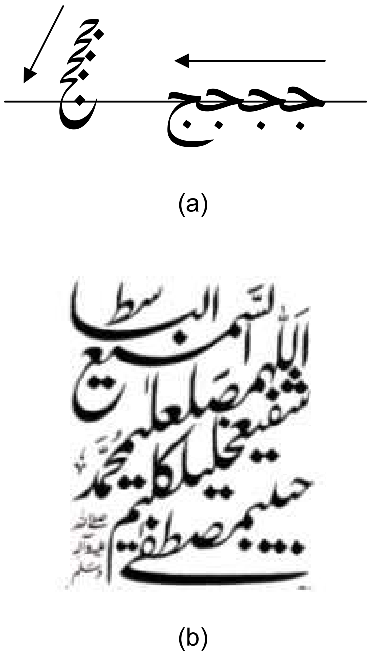


Figure 9: Bidirectional System for Writing Urdu in Nasta'leeq *("Today 14 October 2003 is"* i.e. "Today is 14th October 2003" is Written)

Additionally Nasta'leeq is a cursive writing style in which letters in a word connect at delicate joints in a diagonal manner, from top right to bottom left, according to well defined joining rules. This diagonal writing style gives rise to one of the defining features of Nasta'leeq - the vertical overlap of characters within and across ligatures (sequence of connected characters). Figure 10 (a) shows an example of Nasta'leeq's diagonality versus the horizontal behavior of Naskh, another popular writing style for Arabic script normally used for writing Arabic language. In both cases, the letter 'Jeem' is repeated four times (forming a nonsense string). A horizontal baseline is also drawn. This example also shows how Nasta'leeq is able to conserve space. Though both ligatures present the same information, the diagonal nature of Nasta'leeq makes the ligature more vertical but lesser in width compared to the ligature formed by Naskh writing style. Figure 10 (b) shows an artistic use of this diagonality feature of Nasta'leeq in a calligraphic composition.



(a)

(b)

Figure 10: (a) Nasta'leeq's (left) Diagonal vesus Naskh's (right) Horizontal Writing (b) Calligraphic Composition Showing Nasta'leeq Diagonality (Source Unknown)

Nasta'leeq is also non-monotonic, where certain letters contain a stroke which goes back and beyond the previous character. This is also true in writing where overlapping may be done to conserve space. This is shown in Figure 11 below, in which the strokes for letters 'Bari Yay' and ‘Jeem’ go back (towards right) beyond the previous characters. In both cases, letter ‘Bay’ is written as the first letter.

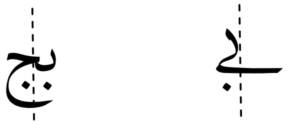


Figure 11: Negative Overlap of ‘Bari Yay’ and ‘Jeem’ in Nasta'leeq (Vertical Dotted Line Marks the End of First Character)

Nasta’leeq style is complex as letters in a word assume different shapes according to the context in which they occur. For less than 50 characters a total of about a thousand different shapes are coded in Nafees Nasta'leeq to cater to the context sensitivity of characters. Special rules are written to choose the right shape of a character in a given context. Figure 12 shows four such rules which choose the appropriate shape for letter ‘Jeem’ written in Nafees Nasta'leeq font. Second rule can be read as “’Jeem’ takes JeemInitial4 shape (given in middle column) when it comes before RayFinal1”. Actual context is given in the right-most column.

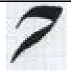
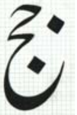
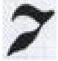
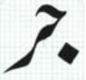
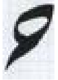
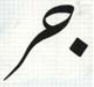
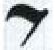
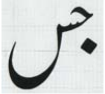
ج → ج Initial3 / ___ ج Final | ___<ج> <NS>*<All>	 ج Initial3	
ج → ج Initial4 / ___ ر Final1	 ج Initial4	
ج → ج Initial5 / ___ ر Final2	 ج Initial5	
ج → ج Initial6 / ___ س Final1 | ___<س><NS>*<All>	 ج Initial6	

Figure 12: Context Dependent Rules of Nafees Nasta'leeq for Urdu

Marks are indispensable in Nasta'leeq for Urdu. The marks can either be ‘Nuqta’ (dot), ‘aerab’ (vowel diacritics) or honorifics or other special symbols. ‘Nuqta’ marks are an integral and defining part of some characters and are mandatory. Other marks are mostly optional. These marks are also used in other Arabic Scripts (see Figure 7 above) but pose a significantly greater challenge in Nasta'leeq because of its diagonality and shorter horizontal span. A calligrapher can artistically move these marks around while writing. However, a computer font is normally dumb and has fixed positions for these marks relevant to the base character often causing base-mark or mark-mark collisions. Theses collisions can be avoided by writing rules to move the marks in case a clash is anticipated. There are many such possibilities. More than 2000 rules have been written in Nafees Nasta'leeq to avoid the ‘Nuqta’-clashes. Figure 13 below shows (a) base-mark type ‘Nuqta’ clash (circled), and (b) how a calligrapher avoids the clash by shifting the ‘Nuqta’ marks (the circled ‘Nuqta’ marks are moved in the direction of the arrow, away from previous character).

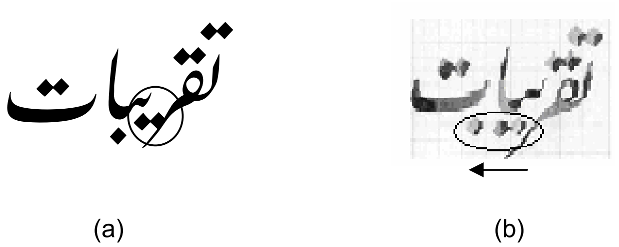


Figure 12: (a) 'Nuqta'-Clash, and (b) Clash Removal by Shifting 'Nuqta'

In addition to these, there are significant spacing and justification requirements in Nasta'leeq, which OTF formalism does not handle and may eventually be programmed by developing rule or heuristics based expert systems.

Success Story

With a concentrated effort of a five person team for eighteen months funded by Small Grants Program (by IDRC, UNDP APDIP and APNIC), Nafees Nasta'leeq, a character based Nasta'leeq font for Urdu was developed and released on 14th August 2003. This font is freely available from www.crulp.nu.edu.pk or www.crulp.org. Because this font is based on Unicode standard and OTF formalism and is freely available, it is now possible for Urdu speakers to publish and access Urdu over internet without resorting to images. This would allow Urdu speakers to use the internet and other computing applications as well and remove the English language barrier.

In process of development of Nafees Nasta'leeq font, it has also been possible to completely document the details of Nasta'leeq writing style. This is a calligraphic achievement as these rules, which had been passed down for generations, had never been documented in detail.

The research center has also been able to build significant HR capacity to develop fonts by doing this work.

Finally, the font also successfully depicts Nafees Shah's Lahore style of Nasta'leeq and thus preserves his tradition.

Lessons Learnt

In the course of development of Nafees Nasta'leeq, R&D team faced multiple challenges. As the project leader, the author also went through a learning process which looked not only on the development challenges, but also at other issues involved in achieving the ultimate objective for the work: to enable masses in Pakistan access to computing in a language they understand. These findings are listed below.

1. Open Type Font formalism is much more advanced and powerful compared to the simple collection of glyphs in True Type Font formalism. OTF is a very effective formalism to realize complex Asian scripts. However, it still has some problems.

    - Though Uniscribe, the OTF rendering engine by Microsoft, is the most mature among existing rendering engines, it still does not support all functionality defined in OTF formalism.
    - OTF formalism allows limited amount of space for writing rules. This space is sufficient for many writing systems and styles. However, due to the complexity of Nasta'leeq, all this space was utilized in creating Nafees Nasta'leeq. Still more rules were required, but were not incorporated because of space constraints.

- As the number of rules increase, the performance of Uniscribe to render the OTF font depreciates. Thus, large documents are very slowly displayed.
- Uniscribe and OTF are based on Unicode standard which is not supported by Linux and older versions of Microsoft (e.g. Windows 95/98). Thus, there is still limited usage of OTF based fonts, though this will change over next few years.
- Nasta'leeq is a very complex writing style and has significant artistic and calligraphic requirements. The artistic nature of Nasta'leeq is created through very complex balancing rules of ligatures and words and by overlapping or stretching different characters. Developing Nasta'leeq font to compete with the creation of a calligrapher's mind is a tough task. Nafees Nasta'leeq has achieved the functionality of the font but still lacks the beauty of creation of Nafees Shah, the master after which the font has been named. Open Type Font formalism is not powerful enough to handle this aspect of the font.
- Linux does not have a mature rendering engine for OTF. This limits the usage of OTF font on Linux platform. The efforts on developing this support on Linux are slow. This is an area which needs to be urgently addressed.

2. Apart from development challenges, there were significant other lessons learnt. As a computer scientist, project leader has limited exposure to the 'real world' scenarios. Naively, it was thought that if the font is built it will be automatically used by the masses. Interestingly, in the project budget there was no initial allocation for marketing the font. Once the font was developed, the development team realized how crucially a marketing budget was required. Developers do ineffective marketing. Though email announcement were made, it was not reaching the masses. Some budget was re-appropriated for this purpose. However, one of the significant lessons learnt is that effective and planned marketing needs to be done to really enable such products to be effectively disseminated. Dissemination of information can be achieved by partnering developers with social scientists and through proper planning and resource allocation.

3. Font is only one piece of LICT4D.asia puzzle. To provide access of developing populations in local languages for effective use, much more research and development in areas of script, speech and language processing applications is required. Special focus also needs to be put on developing localization capability on Linux and open sources to give cheap alternative to poor populations. Proper funding needs to go into this.

4. HR capacity needs to be built to enable LICT development for Asia. Developing countries do not have enough HR capability to do much of the advanced R&D to develop local language software.

5. Governments need to focus on LICT as an important tool and effectively address development of LICT in their IT related policies.

Conclusions

Most populations in Asia require urgent access information for their development. Internet is the main repository and vehicle for access of information. However, language barriers pose a significant problem to this access. This is an urgent and important issue and must be addressed to give information access to poor populations of Asia, the access which is their basic human right! This may be effectively achieved by computer and social scientists working closely together to develop technology and effective dissemination mechanisms for the technology, and public sector supporting such work.

Publication Note and Acknowledgements

This paper was originally presented at 12th AMIC ANNUAL CONFERENCE: E-WORLDS: GOVERNMENTS, BUSINESS AND CIVIL SOCIETY, NOVEMBER 2003, SINGAPORE. The work presented has been partially supported by a Small Grants Program by IDRC, APDIP UNDP and APNIC. The work on Nasta'leeq has been completed by the Nafees Nasta'leeq team at CRULP, NUCES.

References

[1] “**Info-structure in developing countries,**” UNESCO’s contribution for the World Summit on the Information Society (WSIS) (8 February 2002).

[2] “**Statement Presented by Minister of Information and Communications Technology of Thailand,”** Regional Conference for WSIS in Tokyo Jan 12-15, 2002).

[3] **“Access Features Information Technology for Rural Community Development**,” http://www.ncsa.uiuc.edu/News/Access/Stories/InfoTechnology/information.html

[4] http://www.worldbank.org/data/countryclass/classgroups.htm#top.

[5] www.worldpop.org/datafinder.htm .

[6] “**Is it time for multilingual publishing online?”** 25th February, 2001, www.onlinepublishingnews/secure/assets.htm.

[7] “**Information and Communication Technology in Developing Countries of Asia**,” Brahm Prakash, http://www.adb.org/Documents/Conference/Technology_Poverty_AP/adb6.pdf

[8] http://www1.linkclub.or.jp/~jafae/Englishes/Greetings.htm.

[9] **Tools and Statistics Unit**, http://www.unhabitat.org/habrdd/asia.html

[10] “**The effect of native language on Internet usage**”, Telecommunications Policy Research Conference (TPRC) 29th Research Conference on Communication, Information and Internet Policy, October 27-29, 2001, Alexandria, Virginia.

[11] "**Content and ICTs: challenges, innovation and prospects**" Digital Review of Asia Pacific 2003/2004, www.digital-review.org.

[12] http://logos.uoregon.edu/explore/orthography/chinese.html.

[13] Florian Coulmas, 1999 “**The BlackWell Encyclopedia of Writing Systems**”, BlackWell Publishers Inc, Massachusettes, USA.

[14] www.ethnologue.com.